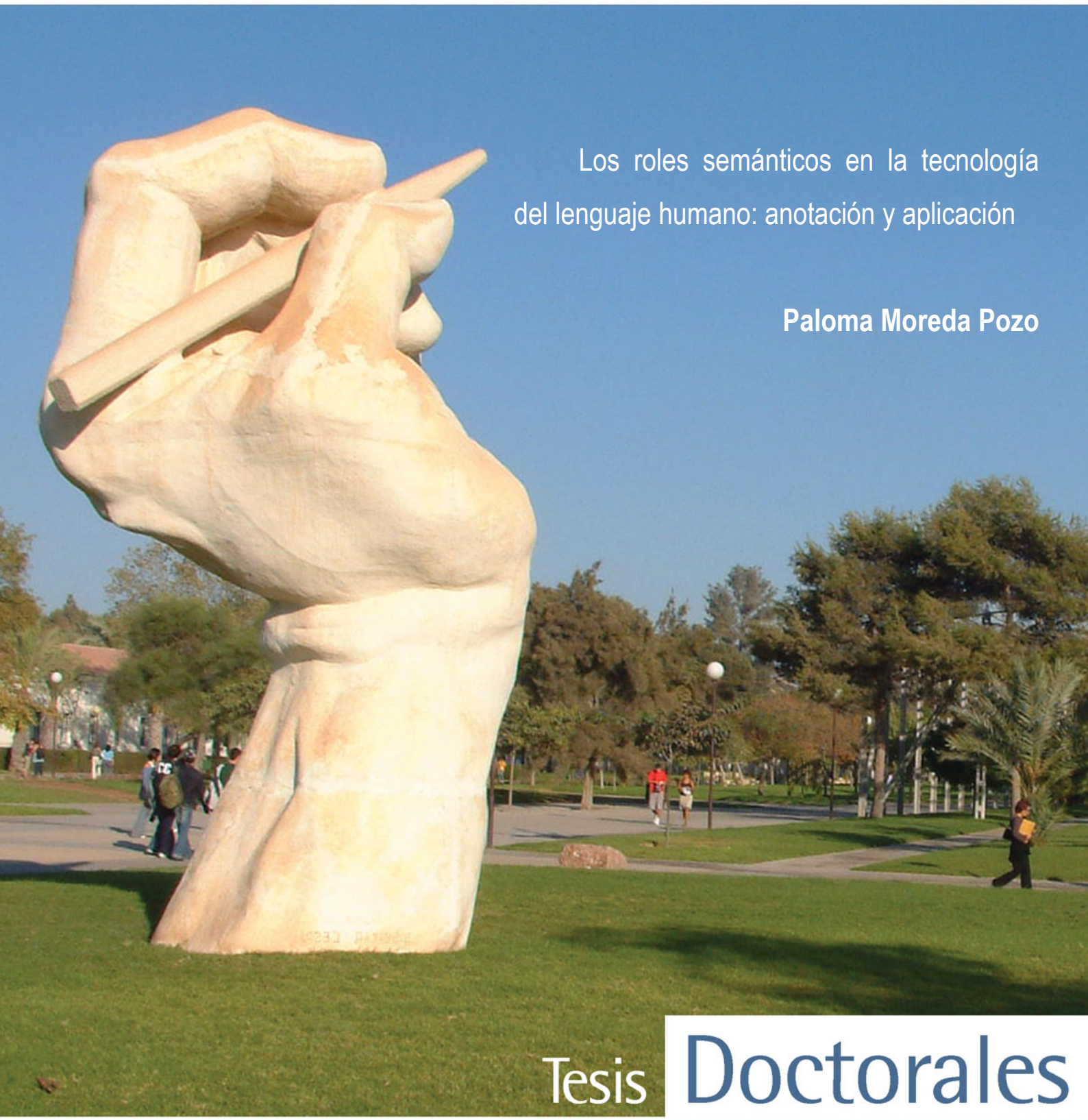
Universitat d'Alacant
Universidad de Alicante



Los roles semánticos en la tecnología del lenguaje humano: anotación y aplicación

Paloma Moreda Pozo

Tesis Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE

Los Roles Semánticos en la Tecnología del Lenguaje Humano: Anotación y Aplicación.

Universitat d'Alacant
Universidad de Alicante

Tesis Doctoral

Paloma Moreda Pozo

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Los Roles Semánticos en la Tecnología del Lenguaje Humano: Anotación y Aplicación.

Tesis Doctoral

Paloma Moreda Pozo

Dirigida por Dr. Manuel Palomar Sanz

Mayo 2008

Universitat d'Alacant
Universidad de Alicante

Índice general

1. Introducción 1

1.1. Organización de la Tesis 12

2. Roles Semánticos: Estado de la cuestión 15

2.1. Análisis de propuestas de conjuntos de roles semánticos 17

2.1.1. Propuesta de Gruber 19

2.1.2. Propuesta de Fillmore. Gramática de casos . 19

2.1.3. Propuesta de Sgall et al. Descripción Generativa Funcional 21

2.1.4. Propuesta de Celce-Murcia 23

2.1.5. Propuesta de Schank. Teoría de la dependencia conceptual 23

2.1.6. Propuesta de Folley y Van Valin. Macropapeles de la gramática del rol y la referencia . 24

2.1.7. Propuesta de Jackendoff 26

2.1.8. Propuesta de Dowty 27

2.1.9. Propuesta del proyecto FrameNet 28

2.1.10. Propuesta del proyecto PropBank 30

2.1.11. Propuesta de de roles semánticos para sistemas de BR 33

Universitat d'Alacant
Universidad de Alicante

2.1.12.Otras propuestas 36
2.2. Recursos lingüísticos basados en roles semánticos . . 39
2.2.1. Proyecto PropBank 39
2.2.2. Proyecto FrameNet 46
2.2.3. Otros recursos lingüísticos 54
2.3. Relaciones entre recursos 67

3. Enfoques para el tratamiento de Roles Semánticos 71
3.1. Enfoques basados en corpus 72
3.1.1. Aprendizaje automático supervisado 73
3.1.2. Aprendizaje automático semi-supervisado . . . 85
3.1.3. Aprendizaje automático no supervisado 88
3.1.4. Selección de características 90
3.2. Enfoques basados en conocimiento 101
3.2.1. Representación basada en reglas 105
3.2.2. Lógica de predicados 105
3.2.3. *Frames* 106

4. Sistemas de Anotación Automática de Roles Semánticos 107
4.1. Enfoques basados en corpus 108
4.1.1. Aprendizaje automático supervisado 108
4.1.2. Aprendizaje automático semi-supervisado . . . 122
4.1.3. Aprendizaje automático no supervisado 122
4.2. Enfoques basados en conocimiento 126
4.2.1. Representación basada en reglas 126
4.2.2. Representación basada en *frames* 130

Universitat d'Alacant
Universidad de Alicante

4.3. Campañas internacionales de evaluación de SRL . . . 130
4.3.1. CoNLL shared task 130
4.3.2. Senseval 138

5. Aportación a la anotación automática de Roles Semánticos 151
5.1. Introducción 151
5.2. SemRol: Una herramienta de anotación automática de roles semánticos 153
5.2.1. Corpus 154
5.2.2. Conjunto de roles semánticos 156
5.2.3. Estrategia de anotación 157
5.2.4. Algoritmo de aprendizaje 160
5.2.5. Información utilizada 165
5.2.6. Arquitectura de SemRol 171
5.3. Módulo de procesamiento *off-line* de SemRol 175
5.3.1. Características utilizadas 177
5.3.2. Máquina de aprendizaje 180
5.3.3. Mejor Conjunto de características 186
5.4. Módulo de procesamiento *on-line* de SemRol 189
5.5. Evaluación de SemRol 191
5.5.1. Proceso de ajuste 192
5.5.2. Clasificador por sentidos frente Clasificador único 194
5.5.3. Clasificador individual frente Clasificador global 199
5.5.4. Comparación con otros sistemas de anotación 200

Universitat d'Alacant
Universidad de Alicante

6. Los Roles Semánticos en aplicaciones de Búsqueda de Respuestas 203

6.1. Sistemas de Búsqueda de Respuesta 206

6.2. Uso de roles semánticos en sistemas de BR 209

6.2.1. Conjunto de roles semánticos utilizados 210

6.2.2. Papel de los roles semánticos 211

6.2.3. Principales conclusiones 217

6.3. SemRol en sistemas de BR 219

6.3.1. Sistema de BR desarrollado 219

6.3.2. Extracción de respuestas basada en roles semánticos 225

6.4. Análisis de la utilidad de los roles semánticos en sistemas de BR 233

6.4.1. Extracción de respuesta basada en Reglas frente a Patrones 234

6.4.2. Comparación con sistemas de BR basados en NE 235

6.4.3. Comparación con otros sistemas de BR basados en roles 238

6.5. Ejemplo de construcción de patrones semánticos . . . 239

7. Conclusiones y trabajos futuros 243

7.1. Conclusiones 243

7.2. Aportaciones al conocimiento de la investigación en roles semánticos 246

7.3. Lista de publicaciones relevantes 249

7.4. Trabajo en progreso y futuro 253

8. Anexo 257

Universitat d'Alacant
Universidad de Alicante

Bibliografía 291

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Índice de cuadros

2.1. Resumen de las principales propuestas de conjuntos de roles 18
2.2. Detalle del conjunto de roles propuesto en (Gruber, 1965) 19
2.3. Primer conjunto de roles propuesto por Fillmore (1968) 20
2.4. Recopilación de roles temáticos propuestos por Fillmore en sus diferentes trabajos 20
2.5. Tipos de relaciones de dependencia en FDG (Hajič, 2004) 22
2.6. Detalle de los roles temáticos propuestos por (Celce-Murcia, 1972) 23
2.7. Casos conceptuales propuestos por (Schank, 1972) 24
2.8. Versión inicial del conjunto de roles propuesto por (Jackendoff, 1990) 27
2.9. Conjunto de roles refinado propuesto por (Jackendoff, 1990) 27
2.10. Propiedades de un proto-agente según Dowty (1991) . . 28
2.11. Propiedades de un proto-paciente según Dowty (1991) . 28
2.12. Conjunto de roles en FrameNet para el marco semántico de la comunicación verbal 30
2.13. Ejemplo de dos conjuntos de roles del verbo *decline* en PropBank 31
2.14. Tendencias de los argumentos numerados de PropBank 32

Universitat d'Alacant
Universidad de Alicante

2.15. Lista de etiquetas de función de adjuntos en PropBank 32
2.16. Resumen de otras propuestas de roles semánticos (1/2) 37
2.17. Resumen de otras propuestas de roles semánticos (2/2) 38
2.18. Resumen de los recursos más utilizados 40
2.19. Ejemplo de un *frameset* en PropBank 42
2.20. Ejemplo de los participantes de un *frameset* de SemFrame versión 2.0 53
2.21. Lista de roles utilizados en el proyecto SenSem 60
2.22. Conjunto de roles utilizados en LCS 61
2.23. Sentidos del verbo *drop* en LCS 62
2.24. Modelo básico para verbos de trayectoria 64
2.25. Procedimientos para relacionar recursos 69
2.26. Correspondencia entre PropBank y la propuesta de Moreda *et al.* (2007) 69

3.1. Otros algoritmos de aprendizaje supervisado utilizados en PLN 75
3.2. Aspectos a establecer en cualquier proceso de selección de características 92
3.3. Características de los principales métodos de selección de características 98
3.4. Otros métodos de selección de características (1/3) 102
3.5. Otros métodos de selección de características (2/3) 103
3.6. Otros métodos de selección de características (3/3) 104

4.1. Detalle de las siglas utilizadas en la columna OBS en los cuadros de resultados 4.2, 4.3, 4.4 121
4.2. Datos sobre la evaluación de sistemas de SRL supervisados 123

Universitat d'Alacant
Universidad de Alicante

4.3. Datos sobre identificación de argumentos de sistemas de SRL supervisados 124
4.4. Datos sobre asignación de roles de sistemas de SRL supervisados 125
4.5. Datos sobre la evaluación de sistemas de SRL no supervisados 126
4.6. Datos sobre la evaluación de sistemas de SRL basados en conocimiento 129
4.7. Resultados de la *shared task* del CoNLL-2004 sobre el conjunto de desarrollo 133
4.8. Resultados de la *shared task* del CoNLL-2004 sobre el conjunto de test 133
4.9. Resultados de la *shared task* del CoNLL-2004 sobre el conjunto de test. Fase de asignación de roles 134
4.10. Resultados de la *shared task* del CoNLL-2005 sobre el conjunto de desarrollo 138
4.11. Resultados de la *shared task* del CoNLL-2005 sobre el conjunto de test 139
4.12. Resultados de la *shared task* del CoNLL-2005 sobre el conjunto de test del corpus Brown 140
4.13. Resultados de la *shared task* del CoNLL-2005 sobre el conjunto de test. Fase de clasificación. 10 mejores sistemas 140
4.14. Resultados de la tarea restrictiva en Senseval-3 143
4.15. Resultados de la tarea no restrictiva en Senseval-3 143
4.16. Resultados de SemEval. Tarea: SRL para catalán y español 145
4.17. Resultados de SemEval. Tarea: SRL para árabe 146
4.18. Resultados de SemEval. Tarea: Estructura semántica . . 148
4.19. Resultados de SemEval. Tarea: SRL para inglés 149

Universitat d'Alacant
Universidad de Alicante

5.1. Características generales de SemRol 153
5.2. Tendencias de los argumentos numerados de PropBank 157
5.3. Lista de etiquetas de función de adjuntos en PropBank 158
5.4. Algunos sentidos y sus roles semánticos para el verbo *give* en PropBank 159
5.5. Resumen del proceso realizado para determinar un tamaño de k adecuado 164
5.6. Detalle de las características utilizadas (1/2) 172
5.7. Detalle de las características utilizadas (2/2) 173
5.8. Lista de argumentos de la oración (E44) 173
5.9. Ejemplo de valores de las características utilizadas para la oración (E44) (1/2) 175
5.10. Ejemplo de valores de las características utilizadas para la oración (E45) (1/2) 176
5.11. Detalle de la información proporcionada por el corpus PropBank para la oración (E45). Oración de un sólo verbo. 180
5.12. Detalle de la información proporcionada por el corpus PropBank para la oración (E46) (2/1). Oración de dos verbos. 181
5.13. Detalle de la información proporcionada por el corpus PropBank para la oración (E46) (2/2). Oración de dos verbos. 182
5.14. Resultados del proceso de selección de características . . 188
5.15. Detalle de las características del clasificador de lugar . . 189
5.16. Comportamiento de las características en la clasificación por sentidos. Algoritmo TiMBL. 193
5.17. Comportamiento de las características en la clasificación única. Algoritmo TiMBL. 193

Universitat d'Alacant
Universidad de Alicante

5.18. Comportamiento de las características en la clasificación por sentidos. Algoritmo ME. 194
5.19. Resultados de los clasificadores por sentidos (vs) y únicos (u) 195
5.20. Comparativa de tiempos de ejecución entre algoritmos de aprendizaje y estrategias de anotación 196
5.21. Comportamiento de los clasificadores para cada tipo de rol cuando se sigue una estrategia de anotación por sentidos del verbo (vs) y cuando no (u). Resultados de $F_{\beta=1}$ medida. 197
5.22. Promedios e incrementos de mejora cuando se sigue una estrategia de anotación por sentidos del verbo (vs) y cuando no (u). Resultados de $F_{\beta=1}$ medida. 198
5.23. Influencia del análisis sintáctico en la anotación de roles 198
5.24. Resultados de los clasificadores específicos para cada tipo de rol 199
5.25. Comparativa de los resultados obtenidos con clasificadores individuales frente a los globales. 200
5.26. Comparación de SemRol con otros sistemas de SRL . . . 201

6.1. Resumen de las principales características de los sistemas de BR que hacen uso de roles semánticos 211
6.2. Resumen del uso de roles semánticos en sistemas de BR 213
6.3. Resultados del uso de roles semánticos en sistemas de BR 217
6.4. Conjunto de relaciones semánticas pregunta-rol_semántico 226
6.5. Correspondencia entre PropBank y la propuesta de Moreda et al. 228
6.6. Resultados para un sistema de BR basado en roles semánticos 234

Universitat d'Alacant
Universidad de Alicante

6.7. Resultados para sistemas de BR basados en roles semánticos y en entidades para respuestas NE y no NE 237

6.8. Comparación de diferentes sistemas de BR basados en roles semánticos 239

6.9. Ejemplos de patrones generados para la pregunta *Where is the actress, Marion Davies, buried?* 239

8.1. Combinaciones con 1 característica. TiMBL. Anotación por sentidos 260

8.2. Combinaciones con 2 características. TiMBL. Anotación por sentidos 261

8.3. Combinaciones con 3 características. TiMBL. Anotación por sentidos 262

8.4. Combinaciones con 4 características. TiMBL. Anotación por sentidos 263

8.5. Combinaciones con 5 características. TiMBL. Anotación por sentidos 264

8.6. Combinaciones con 6 características. TiMBL. Anotación por sentidos 265

8.7. Combinaciones con 7 características. TiMBL. Anotación por sentidos 266

8.8. Combinaciones con 8 características. TiMBL. Anotación por sentidos 267

8.9. Combinaciones con 9 características. TiMBL. Anotación por sentidos (1/2) 268

8.10. Combinaciones con 9 características. TiMBL. Anotación por sentidos (2/2) 269

8.11. Combinaciones con 10 características. TiMBL. Anotación por sentidos (1/5) 270

Universitat d'Alacant
Universidad de Alicante

8.12. Combinaciones con 10 características. TiMBL. Anotación por sentidos (2/5) 271
8.13. Combinaciones con 10 características. TiMBL. Anotación por sentidos (3/5) 272
8.14. Combinaciones con 10 características. TiMBL. Anotación por sentidos (4/5) 273
8.15. Combinaciones con 10 características. TiMBL. Anotación por sentidos (5/5) 274
8.16. Combinaciones con 11 características. TiMBL. Anotación por sentidos (1/4) 275
8.17. Combinaciones con 11 características. TiMBL. Anotación por sentidos (2/4) 276
8.18. Combinaciones con 11 características. TiMBL. Anotación por sentidos (3/4) 277
8.19. Combinaciones con 11 características. TiMBL. Anotación por sentidos (4/4) 278
8.20. Combinaciones con 12 características. TiMBL. Anotación por sentidos (1/3) 279
8.21. Combinaciones con 12 características. TiMBL. Anotación por sentidos (2/3) 280
8.22. Combinaciones con 12 características. TiMBL. Anotación por sentidos (3/3) 281
8.23. Combinaciones con 13 características. TiMBL. Anotación por sentidos (1/3) 282
8.24. Combinaciones con 13 características. TiMBL. Anotación por sentidos (2/3) 283
8.25. Combinaciones con 13 características. TiMBL. Anotación por sentidos (3/3) 284
8.26. Combinaciones con 14 características. TiMBL. Anotación por sentidos (1/5) 285

Universitat d'Alacant
Universidad de Alicante

8.27. Combinaciones con 14 características. TiMBL. Anotación por sentidos (2/5) 286
8.28. Combinaciones con 14 características. TiMBL. Anotación por sentidos (3/5) 287
8.29. Combinaciones con 14 características. TiMBL. Anotación por sentidos (4/5) 288
8.30. Combinaciones con 14 características. TiMBL. Anotación por sentidos (5/5) 289

Universitat d'Alacant
Universidad de Alicante

Índice de figuras

1.1. Proceso de análisis de una oración. 3

1.2. Posibles árboles de análisis sintáctico de la oración *John saw the thief with the binoculars*. 6

2.1. Continuo de relaciones temáticas en RRG 25

2.2. Jerarquía actor-afectado en RRG 26

2.3. Conjunto de roles propuesto por Moreda et al. 33

2.4. Conjunto de roles semánticos utilizados en Sinica Treebank 56

3.1. Ejemplo de un problema linealmente separable en un espacio de dos dimensiones. 80

3.2. Ejemplo de un problema linealmente no separable en un espacio de dos dimensiones. 82

3.3. Función de distribución empírica como una estimación de la función de distribución verdadera. 86

3.4. Espacio de búsqueda para un conjunto de cuatro características. 93

3.5. Algoritmo de búsqueda *forward selection*. 94

3.6. Algoritmo de búsqueda *backward elimination*. 94

3.7. Modelo filtro para selección de características. 95

3.8. Modelo *wrapper* para selección de características. 96

Universitat d'Alacant
Universidad de Alicante

3.9. Arquitectura básica de un sistema basado en conocimiento. 104

5.1. Arquitectura del sistema para anotación de roles semánticos: SemRol. 174

6.1. Uso de roles semánticos en búsqueda de respuestas. . . . 206

6.2. Arquitectura de un sistema de BR basado en roles semánticos. 220

6.3. Reglas utilizadas para identificar las preguntas de tipo lugar. 221

Universitat d'Alacant
Universidad de Alicante

1. Introducción

La lengua ha sido objeto de interés desde la edad antigua y no sólo para lingüistas, sino también para otros colectivos tales como filósofos, psicolingüistas o ingenieros (Moreno *et al.*, 1999c). Cada uno de estos colectivos estudia la lengua desde puntos de vista y propósitos distintos. En el campo de la ingeniería informática el objetivo radica en encontrar mecanismos computacionales efectivos que permitan comprender y generar el lenguaje natural, facilitando la interrelación hombre/máquina y permitiendo una comunicación mucho más fluida y menos rígida que los lenguajes formales. Dicho de forma más sencilla, el objetivo es investigar la realización de aplicaciones informáticas que imiten la capacidad humana de hablar y entender.

Estas investigaciones dieron lugar a la denominada `Tecnología del Lenguaje Humano` (TLH). Este área de la ingeniería, rama de la Inteligencia Artificial (IA), engloba

- El reconocimiento del modo de entrada de la información.
- La capacidad de reconocer, comprender, interpretar y generar lenguaje, conocida como `Procesamiento del Lenguaje Natural` (PLN), linguística informática (LI), o linguística computacional (LC)[1].

[1] Algunos investigadores, como (Martí *et al.*, 2003), realizan pequeñas distinciones entre estos tres conceptos considerando la LC como la línea de investigación general que engloba a las otras áreas: PLN, la parte centrada en los aspectos más aplicados de la LC, como pueden ser la traducción automática de textos o los sistemas de búsqueda de respuestas, y la LI la parte orientada al desarrollo de programas de apoyo a los estudios filológicos, lexicográficos, linguísticos, etc.)

Universitat d'Alacant
Universidad de Alicante

- La realización de aplicaciones finales y desarrollo de la tecnología.

Los primeros intentos de procesamiento del lenguaje natural se remontan a finales de la década de los cuarenta y principios de los cincuenta, y se centran en procesos de traducción automática[2]. En los años 1950, EEUU realizó esfuerzos para obtener ordenadores capaces de traducir textos automáticamente de lenguas extranjeras al inglés, concretamente de revistas científicas rusas. Para traducir un lenguaje en otro, se observó que era necesario entender la sintaxis de ambos lenguajes, al menos en el nivel de morfología (la sintaxis de las palabras) y las frases enteras. Para entender la sintaxis, se debe entender la semántica del vocabulario y la pragmática del lenguaje. De esta manera, lo que empezó como un esfuerzo para traducir textos se convirtió en una disciplina encargada de entender cómo representar y procesar el lenguaje natural utilizando ordenadores.

Por ello, cualquier sistema de PLN que intente simular un comportamiento lingüístico humano, debe tomar conciencia tanto de las estructuras propias del lenguaje, incluyendo las palabras, cómo combinar éstas para formar oraciones, qué significan las palabras, o cómo contribuye el significado de las palabras al significado de la oración; como del conocimiento general acerca del universo de discurso y la capacidad de razonamiento. Todo ello sin olvidar la ambigüedad intrínseca del lenguaje, que es quizá el mayor problema al abordar la tarea de la comprensión computacional del lenguaje.

Universitat d'Alacant
Universidad de Alicante

Atendiendo a las necesidades anteriores y a la clasificación tradicional que los lingüistas han hecho de las formas de conocimiento de la lengua, un sistema computacional divide las fases o niveles de análisis de una oración en: análisis léxico-morfológico, sintáctico, semántico y pragmático o contextual (Moreno *et al.*, 1999c). Estas fases de anotación son dependientes y acumulativas. Por un lado, cada fase necesita de la información proporcionada por la fase anterior; y por otro, la información proporcionada por

[2] http://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales

cada una de las fases incluye y amplía la información que la fase anterior le proporciona. (Ver figura 1.1).

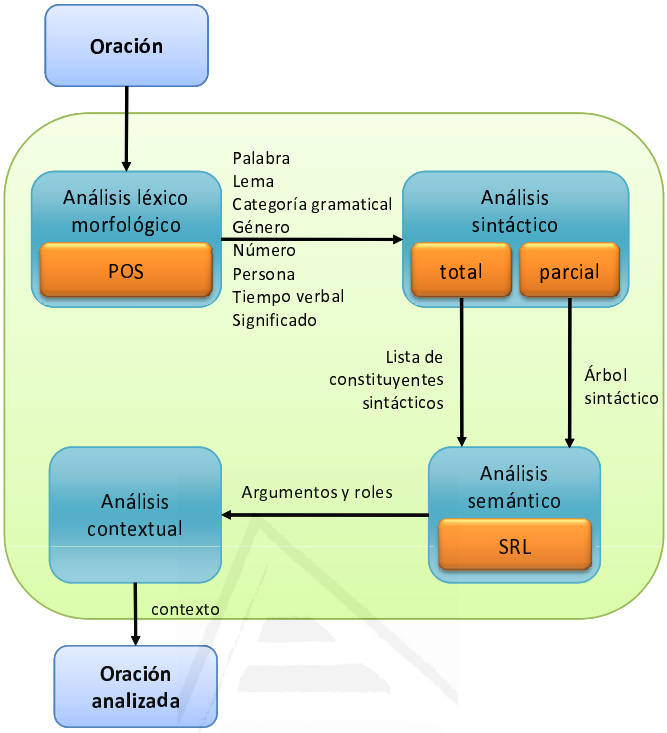


Figura 1.1. Proceso de análisis de una oración.

Universitat d'Alacant
Universidad de Alicante

- **Análisis léxico-morfológico**, también denominado PoS (del inglés *Part-of-Speech tagger*). El objetivo es asignar a cada palabra de la oración, un lema; una categoría gramatical (nombre, verbo, adjetivo, etc.); el género, número, y persona, así como los tiempos y modos verbales, en el caso de verbos; y su significado en la oración. El ejemplo (E2) muestra el resultado del análisis léxico de la oración (E1).

(E1) John saw the thief with the binoculars

(E2) [$_{NNP\ john}$ John] [$_{VBD\ see\ \#1:percibir_por_la_vista}$ saw] [$_{DT\ the}$ the] [$_{NN\ thief\ \#1:criminal}$ thief] [$_{IN\ with}$ with] [$_{DT\ the}$ the] [$_{NNS\ binocular\ \#1:instrumento_optico}$ binoculars]

A este nivel de análisis, la ambiguedad del lenguaje provoca problemas a la hora de:

- Determinar la categoría gramatical de una palabra. Por ejemplo, las oraciones (E3) y (E4) contiene la palabra *work*. Sin embargo, su categoría gramatical es diferente en cada una de ellas. En la oración (E3) *work* es un nombre y en (E4) es un verbo.

  (E3) This [$_{NN}$ piece of work] is very important to you

  (E4) John will [$_{VB}$ work] at the factory tomorrow

- Elegir el significado de una palabra de entre todos sus posibles significados. Por ejemplo, las oraciones (E5) y (E6) contienen ambas la palabra *bank*. Sin embargo, su significado varía de una oración a otra. En el caso de la oración (E5), *bank* representa una pila de objetos similares; y en el de la oración (E6), representa una entidad financiera.

  (E5) John threw a bank$_{\#3:pila}$ of newspapers

  (E6) John came into the bank$_{\#2:entidad_financiera}$

Universitat d'Alacant
Universidad de Alicante

- **Análisis sintáctico**. Analiza la secuencia de unidades léxico-morfológicas de cada oración produciendo una representación de su estructura, normalmente, en forma de árbol. Esta estructura sintáctica indica cómo las palabras se agrupan en otros constituyentes de la oración (sintagmas nominales, preposicionales, verbales, etc.), qué palabras modifican a otras, y qué palabras tienen una importancia central en la oración, así como el tipo de relación que existe entre constituyentes.

En ocasiones, en este tipo de análisis se sacrifican la completitud y profundidad del análisis, limitándolo a la identificación de los constituyentes sintácticos sin tratar las dependencias o relaciones entre los mismos. A cambio se obtienen mayor velocidad y robustez, dado que siempre se obtiene una representación de la oración aunque sea parcial y menos valiosa. Estos dos enfoques han dado lugar a los denominados análisis global o completo y análisis parcial o superficial, respectivamente.

La limitación del análisis parcial no es más que una consecuencia de la ambiguedad, denominada ambiguedad estructural, con la que los analizadores sintácticos se encuentran al determinar qué palabras se agrupan formando los constituyentes de una oración y las relaciones existentes entre ellos. Por ejemplo, en la oración (E1) es difícil determinar si *John* utilizó los prismáticos para ver al ladrón, o si el ladrón al que vió *John*, llevaba unos prismáticos. Por tanto, cualquiera de los árboles de la figura 1.2 sería posible. Sin embargo, si se realiza un análisis parcial de la oración (ver ejemplo (E7)) el problema de la ambiguedad estructural no se contempla.

(E7) [$_{NP}$ John] [$_{VP}$ saw] [$_{NP}$ the thief] [$_{PP}$ with] [$_{NP}$ the binoculars]

- **Análisis semántico**, también conocido como interpretación semántica. Tiene por objetivo identificar relaciones entre palabras de un texto, dando lugar a estructuras que reflejan varios niveles de interpretación semántica del texto (Shi & Mihalcea, 2005). Estas nuevas estructuras que representan el significado de la oración se obtienen a partir de la estructura producida por el proceso sintáctico. Para ello, es necesario desarrollar un modelo estructural, de manera que primero se definen las unidades básicas de representación del significado y de qué forma éstas se pueden combinar, para posteriormente construir el significado de proposiciones u oraciones haciendo uso del principio de composicionalidad. Dicho principio establece que el significado de una oración, proposición o cualquier otra estructura sintáctica, se construye a partir del significado de sus constituyentes.

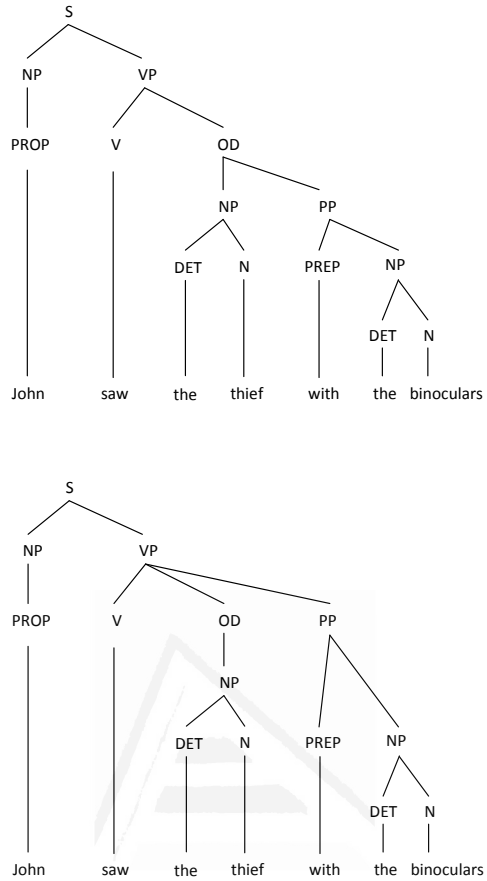


Figura 1.2. Posibles árboles de análisis sintáctico de la oración *John saw the thief with the binoculars.*

Universitat d'Alacant
Universidad de Alicante

Una interpretación superficial de este principio de composicionalidad podría hacer pensar que, dado que las oraciones están formadas por palabras y que éstas son las portadoras primarias del significado del lenguaje, el significado de una oración vendría dado por el significado de las palabras que la forman. Sin embargo, análisis más profundos concluyen que el significado de una oración no se basa solamente en las palabras que lo forman, sino también en el orden, agrupación y relaciones entre palabras de la oración (Jurafsky & Martin, 2000b).

El mayor problema con el que se enfrentan los analizadores semánticos es el hecho de que patrones sintácticos similares pueden introducir diferentes interpretaciones semánticas, y significados similares pueden ser realizados sintácticamente en muchas formas diferentes (Jurafsky & Martin, 2000a). Para tratar con el gran número de casos donde la misma relación sintáctica introduce diferentes relaciones semánticas, es necesario conocer cómo establecer relaciones entre la sintaxis y la semántica (Shi & Mihalcea, 2005). Desde un punto de vista lingüístico, la asignación de roles semánticos a los diferentes argumentos verbales de una oración es una tarea clave a la hora de tratar la interfaz entre la sintaxis y la semántica (Martí & Llisterri, 2002). Este hecho ha dado lugar a que los roles semánticos se hayan constituido en la herramienta utilizada habitualmente en la interpretación semántica.

Un papel o `rol semántico`[3] es la relación entre un constituyente sintáctico (generalmente, aunque no siempre, argumento del verbo) y un predicado (generalmente, aunque no siempre, un verbo). Un rol identifica el papel de un argumento del verbo en el evento que dicho verbo expresa, por ejemplo, un agente, un paciente, un beneficiario, etc., o también adjuntos, como causa, manera o temporal. Dicho de otra manera, un rol semántico es el papel dado por el predicado a sus argumentos.

Consideremos las siguientes oraciones (E8) y (E9):

Universitat d'Alacant
Universidad de Alicante

(E8) [$_{agent}$ John] saw [$_{thing_viewed}$ the thief with the binoculars][4]

(E9) [$_{agent}$ Mary] hit [$_{thing_hit}$ John] [$_{manner}$with a baseball] [$_{temporal}$yesterday] [$_{location}$in the park]

Las palabras de la oración (E9) se agrupan formando cinco constituyentes sintácticos, cada uno de ellos con un rol semántico diferente. El constituyente sintáctico "*Mary*" tiene el rol agente,

[3] También denominados roles temáticos o $\Theta-$roles

[4] Este análisis semántico es obtenido suponiendo el primer árbol de análisis sintáctico mostrado en la figura 1.2

y los constituyentes, "*John*" y "*with a baseball*" tienen los roles paciente e instrumento, respectivamente. Además, los constituyentes "*in the park*" y "*yesterday*" tienen los roles lugar y tiempo, respectivamente.

Es importante destacar, que los posibles roles que pueden jugar los constituyentes sintácticos de una oración varían dependiendo del significado del verbo en esa oración. Considerar las dos oraciones siguientes:

(E10) Mary hit John with a baseball

(E11) Mary hit 300 points

Ambas oraciones hacen uso del verbo *hit*, pero en cada una de ellas el significado del verbo es diferente. En el ejemplo (E10) *hit* tiene sentido #2: *golpear_contra* de WordNet, mientras que en el ejemplo (E11) el sentido de WordNet es #8: *ganar_puntos_en_un_juego*. Como consecuencia, los roles jugados por los argumentos de ambas oraciones son diferentes. En la oración (E10), "*Mary*" tiene el rol de la persona que golpea, "*John*" el rol de la persona golpeada y ""*with a baseball*" el rol del objeto utilizado para golpear. En la oración (E11), "*Mary*" tiene el rol de la persona que gana los puntos y "*300 points*" el de los puntos ganados.

El proceso por el cual se determina el papel que los argumentos de los verbos juegan en una oración, recibe el nombre de anotación de roles semánticos (en inglés, *Semantic Role Labeling* -SRL-). El objetivo en SRL es identificar, para cada uno de los verbos de una oración, todos los constituyentes que juegan algún papel semántico, determinando el rol concreto de cada uno de ellos respecto al verbo. Este proceso se caracteriza por (Dowty, 1991)[5]:

[5] Además de completitud, unicidad y diferenciación, Dowty añade **Independencia**. Según esta característica cada rol tiene una definición semántica que se aplica a todos los verbos en todas las situaciones. De esta manera, estas definiciones no dependen del significado del verbo particular. Sin embargo, como se verá más

- *Completitud*. Todo argumento de un verbo tiene asignado un rol.
- *Unicidad*. A cada argumento de un verbo se le asigna únicamente un rol. Existen unas pocas excepciones para esta característica como muestra la oración (E12). En esta oración "*John*" podría jugar dos papeles diferentes: el rol agente, puesto que inicia el movimiento, o el rol tema, puesto que se trata del objeto que se mueve (Mora, 2001).

  (E12) John ran into the house

- *Diferenciación*. Cada argumento de cada verbo se distingue del resto de argumentos por el rol que tiene asignado. Al igual que en la característica anterior, cabe destacar ciertas excepciones como muestra la oración (E13). En esta oración es difícil determinar cuál de los dos argumentos, "*John*" o "*Mary*", es el que tiene el rol agente (Mora, 2001).

  (E13) John met with Mary

Atendiendo a estas características se puede concluir que, en general, en una oración cada rol semántico es asignado a un único constituyente y cada constituyente juega un único rol. O lo que es lo mismo, dada una oración no puede haber un constituyente que juegue más de un rol, ni dos constituyentes que jueguen el mismo papel semántico.

Una de las consecuencias más beneficiosas de esta conclusión, y que precisamente hace de los roles semánticos una herramienta útil en el análisis semántico es, que aunque cambie el orden de los constituyentes o incluso la voz o el tiempo verbal de la oración, los roles semánticos de los argumentos se mantienen. Por ejemplo, consideremos la oración anterior (E9), si la cambiamos por cualquiera de las oraciones mostradas en los ejemplos (E14) a (E20):

adelante, esta característica únicamente tiene sentido para algunos conjuntos de roles

Universitat d'Alacant
Universidad de Alicante

(E14) [$_{TEMP}$ Yesterday], [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{INSTRUMENT}$ with a baseball] [$_{LOC}$ in the park]

(E15) [$_{PACIENT}$ John] was hit [$_{AGENT}$ by Mary] [$_{TEMP}$ yesterday] [$_{INSTRUMENT}$ with a baseball] [$_{LOC}$ in the park]

(E16) [$_{TEMP}$ Yesterday], [$_{PACIENT}$ John] was hit [$_{INSTRUMENT}$ with a baseball] [$_{AGENT}$ by Mary] [$_{LOC}$ in the park]

(E17) [$_{INSTRUMENT}$ With a baseball], [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{TEMP}$ yesterday] [$_{LOC}$ in the park]

(E18) [$_{TEMP}$ Yesterday] [$_{PACIENT}$ John] was hit [$_{AGENT}$ by Mary] [$_{INSTRUMENT}$ with a baseball] [$_{LOC}$ in the park]

(E19) [$_{LOC}$ In the park], [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{INSTRUMENT}$ with a baseball] [$_{TEMP}$ yesterday]

(E20) [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{INSTRUMENT}$ with a baseball] [$_{LOC}$ in the park] [$_{TEMP}$ yesterday]

o incluso si la traducimos al castellano y alteramos el orden de los constituyentes (E21):

(E21) [$_{TEMP}$ Ayer] [$_{AGENT}$ María] golpeó [$_{PACIENT}$ a Juan] [$_{LOC}$ en el parque] [$_{INSTRUMENT}$ con una pelota de béisbol]

se obtiene que en cualquiera de los casos "*Mary/María*" continúa jugando el rol agente, "*John/Juan*" el rol paciente, "*with a baseball/con una pelota de béisbol*" el rol instrumento, "*in the park/en el parque*" el rol de lugar y "*yesterday/ayer*" el rol temporal.

Universitat d'Alacant
Universidad de Alicante

Todo ello hace de SRL una tarea clave para tareas de PLN que sufran de limitaciones semánticas. Por ejemplo, los sistemas de búsqueda de respuestas, por sus características, requieren información lingüística para afrontar con garantías la tarea de localización de la respuesta correcta. Entre la información lingüística requerida, los roles semánticos juegan un papel fundamental dado que con ellos se puede responder a preguntas como "quién", "cuándo", "dónde", etc. Considerar, por ejemplo, las preguntas (E22 y (E23):

(E22) Who hit John with a baseball yesterday in the park?

(E23) Where did Mary hit John with a baseball yesterday?

un sistema de búsqueda de respuestas que hiciera uso de roles semánticos podría responderlas con cualquiera de las oraciónes (E9), (E14) a (E20). El rol agente,"*Mary*", respondería a la pregunta (E22), mientras que el rol de lugar, "*in the park*", respondería a la pregunta (E23).

- **Análisis pragmático o contextual**. Utiliza la estructura semántica obtenida en el análisis anterior para desarrollar la interpretación final de la oración, en función de las circunstancias del contexto. A este nivel se analizan los mecanismos de coherencia del discurso, es decir, los elementos lingüísticos que el emisor utiliza para comunicar al receptor cuál es su interés discursivo, o que el tema que había iniciado en párrafos anteriores aún continua activo. Estos mecanismos cubre aspectos tales como la identificación de objetos referenciados por determinados constituyentes de la frase (sintagmas nominales, pronombres, elementos elididos,etc.), análisis de aspectos temporales, identificación de la intención del hablante (temas y focos), así como el proceso inferencial requerido para interpretar apropiadamente la oración dentro del dominio de aplicación (Mitkov, 2002; Mitkov *et al.*, 2007).

Universitat d'Alacant
Universidad de Alicante

1.1 Organización de la Tesis

La aportación de este trabajo se centra en el análisis o interpretación semántica, y por tanto en el proceso de anotación de roles semánticos y su aplicación a otras tareas de PLN. Para ello, en primer lugar se realizará un estudio exhaustivo tanto de los diferentes conjuntos de roles semánticos propuestos por diferentes autores, incluyendo una propuesta propia diseñada especialmente para dar soporte a tareas de búsqueda de respuestas; como de los recursos desarrollados hasta el momento que hacen uso de tales conjuntos de roles semánticos, y de las correspondencias que se pueden establecer entre estos recursos (capítulo 2).

A continuación, se analizarán los principales enfoques utilizados por los sistemas automáticos de SRL (capítulo 3); y se presentará información detallada y comparada de estos sistemas atendiendo al corpus que utilizan, al conjunto de roles que determina dicho corpus, a la información proporcionada por los niveles de análisis léxico-morfológica y sintáctica que es utilizada, a la estrategia de etiquetado, al algoritmo de aprendizaje para enfoques basados en corpus, y a los resultados obtenidos (capítulo 4). Esta información se completará con los resultados y principales conclusiones extraídas de las campañas internacionales de evaluación de sistemas de SRL.

También se abordará el desarrollo, evaluación y comparación de un sistema propio de SRL automático, denominado SemRol. SemRol se caracteriza por poseer un fuerte componente de análisis que da lugar a que el proceso de anotación de roles semánticos se realice desde dos perspectivas diferentes y novedosas: clasificación por sentidos vs única, y clasificación global vs individual. Este análisis profundiza en la influencia de la información utilizada en el proceso de anotación de roles semánticos. Como resultado, el estudio determina qué información es útil en el proceso y cuál no (capítulo 5).

En un siguiente paso, la herramienta presentada, SemRol, será utilizada para demostrar la validez de los roles semánticos en sistemas de búsqueda de repuestas (capítulo 6). Con este fin,

en primer lugar se estudiarán las principales características de sistemas similares desarrollados hasta el momento; y posteriormente, se analizarán y evaluarán los resultados de dos novedosos módulos de extracción de respuestas basados en roles semánticos. El primero, un módulo que determina la lista de respuestas candidatas a partir de un conjunto de reglas semánticas, las cuales establecen, dada una pregunta, el tipo de respuesta esperado. El segundo, un módulo que utiliza una base de datos de patrones semánticos previamente generados, para identificar respuestas candidatas.

Para terminar, se presentará un resumen de las principales conclusiones de este trabajo, así como un detalle de las aportaciones más importantes al conocimiento de la investigación en roles semánticos y una lista analizada de las publicaciones más relevantes relacionadas con el trabajo. Finalmente, se comentarán los principales trabajos, tanto en curso como futuros (capítulo 7).

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

2. Roles Semánticos: Estado de la cuestión

Un `rol semántico` es la relación entre un constituyente sintáctico (generalmente, aunque no siempre, argumento del verbo) y un predicado (generalmente, aunque no siempre, un verbo). Ejemplos de roles semánticos son agente, paciente, beneficiario, etc., o también adjuntos, como causa, manera o temporal.

Considerar, por ejemplo, la siguiente oración:

(E24) [$_{agent}$ Mary] hit [$_{thing_hit}$ John] [$_{manner}$with a baseball] [$_{temporal}$yesterday] [$_{location}$in the park]

Las palabras de esta oración se agrupan formando cinco constituyentes sintácticos, cada uno de ellos con un rol diferente. El constituyente sintáctico "*Mary*" tiene el rol agente, y los constituyentes, "*John*" y "*with a baseball*" tienen los roles paciente e instrumento, respectivamente. Además, "*in the park*" tiene el rol de lugar, y el constituyente "*yesterday*" el rol temporal.

A diferencia del nivel sintáctico, donde hay más o menos acuerdo entre la comunidad científica sobre los constituyentes sintácticos y su definición, con los roles semánticos no hay acuerdo alguno sobre qué roles semánticos existen, ni cuáles son las características de cada uno de ellos. En consecuencia, hasta la fecha no ha sido posible definir un conjunto de roles semánticos estándar, aceptado por todos y adecuado para cualquier aplicación. Las causas de esta situación se centran principalmente en (Mora, 2001):

- *Límites*. Cómo y dónde establecer los límites entre tipos de roles dentro de un mismo conjunto. Por ejemplo, considerar los

roles instrumento y tema, y las dos oraciones siguientes (E25) y (E26).

(E25) Load the truck with these rocks

(E26) Load these rocks onto the truck

Se podría considerar que en la oración (E25) "*the truck*" es el tema, es decir, el objeto afectado por el evento, y "*with these rocks*" el instrumento utilizado en el evento; mientras que en la oración (E26) "*onto the truck*" sería el instrumento utilizado y "*these rocks*" el tema.

- *Granularidad.* Existe una total falta de acuerdo respecto a cuántos y cuáles son los roles que se necesitan y con qué nivel de detalle.
- *Organización.* Falta de organización interna, puesto que generalmente el conjunto de roles considerado tiene la forma de lista no estructurada.

Esta situación ha dado lugar a una diversidad de propuestas importante. Esto obliga, a su vez, a que al anotar un corpus con roles semánticos el primer paso sea especificar qué roles se van a anotar y, después, definir las características que describen a cada uno de ellos. Con el objetivo de superar estas limitaciones, el trabajo aquí presentado propone un nuevo conjunto de roles. Dicho conjunto ha sido desarrollado atendiendo a principios de aplicabilidad, generalidad, jerarquía y conexión con otras propuestas de anotación.

El apartado 2.1 muestra dicha propuesta, junto con una recopilación de las principales propuestas de conjuntos de roles semánticos realizadas hasta el momento. Además, la diversidad de recursos lingüísticos que estos conjuntos han generado se presentan en el apartado 2.2. Finalmente, las relaciones definidas entre los recursos lingüísticos con el objetivo de conseguir independencia respecto al recurso utilizado se resumen en el apartado 2.3.

Universitat d'Alacant
Universidad de Alicante

2.1 Análisis de propuestas de conjuntos de roles semánticos

Los roles semánticos son una de las clases de construcciones más antiguas de la teoría lingüística. Sin embargo, hasta la fecha los lingüistas no han alcanzado un consenso acerca del inventor exacto de los roles semánticos ni acerca de su naturaleza o su situación en la teoría lingüística. Mientras que para algunos la primera mención a los roles data de miles de años atrás con la teoría de Panini y sus *karakas*[1] (Kiparsky, 2002), para otros, no fue hasta los años sesenta, cuando Jeffrey Gruber y Charles Fillmore enumeraron las primeras listas de roles, proporcionando un estudio detallado sobre observaciones sintácticas y semánticas conectadas con ellos.

Longevidad engendra variedad. Por ello, podemos encontrar una gran diversidad de propuestas de conjuntos de roles semánticos. El espectro de tales propuestas varía, desde conjuntos muy específicos, dependientes del dominio o del verbo, a conjuntos muy generales. Entre medias, toda una variedad de teorías con una media aproximada de 10 roles. Incluso varía el origen de las propuestas, si bien, se puede establecer como norma, que los conjuntos de roles más abstractos han sido propuestos por lingüistas mientras que los más específicos han sido propuestos por ingenieros (Gildea & Jurafsky, 2002).

De entre todas las propuestas realizadas, a continuación se presentan las más destacadas en orden cronológico. Un resumen de las características más importantes de tales propuestas, se puede ver en el cuadro 2.1. En concreto, el cuadro muestra si el conjunto de roles es de dominio general o no (columna dominio general), si es un conjunto único o varía, por ejemplo, para cada verbo (columna conjunto universal), si tiene una organización jerárquica o no (columna jerarquía), y si es específico de alguna lengua (columna lengua general).

[1] Concepto de la teoría de Panini similar al concepto de rol temático http://en.wikipedia.org/wiki/Karaka Consultado en marzo 2008

[2] Idioma hablado en Sudáfrica.

Universitat d'Alacant
Universidad de Alicante

Década	Propuesta	Dominio General	Conjunto Universal	Jerarquía	Lengua General
4th A.C.	Panini	Si	Si	No	Sánscrito
60	(Gruber, 1965)	No	Si	No	Si
	(Fillmore, 1968)	Si	Si	No	Si
	(Sgall *et al.*, 1986)	Si	Si	No	Si
70	(Celce-Murcia, 1972)	Si	Si	No	Si
	(Schank, 1972)	Si	Si	No	Si
	(Contreras, 1976)	Si	Si	No	Si
80	(R.D. Van Valin, 2005)	No	Si	Si	Si
	(Sowa, 1984)	Si	Si	No	Si
	(Pollard & Sag, 1988)	Si	Por verbo	No	Si
	(Machobane, 1989)	Si	Si	Si	Sesotho[2]
90	(Jackendoff, 1990)	Si	Si	No	Si
	(Dowty, 1991)	Si	Si	No	Si
	(Grimshaw, 1990)	Si	Si	Si	Si
	(Chierchia & McConell-Ginet, 1990)	Si	Si	No	Si
	(Brown & Miller, 1991)	Si	Si	No	Si
	(Frawley, 1992)	Si	Si	Si	Si
	(Palmer, 1994)	Si	Si	No	Si
	(Haegeman, 1991)	Si	Si	No	Si
	(González, 1997)	Si	Si	No	Si
	(Wechsler, 1995)	Si	Por verbo	No	Si
	(Guitar, 1998)	Si	Si	No	Español
	CyC Upper	Si	Si	No	Si
	MUC	No	No	No	Inglés
	P.Treebank II	Si	Si	No	Si
	(Gomez, 1998)	Si	Si	Si	Si
00	FrameNet	Si	Por marco	No	Si
	PropBank	Si	Por sentido	No	Si
	(Stallard, 2000)	No	Si	No	Si
	(Busser & Moens, 2003)	Si	Si	No	Si
	(Girju *et al.*, 2004)	No	Si	No	Si
	(Bethard *et al.*, 2004)	No	Si	No	Si
	VerbNet	Si	Por clase	No	Si
	(Moreda *et al.*, 2007)	No	Si	Si	Si

Cuadro 2.1. Resumen de las principales propuestas de conjuntos de roles

Universitat d'Alacant
Universidad de Alicante

2.1.1 Propuesta de Gruber

Gruber (1965) propone un conjunto de roles específico para el dominio de la localización espacial y el movimiento (Kailuweit, 2006). El cuadro 2.2 muestra el detalle de este conjunto de roles y una breve descripción para cada uno de ellos.

Rol	Descripción
Theme	Objeto en movimiento o que está siendo localizado
Agent	Instigador de una acción o estado
Location	Lugar
Source	Objeto desde el cual se produce el movimiento
Path	Camino
Goal	Objeto hacia el cual se dirige el movimiento

Cuadro 2.2. Detalle del conjunto de roles propuesto en (Gruber, 1965)

2.1.2 Propuesta de Fillmore. Gramática de casos

Fillmore (1968) desarrolló la teoría denominada de gramáticas de caso (en inglés, *case grammar*). Según esta teoría, la oración, en su estructura básica, consta de un verbo y de un conjunto de casos (en inglés, *deep case*) o roles semánticos, los cuales establecen una relación entre el verbo y los sintagmas nominales de la oración, de forma que cada una de esas relaciones sólo ocurre una vez en una oración simple (Wasow, 2003). Cada verbo selecciona un determinado número de casos, dando lugar a su marco de caso (en inglés, *case frame*).

Su objetivo fue establecer un conjunto de roles homogéneo y de propósito general. Sin embargo, modificó sus listas varias veces sin llegar a definir un conjunto definitivo (Kailuweit, 2006). En Fillmore (1968) identificó seis roles, cuyo detalle y descripciones podemos ver en el cuadro 2.3. En Fillmore (1969) identificó siete

Rol	Descripción
Agent	Instigador de la acción identificada por el verbo
Instrument	Objeto o fuerza inanimada envuelto casualmente en la acción o estado identificado por el verbo
Dative	Objeto animado afectado por el estado o la acción identificada por el verbo
Factitive	Objeto que es resultante de la acción o estado identificado por el verbo o que es entendido como parte del significado del verbo
Locative	Posición u orientación espacial del estado o acción identificada por el verbo
Object	Cualquier cosa representable por un nombre, cuyo rol en la acción o estado identificado por el verbo es identificado por la interpretación semántica del verbo en sí mismo

Cuadro 2.3. Primer conjunto de roles propuesto por Fillmore (1968)

roles, cinco de ellos (*agent*, *object*, *result/factitive*, *instrument* y *experiencer/dative*) comunes a la lista anterior (Wasow, 2003). Una recopilación de los roles semánticos de todas sus propuestas se puede ver en el cuadro 2.4.

Rol	Descripción
Agent	El causante de un evento
Experiencer	El que experimenta un evento
Force	El causante involuntario de un evento
Theme	El participante en un evento afectado por el mismo de forma más directa
Result	El producto final de un evento
Content	La proposición o contenido de un evento proposicional
Instrument	El instrumento utilizado en un evento
Beneficiary	El beneficiario de un evento
Source	El origen del objeto en un evento de traslado
Goal	El destino de un objeto en un evento de traslado

Cuadro 2.4. Recopilación de roles temáticos propuestos por Fillmore en sus diferentes trabajos

Universitat d'Alacant
Universidad de Alicante

2.1.3 Propuesta de Sgall et al. Descripción Generativa Funcional

La teoría de Descripción Generativa Funcional (en inglés, *Functional Generative Description* -FDG-), desarrollada por Petr Sgall y sus colaboradores en Praga desde los años 60 (Sgall *et al.*, 1986), consiste en analizar las oraciones en base a dependencias. En FDG se trabaja con la representación tectogramática de las oraciones (Sgall, 2001).

Una representación tectogramática de una oración básicamente tiene forma de árbol de dependencias. De esta manera, a cada oración se le asigna una estructura de árbol con nodos y arcos etiquetados. Los nodos, que representan a las palabras de la oración con significado semántico, tiene asignado un marco de valencia (en inglés, *valency frame*) el cual incluye información sobre sus valores morfológico y léxico. Los arcos en el árbol denotan las relaciones de dependencia, denominadas *functors*, entre las palabras de la oración.

Hay dos tipos de relaciones de dependencia:

- **Participantes internos** (en inglés, *inner participants*) o argumentos, los cuales pueden ser obligatorios u opcionales. A su vez, se clasifican en:
  - Sintácticos: ACT(or), siempre el primer participante, PAT(tient), el segundo.
  - Semánticos: ADDR(essee), EFF(ect), ORIG(in);
- **Modificaciones libres** (en inglés, *free modifications*) o adjuntos, como *location*, *time*, *manner* o *intention* (Baker *et al.*, 2004). Son opcionales.

Información más detallada sobre las posibles relaciones de dependencia se muestra en el cuadro 2.5 (Hajič, 2004).

Aunque la posición central en una oración la ocupa, normalmente, un verbo, esta representación también incluye nombres y adjetivos.

Universitat d'Alacant
Universidad de Alicante

Tipo de relación	Descripción
Participantes internos	ACT - Actor
	PAT - Paciente
	ADDR - Dirección
	ORIG - Origen
	EFF - Efecto
Time	TWHEN - Cuándo
	TTILL - Hasta cuándo
	TSIN - Desde cuándo
	TFHL - Durante cuánto
	TFRWH - Desde cuándo
	TOWH - Hasta cuándo
	TPAR - Eventos paralelos
	THO - Cuántas veces
Location	LOC - Lugar
	DIR1 - Desde dónde
	DIR2 - Por dónde
	DIR3 - Hasta dónde
Manner	MANN - Manera
	MEANS - Medio de alcanzar algo
	RESL - Resultado
	REG - De acuerdo a
	CRIT - Criterio o norma
	EXT - Extensión
	ACMP - Acompañamiento
	DIFF - Diferencia
	CPR - Comparación
Implication	CAUS - Causa
	COND - Condición
	AIM - Objetivo
	INTT - Intención
Other	BEN - Benefactor
	SUBS - Sustitución
	HER - Herencia
	CONTRD - Contradicción
	RSTR - Atributo general
	AUTH - Autoría
	APP - Accesorio
	MAT - Material
	ID - Identidad
	COMPL - Complemento

Cuadro 2.5. Tipos de relaciones de dependencia en FDG (Hajič, 2004)

Universitat d'Alacant
Universidad de Alicante

2.1.4 Propuesta de Celce-Murcia

Como continuación a la propuesta de la teoría de la gramática de casos de Fillmore (ver apartado 2.1.2), Celce-Murcia (Celce-Murcia, 1972; Celce-Murcia, 1976) propone que todos los argumentos de cualquier verbo pueden ser clasificados como miembros de cinco relaciones de caso. Un detalle de dicho conjunto de relaciones puede verse en la tabla 2.6.

Rol	Descripción
Causal Actant	El causante de la acción
Theme	El participante en un evento afectado por el mismo
Locus	Lugar
Source	Origen
Goal	Destino

Cuadro 2.6. Detalle de los roles temáticos propuestos por (Celce-Murcia, 1972)

2.1.5 Propuesta de Schank. Teoría de la dependencia conceptual

La propuesta de Schank (Schank, 1972), denominada teoría de la dependencia conceptual (en inglés, *conceptual dependency*), es un modo de representar la información en el nivel conceptual según el cual las relaciones entre conceptos son dependencias.

De la misma manera que a nivel léxico las palabras se unen formando oraciones, según Schank, a nivel conceptual, los conceptos se unen formando conceptualizaciones. Una conceptualización consta de un actor, una acción y un conjunto específico de casos conceptuales. Los posibles casos conceptuales son: *objective*, *directive*, *instrumental* y *recipient*. Ver cuadro 2.7.

Caso	Descripción
Objective	Objeto que sufre la acción
Directive	Dirección o localización de la acción
Instrumental	Lo utilizado para llevar a cabo la acción
Recipient	El que recibe un objeto como resultado de la acción

Cuadro 2.7. Casos conceptuales propuestos por (Schank, 1972)

2.1.6 Propuesta de Folley y Van Valin. Macropapeles de la gramática del rol y la referencia

En la teoría de de la gramática del rol y la referencia (en inglés, *Role and Reference Grammar* -RRG-), desarrollada en los años 80, se proponen dos grupos de roles semánticos (R.D. Van Valin, 2005):

- Las **Relaciones temáticas específicas** corresponden a roles semánticos como los propuestos por Gruber y Fillmore (ver apartados 2.1.1 y 2.1.2, respectivamente), tales como agente, tema, posición etc. Todas las relaciones temáticas se definen en términos de posiciones de argumentos para verbos de estado y actividad. El detalle de estas relaciones puede verse en la figura 2.1.

  Van Valin destaca, que si bien podría dar la impresión de que RRG propone una gran cantidad de relaciones temáticas, sin embargo, sólo hay cinco distinciones relevantes que corresponden a las cinco posibles posiciones de los argumentos.

  Además, en realidad, en RRG las etiquetas correspondientes a roles semánticos tradicionales se mantienen como meras etiquetas para las posiciones en un continuo semántico constituido por las posiciones argumentales de los predicados de actividad y estado, con agente en un extremo y paciente en el otro.

  Con la excepción de agente, cada una de las relaciones temáticas listada bajo una posición de argumento particular representa una subclase distinta de verbo de estado o actividad. Por ejemplo, la relación temática *stimulus*, representa al segundo argumento de un predicado de estado de dos argumentos.

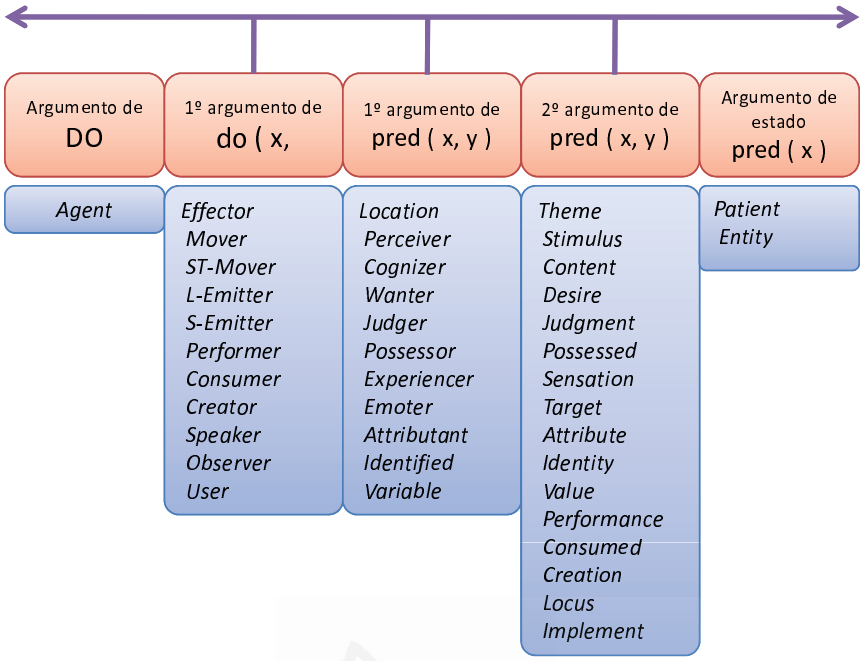


Figura 2.1. Continuo de relaciones temáticas en RRG

- Roles semánticos generalizados, también denominados **macroroles semánticos**. Se definen dos macroroles, actor y *undergoer* o afectado. Se denominan macroroles porque cada uno de ellos incluye un número de relaciones temáticas específicas. Las relaciones temáticas en la columna de la izquierda funcionan como actor, y las relaciones en la columna de la derecha como afectado. Por tanto, actor y afectado son, en realidad, generalizaciones a través de las relaciones temáticas en cada columna. Por ejemplo, actor es una generalización de agente, instrumento y otros roles; y afectado es generalización de paciente, tema, recipiente y otros roles.

  La relación jerárquica entre las relaciones temáticas y los macroroles se muestra en la figura 2.2.

Lo que esta jerarquía muestra es que dado un verbo, el argumento más a la izquierda será el actor y el más a la derecha

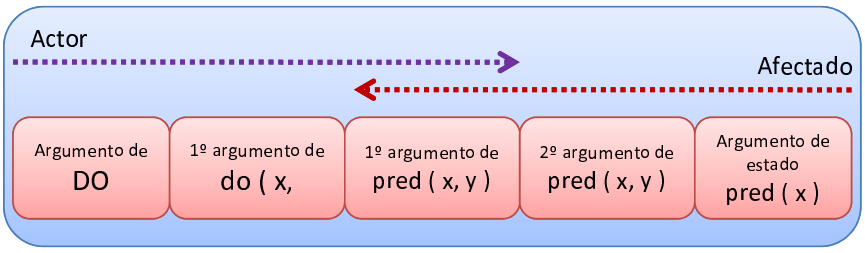


Figura 2.2. Jerarquía actor-afectado en RRG

el afectado. El actor se define como el argumento que expresa el participante que realiza, efectúa, instiga o controla la situación denotada por el predicado. El afectado, como el argumento que expresa el participante que es fuertemente afectado por los participantes en algún modo (Folley & Valin, 1984).

Si el predicado es transitivo se le asigna el macropapel actor al argumento más activo en la jerarquía actor-afectado y el macropapel afectado al argumento más pasivo. Las actividades intransitivas sólo constan del macropapel actor para el argumento más activo y los estados intransitivos del macropapel afectado para el argumento más pasivo (Valin & Polla, 1997).

2.1.7 Propuesta de Jackendoff

Jackendoff (1990) continuó las observaciones y conclusiones de Gruber (ver apartado 2.1.1) bajo la idea de que el conjunto de roles utilizado por Gruber para localización espacial y movimiento podía ser generalizado a muchos otros campos semánticos, dado que muchos verbos y preposiciones aparecen en dos o más campos semánticos. Como consecuencia de estos trabajos propuso su propio conjunto de roles el cual se muestra en el cuadro 2.8.

Trabajos posteriores hicieron que Jackendoff refinara y modificara este conjunto inicial. Estos trabajos dieron lugar a la que se ha denominado teoría de la semántica conceptual (en inglés, *conceptual semantics*). Según dicha teoría el significado de una expresión lingüística se representa mediante una estructura con-

Rol	Descripción
Theme	Objeto en movimiento o que está siendo localizado
Source	Objeto desde el cual se produce el movimiento
Target	Objeto hacia el cual se dirige el movimiento
Agent	Instigador de un estado o acción

Cuadro 2.8. Versión inicial del conjunto de roles propuesto por (Jackendoff, 1990)

ceptual, la cual está formada por constituyentes conceptuales. Cada uno de estos constituyentes conceptuales comprende una o más primitivas o funciones semánticas, como por ejemplo, *GO* (para eventos), *BE* (para estados), *CAUSE*, *TO*, *FROM*, etc. En este marco, los roles temáticos son relaciones estructurales dentro de estructuras conceptuales. Como tales, Jackendoff redefine los roles anteriores tal y como se muestra en el cuadro 2.9.

Nivel	Rol	Descripción
Temático	*Theme*	El primer argumento de cualquiera de las funciones de posición o movimiento
	Source	El argumento de *FROM*
	Target	El argumento de *TO*
	Agent	El primer argumento de *CAUSE*
De la acción	*Actor*	El primer argumento de la función de afecto
	Patient/Beneficiary	El segundo argumento de la función de afecto

Cuadro 2.9. Conjunto de roles refinado propuesto por (Jackendoff, 1990)

Universitat d'Alacant
Universidad de Alicante

2.1.8 Propuesta de Dowty

Con el objetivo de solucionar los problemas de organización interna y de determinación de límites que presentan algunos conjuntos de roles, tal y como se comentó al principio del capítulo, Dowty (1991), partiendo de los trabajos realizados por Folley y Van Valin (1984) (ver sección 2.1.6), desarrolla una aproximación basada en dos tipos de roles prototípicos que él denomina proto-agente y proto-paciente. Cada uno de estos protoroles se carac-

teriza por un conjunto de propiedades (ver cuadros 2.10 y 2.11, respectivamente). Un argumento de un verbo será proto-agente o proto-paciente dependiendo del número de propiedades de agente o paciente que cumpla.

PROTO-AGENTE
Supone voluntad en el evento o estado
Causa un evento o cambia el estado de otro participante
Movimiento (relativo a la posición de otro participante)
Existe independientemente del evento denotado por el verbo
Sentience (y/o *perception*)

Cuadro 2.10. Propiedades de un proto-agente según Dowty (1991)

PROTO-PACIENTE
Experimenta cambio de estado
Causalmente afectado por otro participante
Parado respecto al movimiento de otro participante
No existe independientemente del evento
Incremental theme

Cuadro 2.11. Propiedades de un proto-paciente según Dowty (1991)

2.1.9 Propuesta del proyecto FrameNet

El proyecto FrameNet (Fillmore, 2002), del cual se hablará en detalle en la sección 2.2.2, propone roles, denominados elementos de marco o de *frame*, ni tan específicos como los miles de roles potenciales específicos para cada verbo, ni tan generales como las propuestas de conjuntos de 10 roles. En FrameNet consideran que los roles de propósito general no cubren todas las necesidades para los marcos o *frames* semánticos, y por ello definen nombres de roles específicos para cada marco. Como ejemplo (Johnson *et al.*, 2002), el cuadro 2.12 muestra los roles o elementos de *frame*, y sus descripciones, identificados para el *frame* semántico de la comunicación verbal. Un resumen de algunos *frames*, sus elementos de *frame* y sus relaciones se puede ver en Fillmore y Baker (2001).

Los *frames* o marcos semánticos son representaciones esquematizadas de situaciones del mundo real, en base a los cuales se organiza la información. Un *frame* incluye un conjunto de unidades léxicas, una lista de roles o elementos de *frame* y un conjunto de ejemplos.

Las unidades léxicas se definen como los pares palabra-sentido que más frecuentemente evocan el marco semántico al cual pertenecen. Varias unidades léxicas, pueden evocar un mismo marco semántico y por tanto, compartir roles. Por ejemplo, los verbos *to give* y *to receive* en las oraciones de los ejemplos (E27) y (E28) evocan el marco semántico *transaction* y por tanto, comparten los roles agente, paciente y receptor.

(E27) [$_{agente}$ John] gave [$_{receptor}$ Mary] [$_{paciente}$ the book]

(E28) [$_{receptor}$ Mary] received [$_{paciente}$ the book] from [$_{agente}$ John]

Por otro lado, diferentes sentidos de una misma palabra pueden pertenecer a *frames* diferentes. Por ejemplo el verbo *to argue*, pertenece a los *frames Quarreling* y *Reasoning* (Lopatková, 2003).

A diferencia de otros conjuntos de roles semánticos considerados por el resto de investigadores según los cuales los roles semánticos suelen ser argumentos de verbos, los elementos de *frame* pueden ser argumento de cualquier predicado, incluyendo verbos, nombres y adjetivos.

Ciertos roles son considerados como elementos de *frame* núcleo para un marco en particular, en el sentido de que ellos siempre están presentes conceptualmente. Frente a estos se encuentran los que no están siempre expresados en cada ocurrencia de un predicado que evoque el *frame*. Sería el caso de roles como tiempo o lugar.

Además de los roles, cada marco semántico incluye ejemplos, anotados a mano, para los diferentes elementos de *frame*. (Ver cuadro 2.12).

Universitat d'Alacant
Universidad de Alicante

Rol	Descripción
Speaker	Persona que realiza el acto de comunicación verbal
Addressee	Destinatario de un mensaje verbal
Message	Propósito comunicado
Topic	Asunto de un mensaje
Medium	Canal físico de comunicación
Code	Lenguaje u otro código utilizado para comunicar
[Others (Speaker, NP, Ext)] assert [that anthropology is the tree and sociology the brach (Message, Sfin, Somp)]	

Cuadro 2.12. Conjunto de roles en FrameNet para el marco semántico de la comunicación verbal

Para más información sobre el proyecto consultar el apartado 2.2.2.

2.1.10 Propuesta del proyecto PropBank

En el proyecto Proposition Bank (PropBank) (Palmer *et al.*, 2005), del cual se hablará en detalle en la sección 2.2.1, el conjunto de roles correspondiente a un uso de un verbo se denomina *roleset*. Dicho conjunto está asociado a un conjunto de *frames* o marcos sintácticos, dando lugar a un denominado *frameset*. El criterio para distinguir *framesets* se basa en semántica, de manera que dos significados de un verbo se sitúan en *framesets* diferentes si toman diferente número de argumentos. En consecuencia, un verbo polisémico puede tener más de un *frameset* cuando las diferencias en significado son suficientemente distintas como para requerir un conjunto de roles diferentes, uno por cada *frameset*. Un ejemplo se muestra en el cuadro 2.13.

Según Palmer *et al.*, (2005), dada la dificultad de definir un conjunto universal de roles semánticos o temáticos que cubran todos los tipos de predicados, en PropBank, los argumentos semánticos de un verbo son numerados, comenzando por 0 y hasta 5, expresando la proximidad semántica respecto al verbo. El uso de argumentos numerados se debe a que están a mitad de camino entre muchos puntos de vista teóricos diferentes. Por otra parte, los propios autores destacan que tales argumentos numerados

Frameset		*Frameset*	
DECLINE.01: descender gradualmente		DECLINE.02: rechazar	
ROL	DESCRIPCIÓN	ROL	DESCRIPCIÓN
Arg1	Entidad que desciende	Arg0	Agente
Arg2	Cantidad que desciende	Arg1	Cosa rechazada
Arg3	Punto de partida		
Arg4	Punto de llegada		

Cuadro 2.13. Ejemplo de dos conjuntos de roles del verbo *decline* en PropBank

pueden ser mapeados fácil y consistentemente a cualquier teoría de estructura de argumentos.

No se ha intentado que las etiquetas de los argumentos tengan el mismo significado, de un sentido de un verbo, a otro. Por ejemplo, el rol jugado por arg2 en un sentido de un predicado dado, puede ser jugado por arg3 en otro sentido.

Como mucho pueden aparecer 6 argumentos numerados, de 0 a 5, aunque la mayoría de los *rolesets* tienen de dos a cuatro roles numerados. Aunque muchos lingüistas considerarían cualquier argumento mayor que arg2 o arg3 como adjuntos, en PropBank consideran que éstos ocurren con suficiente frecuencia con sus verbos respectivos, o clases de verbos, como para que se les asigne un número. De esta manera, además, persiguen asegurar una anotación consistente.

Para un verbo en particular, arg0 es generalmente el argumento que muestra las características de un proto-agente de los de Dowty (ver apartado 2.1.8), mientras que arg1 es un proto-paciente o tema. Como muestra el cuadro 2.14, para argumentos de número mayor no se pueden hacer generalizaciones (Baker *et al.*, 2004), aunque se hizo un esfuerzo por definir roles de forma consistente a través de los miembros de las clases de VerbNet (Kipper, 2005) (más información sobre VerbNet en el apartado 2.2.3).

Existe un rol específico especial etiquetado como argA. Dicha etiqueta es utilizada para capturar el agente de una acción inducida que ocurre con los verbos volitivos de movimiento. Tal es el caso de la oración (E29).

Universitat d'Alacant
Universidad de Alicante

Rol	Tendencias
Arg0	Agente
Arg1	Objeto directo/tema/paciente
Arg2	Objeto indirecto/beneficiario/instrumento/atributo/estado final/extensión
Arg3	Punto de partida, origen/beneficiario/instrumento/atributo
Arg4	Punto de llegada, destino

Cuadro 2.14. Tendencias de los argumentos numerados de PropBank

(E29) Mary volunteered John to clean the garage

Además de los roles numerados específicos de cada verbo, los verbos pueden tomar cualquiera del conjunto de roles generales o adjuntos definidos en PropBank. Se etiquetan como argM, más una etiqueta de función. Una lista detallada de los mismos puede verse en el cuadro 2.15.

Rol	Descripción
LOC	Lugar
EXT	Extensión (argumento numérico)
DIS	Conectiva del discurso
ADV	Propósito general
NEG	Marca de negación
MOD	Verbo modal
CAU	Causa
TMP	Tiempo
PNC	Propósito
MNR	Modo
DIR	Dirección
PRD	Predicación secundaria (indica que existe relación entre los argumentos, o lo que es lo mismo, que el argumento en cuestión actúa como un predicado para algún otro argumento de la oración. Ej.: *Mary called John an idiot*, relación entre "Jonh" y "an idiot")

Cuadro 2.15. Lista de etiquetas de función de adjuntos en PropBank

Universitat d'Alacant
Universidad de Alicante

2.1.11 Propuesta de de roles semánticos para sistemas de BR

Con el objetivo de desarrollar un recurso útil para tareas de PLN, en concreto, tareas de búsqueda de respuestas, en este trabajo se propone un conjunto de roles semánticos organizados jerárquicamente que responda a posibles entidades semánticas por las que se puede preguntar en una consulta a partir del verbo (Moreda *et al.*, 2007). Ver figura 2.3.

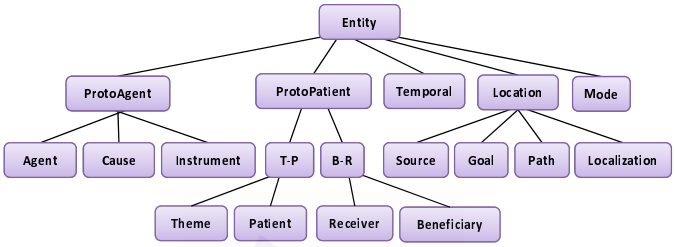


Figura 2.3. Conjunto de roles propuesto por Moreda et al.

Los sistemas de búsqueda de respuestas, por sus características, requieren información lingüística para afrontar con garantías la tarea de localización de la respuesta correcta. Entre la información lingüística requerida, los roles semánticos juegan un papel fundamental. Con la información que los roles proporcionan se podría responder a preguntas como "quién", "cuándo", "dónde" o "qué".

Considerar, por ejemplo, las preguntas (E30 y (E31):

(E30) Who hit John with a baseball yesterday in the park?

(E31) Where did Mary hit John with a baseball yesterday?

un sistema de búsqueda de respuestas que hiciera uso de roles semánticos podría responderlas con la oración (E32). El rol agen-

te,"*Mary*", respondería a la pregunta (E30), mientras que el rol de lugar, "*in the park*", respondería a la pregunta (E31).

(E32) [$_{agent}$ Mary] hit [$_{thing_hit}$ John] [$_{manner}$with a baseball] [$_{temporal}$yesterday] [$_{location}$in the park]

Los principios seguidos a la hora de establecer estos roles son los siguientes (Navarro *et al.*, 2004):

- *Principio de aplicabilidad.* El objetivo no es demostrar ni justificar ninguna teoría concreta sobre roles semánticos, sino desarrollar un recurso útil para tareas de PLN. Por ello, no se pretende definir unos roles semánticos universales, sino establecer un conjunto de roles semánticos que tenga una aplicación clara a búsqueda de respuestas. En consecuencia, estos roles semánticos responderán a posibles entidades semánticas por las que se puede preguntar en una consulta a partir del verbo.
- *Principio de generalidad.* La lista de roles definidos son roles generales, aplicables a diferentes verbos que compartan rasgos semánticos similares, es decir, a toda una clase verbal.
- *Principio de conexión* con otras propuestas de anotación. Proponer un nuevo conjunto de roles semánticos no serviría de nada si los roles propuestos no se relacionaran con los roles de otras propuestas similares. Así, la lista de roles propuesta está basada en los roles generales de PropBank (ver apartado 2.1.10) y VerbNet (ver apartado 2.2.3) y tiene en cuenta los utilizados en FrameNet (ver apartado 2.1.9). De esta manera, el conjunto de roles resultante quedará relacionado con conjuntos de roles similares.
- *Principio de jerarquía.* Teniendo en cuenta la propuesta de Dowty (1991) (ver apartado 2.1.8), se considera que es posible establecer una jerarquía de roles semánticos. Esto hace al conjunto de roles más consistente dado que no es una simple lista de roles que puede asumir un argumento verbal, sino que, según el contexto, puede ser semánticamente más generales o más específicos. En consecuencia, si un argumento pudiera etiquetarse

con dos roles semánticos, es posible resolver la ambiguedad con el rol de nivel superior que incluya ambos roles.

Esta jerarquía se define (ver figura 2.3):

- El nivel más **general** es aquél que no tiene ninguna información semántica. El rol `entity` sólo indica la presencia de un argumento.
- En un **primer nivel** de concreción semántica están los roles de carácter universal, como `tiempo`, `lugar` o `modo`, junto al conjunto de roles relacionados con el agente, `proto-agente`, y el conjunto de roles relacionados con el paciente, `proto-paciente`.
- En un **segundo nivel** se sitúan los roles específicos de cada uno de estos:
  - Roles que suelen actuar como **argumentos**. Estos roles se pueden relacionar, a grandes rasgos, con las funciones sintácticas de sujeto, objeto directo y objeto indirecto de las oraciones transitivas, respectivamente:
    - `Agente-Causa`. Argumento que denota la entidad que desde un punto de vista general produce la acción o evento (o es la principal entidad del estado) expresado en el verbo.

      En general, este rol responde a la pregunta "¿quién?". Si tiene el rasgo [+animado] se considera agente, y si tiene el rasgo [-animado] se considera causa. Relacionados con estos roles está también el rol instrumento.
    - `Tema-Paciente`. Argumento que denota la entidad directamente afectada por el verbo.

      Suele responder a la pregunta "¿qué?". Si tiene el rasgo [+animado] se considera paciente, y si tiene el rasgo [-animado] se considera tema.
    - `Beneficiario-Receptor`. Argumento que denota la entidad que resulta beneficiada o afectada indirectamente por el verbo.

Universitat d'Alacant
Universidad de Alicante

Responde a preguntas tipo "¿a/para qué/quién?".

- Roles que suelen aparecer como **adjuntos** (si bien hay determinados verbos que los exigen como argumentos):
  - `Tiempo`. Sólo se anota si aparece un sintagma que especifique de manera explícita el tiempo en el que la acción/estado del verbo se desarrolla. Responde a la pregunta "¿cuándo?".
  - `Lugar`. Pueden hacer referencia tanto a lugares físicos como a lugares abstractos. Responde a la pregunta "¿dónde?".
  - `Modo`. Complemento similar a los anteriores que indica el modo o manera en que se lleva a cabo la acción, evento o estado del verbo. Responde a la pregunta "¿cómo?".

- En algunos casos existe un **tercer nivel**, donde aún se especifican más subroles. En concreto, el rol de lugar puede especificar, a su vez, tres sub-roles: `origen` (lugar "desde donde"), `meta` (lugar "a donde") y `trayectoria` (lugar "por donde").

2.1.12 Otras propuestas

Hasta el momento se han presentado las propuestas de conjuntos de roles semánticos más significativas. Sin embargo, la cantidad de propuestas existentes es muchísimo más amplia. Por ello, un resumen de algunas de las más interesantes se muestra en los cuadros 2.16 y 2.17.

Además, el cuadro 2.1 incluye información esquemática sobre tales propuestas. En concreto, se muestra si el conjunto de roles es de dominio general o no (columna dominio general), si es un conjunto único o varía, por ejemplo, para cada verbo (columna conjunto universal), si tiene una organización jerárquica o no (columna jerarquía), y si es específico de alguna lengua (columna lengua general).

Autor	Propuesta de Roles
Panini	*Agent*, *Goal*, *Recipient*, *Instrument*, *Locative* y *Source*
(Contreras, 1976)	*Agent*, *patient*, *instrument*, *beneficiary*, *experiencer*, *possessor*, *cause*, *complement*, *source*, *target*, *location*, *time*, *identifier*
(Sowa, 1984)	37 relaciones conceptuales como *agent*, *cause*, *destination* o *instrument*. Un detalle de las mismas se muestra en http://www.cs.nmsu.edu/ tomohara/thematic-roles/sowa-conceptual-relations.html, consultado en abril 2008
(Machobane, 1989)	*causer*, *agent*, *benefactive*, *experiencer*, *goal* (animado), *theme*, *goal* (inanimado), *locative*, *instrument*
(Wechsler, 1995; Pollard & Sag, 1988)	Dentro de la teoría *Head-Driven Phrase Structure Grammar* (HPSG)(consultar http://www.ling.ohio-state.edu/research/hpsg/, consultado en marzo 2008) algunos lingüistas, con el objetivo de superar el problema de definir un conjunto de roles y describirlos, empezaron a referirse a los roles como dador, dado, etc.
(Grimshaw, 1990)	Conjunto de roles organizados jerárquicamente: *agent*, *experiencer*, *goal*/*source*/*location* y *theme*. En esta jerarquía el argumento más alto en el árbol de análisis es siempre el *agent*, el siguiente el *experiencer*, y así sucesivamente
(Chierchia & McConell-Ginet, 1990)	*Agent*, *theme*, *experiencer*
(Brown & Miller, 1991)	*Agent*, *patient*, *instrument*, *benefactive*, *manner*, *place*, *location*, *range*, *result*, *dative*, *goal*, *source*, *path*, *attribute*, *neutral*
(Frawley, 1992)	Jerarquía de cuatro tipos de roles: *logical actors*, (*agent*, *author*, e *instrument*); *logical recipients* (*patient*, *experiencer* y *benefactive*), *spatial roles* (*theme*, *source*, y *goal*) y *non-participant roles* (*locative*, *reason*, y *purpose*)
(Palmer, 1994)	*Agent*, *patient*, *beneficiary*, *instrumental*, *locative*
(Haegeman, 1991)	*Agent*/*actor*, *patient*, *theme*, *benefactive*/*beneficiary*, *goal*, *source*, *location*
Penn TreeBank II	*beneficiary*, *direction*, *spatial extent*, *manner*, *location*, *purpose* / *reason* y *temporal*

Cuadro 2.16. Resumen de otras propuestas de roles semánticos (1/2)

Universitat d'Alacant
Universidad de Alicante

Autor	Propuesta de Roles
(González, 1997)	Verbador, verbado y verbatario
Cyc Upper Ontology	La ontología *Cyc Upper Ontology* define un conjunto de 130 relaciones temáticas, entre las que se encuentran *actors*, *beneficiary*, *buyer* o *fromLocation*. Un detalle de las mismas se muestra en http://www.cs.nmsu.edu/ tomohara/thematic-roles/cyc-thematic-roles.html, consultado en abril 2008
(Guitar, 1998)	causa, tema, locus
MUC	Conjunto de roles utilizado para el dominio terrorista en los corpus del DARPA *Message Understanding Conferences* (MUC): *perpetrator*, *target*, *victim*, *location*, *instrument* y *date* (Riloff & Schmelzenbach, 1998)
(Gomez, 1998)	Conjunto de roles específico para cada clase verbal: *distance*, *agent*, *theme*, *goal*, *source*, *instrument*, *inanimate_cause*, etc.
(Stallard, 2000)	Conjunto de roles específico para el dominio de viajes en avión, como hora_salida, ciudad_origen o, ciudad_destino
(Busser & Moens, 2003)	Partiendo de la teoría de la gramática funcional desarrollada por (Halliday, 1994) y sus cinco tipos de procesos (material, relacional, existencial, mental y conductista), De Busser propone patrones de roles funcionales para cada uno de tales tipos. Cada patrón consta del verbo, el cual tiene el rol de proceso; un conjunto de roles participantes y algunos roles circunstanciales opcionales. Ejemplos de roles participantes son, el actor en un proceso material, el cliente en un proceso material, o el receptor en un proceso verbal. Ejemplos de circunstanciales son, circunstanciales de lugar, como lugar y tiempo; o de movimiento, como tiempo, lugar desde y lugar hacia
(Girju *et al.*, 2004)	Conjunto de relaciones semánticas para ayudar en la interpretación de nombres compuestos
(Bethard *et al.*, 2004)	Con el objetivo de mejorar la tarea de responder preguntas de opinión se propone extender los conjuntos de roles utilizados en FrameNet y PropBank con dos nuevos roles: *propositional opinion* y *opinion-holder*
VerbNet	23 roles temáticos: *actor, agent, asset, attribute, beneficiary, cause, location, destination, source, experiencer, extent, instrument, material, product, patient, predicate, recipient, stimulus, theme, time, topic*. Puesto que no pretenden definir un conjunto de roles exhaustivo, el inconveniente que puede plantear utilizar un conjunto limitado de roles es que no cubra todos los posibles argumentos para todas las clases de verbos. Sin embargo, según los propios autores, para los sentidos de verbos incluidos en el léxico, el conjunto considerado ha proporcionado suficiente información (Kipper, 2005). Más información en http://verbs.colorado.edu/ mpalmer/projects/verbnet.html consultado en marzo 2008.

Cuadro 2.17. Resumen de otras propuestas de roles semánticos (2/2)

2.2 Recursos lingüísticos basados en roles semánticos

Los conjuntos de roles semánticos comentados anteriormente han sido utilizados para la creación de diversos recursos lingüísticos, tales como corpus anotados o léxicos. Si, como se ha podido ver, la cantidad de conjuntos de roles semánticos definidos hasta la fecha es alta, una cantidad igualmente grande se traslada al número de recursos desarrollados. Aunque los más destacados se comentan a continuación, un resumen de sus principales características se muestra en el cuadro 2.18, en orden cronológico. En concreto, para cada recurso se puede ver el tipo de recurso que es, corpus, bases de datos, etc. (columna tipo); las lenguas para las que está disponible el recurso (columna lengua); el tipo de construcción (columna proceso); el conjunto de roles utilizado para construirlo (columna roles); la unidad básica para asignación de roles (columna nivel); el tamaño del recurso cuando se conoce (columna tamaño); el origen de los datos utilizados (columna base); y finalmente, las extensiones, si las tienen (columna extensiones).

De todos los recursos desarrollados, dos de ellos destacan por su completitud y usabilidad: los trabajos realizados dentro de los proyectos PropBank y FrameNet. Los siguientes apartados muestran un detalle de tales recursos (apartados 2.2.1 y 2.2.2), respectivamente, así como un resumen del resto de recursos desarrollados hasta la fecha (apartado 2.2.3).

Universitat d'Alacant
Universidad de Alicante

2.2.1 Proyecto PropBank

El objetivo del proyecto PropBank[3] (Palmer *et al.*, 2005) es documentar las realizaciones sintácticas de argumentos de los predicados de un léxico en inglés de propósito general. Tal objetivo se consigue anotando un corpus con roles semánticos. La finalidad de este corpus es proporcionar datos de entrenamiento para etiquetadores y analizadores semánticos estadísticos.

[3] http://verbs.colorado.edu/∼mpalmer/projects/ace.html Consultado en marzo 2008

Recurso	Tipo	Lengua	Proceso	Nivel	Roles	Tamaño	Base	Extensiones
Sinica Treebank	Corpus	Chino mandarín	Manual	Verbo	74 universal	361.834 palabras	Corpus Sinica	
LCS	BD clases verbos	Inglés	Manual	Verbo, nombre	12 universal	4.432 verbos		
FrameNet	Jerarquía clases semánticas	Inglés	Manual	Unidad léxica	Específ. clase	8.900 unidades léxicas		Salsa, Spanish FN, SemFrame, Kictionary, MS-FA, BiFrameNet, Japanese FN
	Corpus	Inglés	Manual	Unidad léxica	Específ. clase	135.000 oraciones	BNC + American Newswire	
PropBank	BD clases verbos	Inglés	Manual	Verbo	Uso de verbo			
	Corpus	Inglés	Aut.+Rev.Man.	Verbo	Específ. sent.verbo		WSJ Penn Treebank	NomBank, BioProp, PB vasco, PB chino, PB ruso, PB árabe, OntoBank
VerbNet	BD verbos	Inglés	Manual	Sentido verbo	Universal	5.257 sentidos de verbos		
PDT	Léxico verbos	Checo	Semi-aut.	sentidos verbos, nombres	FDG			VerbaLex
	Corpus	Checo	Semi-aut.	sentidos verbos, nombres	FDG		Czech National Corpus	
PCEDT	Corpus	Inglés, Checo		sentidos verbos	FDG		PEDT	
IAMTC	Corpus	Japonés, español, árabe, francés, hindi, coreano	Manual	unidades léxicas	LCS		DARPA + MTA	
CESS-ECE	Corpus	Español, catalán, vasco	Semi-aut.	verbos	PropBank	27.421 verbos	3LB	
SemSem	Corpus	Español	Manual	sentido verbo	7 universal		El periódico de Catalunya	
Adesse	BD verbos	Español	Manual	Verbo	Específ. clase		Arthus	
Lexpir	Jerarquía verbos	Español, inglés, catalán	Manual	Sentido verbo	Específ. sent.verbo	3.600 verbos		
HowNet	BD verbos	Inglés, chino	Manual	Verbo	69 universal	14.000 conceptos		
MUC	Corpus	Inglés	Manual	Verbo	6 Específ. terrorismo			
SIMPLE	Corpus	12 lenguas eur.	Manual	Sentido verbo	15 universal	10.000 sent.pals	PAROLE	
(Bethard *et al.*, 2004)	Corpus	Inglés,	Semi-aut.	verbo	Universal	5.139 oraciones	FN+PB	
(Gomez, 2007)	Léxico verbos	Inglés	Semi-aut.	Clase verbal	Específ. clase	3.017 predicados	WordNet	
	Corpus	Inglés	Semi-aut.	Clase verbal	Específ. clase	2.000 oraciones	Enciclopedia World Book	
D-Coi	Corpus	Alemán	Semi-aut.	Verbo	Uso de verbo	500 mill. pals.	Textos alemán contemporáneo	

Cuadro 2.18. Resumen de los recursos más utilizados

Universitat d'Alacant
Universidad de Alicante

Para ello añade información de predicado-argumento, o roles semánticos, a las estructuras sintácticas de la parte del Wall Street Journal (WSJ) utilizada para el proyecto Penn Treebank II[4] (Marcus, 1994). La anotación se realiza mediante un etiquetador automático basado en reglas (para más información consultar capítulo 4) cuya salida es corregida de forma manual. El conjunto de roles utilizado es el comentado anteriormente en el apartado 2.1, un conjunto de roles específico para cada uso de cada verbo, numerados de 0 a 5, más una lista de roles generales o adjuntos, como modo, lugar, causa, etc.

Recordemos que el conjunto de roles correspondiente a un uso de un verbo es denominado en PropBank, *roleset*, el cual está asociado a un conjunto de *frames* o marcos sintácticos, dando lugar a un denominado *frameset*. Un verbo polisémico puede tener más de un *frameset* cuando las diferencias en significado son suficientemente distintas como para requerir un conjunto de roles diferentes, uno por cada *frameset*. El procedimiento general es examinar un número de oraciones del corpus y seleccionar los roles que parece que ocurren con más frecuencia y/o son necesarios semánticamente (Kingsbury *et al.*, 2002).

Para cada rol se incluye un campo descriptor, como por ejemplo cosa aceptada, instrumento, etc., el cual es utilizado durante el proceso de anotación. Además, cada *frameset* se complementa con un conjunto de oraciones anotadas. Un ejemplo de un *frameset* se muestra en el cuadro 2.19. El conjunto de *framesets* de un verbo se denomina fichero de *frames* del verbo.

Para el proceso de anotación, puesto que se parte del corpus WSJ utilizado para el proyecto Penn Treebank, se tienen los árboles de análisis sintáctico revisados manualmente. En dicho proceso de anotación, basado en verbos, se anotan todas las ocurrencias de un verbo cada vez, comenzando por los verbos más comunes en el corpus. En este proceso se decide qué es argumento, a qué *frameset* hace referencia y qué rol juega el argumento. Los argumentos se restringen a nodos en el árbol de análisis.

[4] http://www.cis.upenn.edu/∼treebank/ Consultado en marzo de 2008

Universitat d'Alacant
Universidad de Alicante

Frameset ACCEPT.01: aceptar por voluntad propia	
ROL	DESCRIPCIÓN
Arg0	el que acepta
Arg1	cosa aceptada
Arg3	aceptado-desde
Arg4	atributo
[Arg0 He] [ArgM-MOD would][ArgM-NEG n't] accept [arg1 anything of value] [Arg2 from those he was writing about]. (wsj-0186)	

Cuadro 2.19. Ejemplo de un *frameset* en PropBank

Respecto al volumen de información proporcionado por el recurso, destacar que en WSJ hay 3.300 verbos que referencian 4.500 *framesets*.

Como resultado del proyecto se han obtenido dos recursos:

- Un *léxico* de ficheros de *frames*, hecho a mano, el cual almacena todos los significados de los verbos con su descripción y ejemplos.
- Un *corpus* anotado con un etiquetador automático basado en reglas (Palmer *et al.*, 2001) cuya salida ha sido revisada de forma manual.

Un índice de los verbos anotados y ejemplos puede verse en el portal web del proyecto[5].

El inconveniente que se le ha achacado a PropBank ha sido los problemas que plantea a la hora de hacer generalizaciones e inferencias basadas en etiquetas de roles semánticos. El hecho de que en PropBank cada etiqueta de rol sea específica para cada verbo, hace que no se pueda determinar cuando dos argumentos diferentes tienen el mismo rol semántico. Sin embargo, investigaciones recientes ponen de manifiesto el comportamiento robusto de PropBank así como la posibilidad de llevar a cabo generalizaciones (Zapirain *et al.*, 2008).

[5] http://www.cs.rochester.edu/ gildea/PropBank/ Consultado en marzo 2008.

Universitat d'Alacant
Universidad de Alicante

Siguiendo la anotación propuesta en PropBank se han desarrollado varios recursos, como NomBank, en el que se han anotado los argumentos de nombres; o extensiones a otras lenguas, como vasco, chino, ruso o árabe, y a dominios específicos, como el biomédico en BioProp.

Proyecto NomBank. Como parte de un esfuerzo mayor de añadir capas de anotación adicionales al corpus Penn Treebank II, el proyecto NomBank[6] (Meyers *et al.*, 2004b) tiene como objetivo anotar los conjuntos de argumentos que concurren con nombres en el corpus, de forma paralela a como PropBank lo realizó para verbos.

El conjunto de roles utilizado en la anotación ha sido un subconjunto del conjunto de etiquetas utilizadas en PropBank, en concreto el de las etiquetas arg0, arg1, arg2, arg3 y arg4, y ciertos roles de tipo argM como DIR, LOC, MNR, TMP, EXT, PRP, CAU, ADV y NEG (Meyers *et al.*, 2004a).

En NomBank los diferentes argumentos y adjuntos de cada sentido de los nombres núcleo (en inglés, *head nouns*) se etiquetan con roles. Por ejemplo, para el nombre *destruction* y la oración (E33) el argumento con rol arg0, el agente, es "*Richard*" y el argumento con rol arg1, el paciente, es "*the secret tapes*".

(E33) [$_{Arg0}$ Richard]'s destruction of [$_{Arg1}$ the secret tapes]

Universitat d'Alacant
Universidad de Alicante

Para la construcción de los *frames* de nombres se han utilizado los *frames* de PropBank, así como NOMLEX-Plus, un diccionario de nominalizaciones del Inglés (Meyers *et al.*, 1998). De esta manera, para nominalizaciones verbales anotadas en PropBank, se ha intentado mantener el conjunto de etiquetas utilizado en PropBank. Por ejemplo, el argumento con rol arg1 del sentido #1 del verbo *destroy*, se corresponde con el argumento de rol arg1 del sentido #1 del nombre *destruction*. En el caso de que la nominalización verbal no estuviera presente en PropBank, se ha añadido tanto la información relativa al verbo como la relativa al nombre.

[6] http://nlp.cs.nyu.edu/meyers/NomBank.html Consultado en marzo 2008

En 2006 la versión de NomBank incluía un total de 104.017 instancias de nombres en el corpus, de las cuales 3.290 son palabras diferentes.

Como resultado del proyecto se dispone además del recurso denominado NomBank-Plus, una versión de NomBank que ha sido extendida de forma semiautomática con unas 7.000 entradas.

BioProp. Con el objetivo de desarrollar un sistema de anotación automática de roles semánticos en el domino biomédico, (Tsai *et al.*, 2006) desarrollan de forma semiautomática, un corpus biomédico, denominado BioProp. Dicho corpus es anotado con verbos y argumentos, y sus correspondientes roles semánticos, siguiendo para ello el esquema de anotación de PropBank y las definiciones de VerbNet (para más información consultar apartado 2.2.3).

Biomedical Proposition Bank (BioProp) se construye a partir del corpus GENIA[7] (Kim *et al.*, 2003), una colección de 500 resúmenes que siguen el esquema de anotación del Penn Treebank II.

Para llevar a cabo el proceso de anotación semiautomático es nececesario adecuar los *framesets* de PropBank al dominio biomédico, dado que algunos verbos pueden tener usos diferentes en textos de periódicos y en textos biomédicos. Con este fin fueron seleccionados 30 verbos representativos para el dominio.

Tras la selección de los verbos y la definición de sus *framesets* se lleva a cabo el proceso de anotación automático cuya salida es manualmente revisada.

PropBank vasco. El PropBank vasco (Agirre *et al.*, 2006) persigue añadir una capa de anotación semántica a nivel de roles semánticos, al corpus vasco EPEC. Dicho corpus es una colección de unas 300.000 palabras de vasco estándar. Aproximadamente un tercio del corpus se obtuvo del corpus estadístico vasco *20th*

[7] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA Consultado en marzo de 2008

Universitat d'Alacant
Universidad de Alicante

Century[8] y el resto, de noticias del periódico *Euskaldunon Egunkaria*[9].

Hasta el momento, se ha llevado a cabo la anotación de tres verbos, seleccionados del total de verbos en el corpus: *esan* (en español, decir o contar), *adierazi* (en español, expresar) y *eskatu* (en español, preguntar por). Para llevar a cabo tal anotación se hace uso de tres recursos: el modelo utilizado en el proyecto PropBank; una base de datos, construida de forma manual, con marcos de subcategorización sintáctico-semánticos para verbos vascos (Aldezabal, 2004); y el corpus Basque Dependency Treebank (Aduriz *et al.*, 2003).

Actualmente se está trabajando en la extensión del corpus a su totalidad.

PropBank chino. A fin de llevar a cabo el proceso de anotación del corpus Penn Chinese Treebank[10] (Xue & Palmer, 2003) se ha construido una base de datos léxica, donde cada entrada corresponde a un predicado con sus *framesets* y su correspondiente *roleset*.

PropBank ruso. 4LEX (Civit *et al.*, 2005) es un corpus para ruso, manualmente construido a partir del léxico 3LB-LEX y sus equivalentes rusos. 3LB-LEX es obtenido de forma automática a partir de los verbos presentes en el corpus 3LB[11] (Palomar *et al.*, 2004).

PropBank árabe. El *Arabic PropBank* (APB), actualmente en desarrollo, comprende 200K palabras del corpus Penn Arabic Treebank III v.2 (ATB)[12] (Maamouri *et al.*, 2004). Para construir APB se han seleccionado los verbos que ocurren al menos 12 veces en ATB, lo que supone el 80 % de los datos sobre el total del corpus ATB. Una vez que los verbos son seleccionados, los anotadores construyen *frames* atendiendo al comportamiento sintáctico

[8] http://www.euskaracorpusa.net Consultado en marzo de 2008
[9] http://www.egunero.info Consultado en marzo de 2008
[10] http://www.cis.upenn.edu/ chinese/cpb Consultado en marzo 2008
[11] http://www.dlsi.ua.es/projectes/3lb/ Consultado en febrero de 2008
[12] http://www.ircs.upenn.edu/arabic/ Consultado en abril 2008

Universitat d'Alacant
Universidad de Alicante

y semántico expresado por el verbo y sus argumentos no adjuntos. Si un verbo tiene más de un sentido se divide en más de un *frame*, dependiendo de cómo se relacione con sus argumentos.

Actualmente proporciona anotación de roles semánticos para 454 predicados verbales.

OntoNotes. El proyecto OntoNotes (Hovy *et al.*, 2006), formalmente OntoBank, con el objetivo de proporcionar análisis semántico automatizado, se centra en una representación independiente de dominio, que incluya estructuras predicado-argumento, sentidos de palabras, relaciones ontológicas y correferencias. Como resultado del mismo se obtendrá un corpus mulitilingüe formado por 700 verbos, ricamente anotado, y construido con un acuerdo entre anotadores estimado del 90 %. Una primera versión inicial para inglés y chino está actualmente disponible.

2.2.2 Proyecto FrameNet

El proyecto FrameNet[13] surge como continuación a los trabajos realizados en el proyecto DELIS[14]. En DELIS se llevó a cabo una primera propuesta basada en marcos semánticos, obteniéndose como resultado un léxico para verbos de comunicación y de percepción en inglés, alemán, francés, italiano y danés, utilizando como interlingua los elementos de *frame* (Fillmore *et al.*, 2003).

El objetivo del proyecto FrameNet (Fillmore, 2002) es crear un recurso léxico para inglés con información detallada de las posibles realizaciones sintácticas de elementos semánticos. Dicha información se organiza en marcos semánticos (en inglés *semantic frames*), los cuales se pueden definir como representaciones esquemáticas de situaciones del mundo real.

Los marcos semánticos nacen del hecho de que el significado de las palabras se entiende mejor en referencia a las estructuras conceptuales que las soportan y las motivan. Por ello, los *frames*

[13] http://framenet.icsi.berkeley.edu/ Consultado en marzo 2008
[14] http://www.ims.uni-stuttgart.de/projekte/delis/ Consultado en febrero 2008

Universitat d'Alacant
Universidad de Alicante

no son necesariamente consistentes entre diferentes sentidos de un mismo verbo, sino que son consistentes entre diferentes verbos que comparten estructuras de argumentos similares.

Cada *frame* o marco semántico está formado por:

- El *nombre* del marco semántico y una descripción.
- Las *unidades léxicas* más frecuentes que pueden servir como predicados del *frame*, y que por tanto evocan dicho marco. Una unidad léxica es un par formado por una palabra y su significado, donde una palabra puede ser tanto un verbo, como un nombre o un adjetivo. Dichas unidades léxicas se agrupan en conjuntos de acuerdo a las descripciones semánticas que permiten. Por ejemplo, el marco semántico conversación es invocado por verbos como argumentar, debatir o conversar, o por nombres como disputa o discusión; o el marco semántico relativo a una transacción comercial puede ser evocado por palabras como comprar, vender, pagar, tarifa, cliente o comerciante, entre otras.

  Es importante destacar que una palabra polisémica sólo aparece con el sentido para el cual se haya definido el marco semántico.
- Lista de roles, denominados *elementos de frame* o de marco (en inglés *frame elements* (FE)), cada uno con su descripción, e información de las relaciones entre ellos. Por ejemplo, en el marco conversación, los elementos de marco incluyen protagonista, medio y tema; o por ejemplo, en el caso de una transacción comercial, comprador, vendedor, pago o mercancías, son elementos de ese marco. Se definen a nivel de unidades léxicas, de manera que el conjunto de unidades léxicas que evocan un mismo marco comparten los roles de ese marco.

  Ciertos roles son considerados como elementos de marco núcleo para un marco en particular, en el sentido de que ellos siempre están presentes conceptualmente. Frente a estos se encuentran los que no están siempre expresados en cada ocurrencia de un predicado que evoque el *frame*. Sería el caso de roles co-

Universitat d'Alacant
Universidad de Alicante

mo tiempo o lugar, que generalmente coinciden con adverbios o sintagmas preposicionales.

Para más detalle sobre el conjunto de roles utilizado en FrameNet ver sección 2.1.9.

- *Relaciones* entre marcos, tales como relaciones del tipo herencia, donde *frames* más específicos heredan todas las características de *frames* más generales. Por ejemplo, el marco denominado medios de comunicación hereda del marco comunicación.
- Conjunto de *ejemplos* anotados manualmente. Las oraciones ejemplo se eligen de manera que se asegure cobertura de todas las realizaciones sintácticas de los elementos de *frame*, prevaleciendo oraciones sencillas frente a otras más complejas (Palmer *et al.*, 2005).

  El corpus del que se han extraído las oraciones ejemplo es el British National Corpus (BNC)[15], aunque también se utilizó en menor medida, el corpus American Newswire[16].

Como ejemplo de un marco semántico ver el cuadro 2.12. Y para información más detallada de algunos *frames*, sus elementos de *frame* y las relaciones entre ellos, consultar Fillmore y Baker (2001).

Para el proceso de anotación no se han utilizado los árboles de análisis, si no que se ha partido de la información proporcionada por un *pos-tagger* y un lematizador. En dicho proceso, dada una unidad léxica, se determina en primer lugar el marco semántico al cual pertenece; se eligen oraciones en el corpus que contengan dicha unidad léxica; y se determina el elemento de *frame* o papel jugado por los constituyentes en la oración.

Actualmente, la base de datos de FrameNet consta de 8.900 unidades léxicas y 625 marcos semánticos con más de 135.000 oraciones anotadas[17].

[15] http://www.natcorp.ox.ac.uk Consultado en febrero de 2008
[16] http://americannationalcorpus.org Consultado en febrero de 2008
[17] http://framenet.icsi.berkeley.edu Consultado en marzo 2008

Entre los inconvenientes detectados en FrameNet, se cuentan: i) La poca cobertura con adjuntos, ii) la elección *ad-hoc* y bajo criterios de sencillez de los ejemplos anotados, iii) la especificidad de los roles, lo que hace que haya un gran número de ellos, y por tanto que haya problemas de dispersión de datos para aplicaciones de aprendizaje automático.

Con el objetivo de mejorar FrameNet, (Johansson & Nugues, 2006b) propone aumentar la información que dicho recurso proporciona con otro tipo de información semántica como argumentos nulos o marcadores de aspecto, entre otros.

Actualmente, se están realizando extensiones de FrameNet a otras lenguas como el alemán, proyecto *German FrameNet*, también conocido como proyecto SALSA; el japonés, proyecto *Japanese FrameNet*; el rumano; y el español, proyecto *Spanish FrameNet*. Además, se han realizado también propuestas de métodos automáticos que permitan inducir marcos semánticos a partir de otros recursos léxicos existentes, como es el caso de SemFrame; o propuestas que intentan superar las deficiencias encontradas en FrameNet, como por ejemplo MSFA. Incluso se han llevado a cabo trabajos en dominios específicos, como el caso de Kictionary. A continuación se muestra información más detallada sobre estas extensiones.

FrameNet Alemán. Salsa. El proyecto *Saarbrücken Lexical Semantics Annotation* (SALSA)[18] (Burchardt *et al.*, 2006) tiene como objetivo, entre otros, la creación un corpus para el alemán anotado semánticamente con información sobre roles semánticos. Utiliza para ello el corpus TIGER[19] (Brants *et al.*, 2002) el cual contiene 1,5 millones de palabras de textos de periódicos en alemán anotadas con información sintáctica.

Aunque se basa en los frames de FrameNet, integra también aspectos de otras aproximaciones. Las principales diferencias con

[18] http://gframenet.gmc.utexas.edu/ y http://www.coli.uni-saarland.de/projects/salsa/ Consultado en marzo de 2008

[19] http://www.ims.uni-stuttgart.de/projekte/TIGER/ Consultado en marzo de 2008

FrameNet, además de la lengua, son que la anotación se construye sobre representaciones sintácticas explícitas, que existe la posibilidad de asignar más de un *frame* a la misma expresión, y que sigue una aproximación basada en corpus. En consecuencia, se han de cubrir anotaciones de instancias no contempladas en FrameNet. Esto ha obligado a construir un conjunto de *frames* preliminares, denominados proto-*frames*, dando lugar a una versión *light* de un FrameNet en alemán.

De manera similar a FrameNet, los proto-*frames* de SALSA tienen una definición, un conjunto de roles con nombres al estilo de FrameNet, y oraciones ejemplo anotadas.

FrameNet Español. El proyecto FrameNet Español (en inglés, *Spanish* FrameNet -SFN-)[20] (Subirats & Petruck, 2003) está creando un recurso léxico para el español, basado en la semántica de *frames*. SFN proporciona un cuerpo de oraciones anotadas sintáctica y semánticamente, atendiendo al mismo proceso de anotación seguido en FrameNet (Subirats, 2006).

Utiliza un corpus de 350 millones de palabras que incluye una gran variedad de textos en español de diferentes géneros, principalmente artículos, revisiones de libros y ensayos de humanidades. Este corpus incluye tanto textos del español de América (60 %) como del español europeo (40 %).

La versión inicial de SFN, que está disponible desde julio de 2007, contiene alrededor de 1.000 unidades léxicas, verbos, nombres, adjetivos, adverbios, preposiciones y entidades, representativas de un amplio rango de dominios semánticos.

FrameNet Japonés. El objetivo del proyecto FrameNet japonés (en inglés, *Japanese* FrameNet -JFN-)[21] (Ohara *et al.*, 2004) es crear un léxico basado en corpus para el japonés, descrito en términos de marcos semánticos.

El corpus JFN contiene 8 millones de oraciones extraídas del periódico Mainichi y textos tomados de novelas y ensayos.

[20] http://gemini.uab.es:9080/SFNsite Consultado en marzo de 2008
[21] http://jfn.st.hc.keio.ac.jp Consultado en marzo de 2008

JFN se centra actualmente en analizar verbos básicos en japonés y más concretamente, en verbos y usos de verbos que no hayan sido descritos en detalle en diccionarios ya existentes.

A diferencia del proyecto en español, y aunque en un principio se pensó lo contrario (Ohara *et al.*, 2003), en el caso del japonés es necesaria la revisión y redefinición de los marcos semánticos definidos para inglés.

FrameNet Rumano. El punto de partida en los FrameNet alemán, español y japonés, fue la anotación manual a nivel de elemento de frame de corpus existentes para cada lengua. Para rumano, (Trandabăţ, 2007) propone crear un corpus empezando por la traducción de una parte del corpus de oraciones anotadas en inglés. Para ello define un método de importación de la anotación de FrameNet de inglés a rumano.

El primer paso en el proceso es la traducción por expertos de 110 oraciones del inglés, seleccionadas de forma aleatoria, y de las oraciones incluidas en el frame *Event*. A continuación se realiza una alineación automática de las oraciones, que es revisada antes de pasar al último paso, en el cual se importa la anotación del inglés, la cual también es revisada de forma manual.

Los primeros experimentos han dado lugar a la obtención de unas 1.000 oraciones.

Proyección de FrameNet a otras lenguas. En tareas como la traducción automática o la recuperación de información *cross*-lingüe, disponer de corpus bilingües es esencial. Sin embargo, la construcción de tales corpus para lenguas distintas al inglés son generalmente muy costosos.

Como solución a este problema se proponen mecanismos que permitan, a partir de un corpus anotado en inglés, derivar de forma automática corpus en otras lenguas. Si bien los corpus generados tendrán una calidad inferior a los manualmente generados, ellos podrán ser utilizados como semillas para ciertos métodos de aprendizaje automático, como *bootstrapping* (para más información consultar capítulo 3).

Respecto a corpus anotados con roles semánticos, se han realizado varias propuestas de proyección automática de FrameNet, ya sea mediante alineación de palabras y corpus paralelos como en Johansson y Nugues (2005b), Johansson y Nugues (2006a), Pado y Lapata (2005) y Boas (2005); como mediante la utilización de otros recursos, como por ejemplo, ontologías (Fung & Chen, 2004). De esta manera se han realizado proyecciones de inglés a español, sueco, alemán, francés y chino.

Kicktionary. Kicktionary[22] (Schmidt, 2006) es un diccionario electrónico trilingüe, para alemán, inglés y francés, restringido al dominio futbolístico.

Uno de sus puntos de partida es la teoría de marcos semánticos por lo que atendiendo a su estructura semántica y de argumentos, las unidades léxicas se agrupan en cientos de *frames*. Con el mismo criterio que en FrameNet, una unidad léxica es un par formado por una palabra, ya sea nombre, verbo, adjetivo o expresión idiomática, y su significado.

Un ejemplo de anotación se muestra en la oración (E34).

(E34) [*PLAYERSWITHBALL* Yattara] sidestepped his [*OPPONENTPLAYER* maker] and shot in form an acute angle

El corpus utilizado es una colección de reportajes de encuentros de fútbol en alemán, inglés y francés, extraídos del sitio web de la *Union of European Football Associations* (UEFA)[23]. Para cada lengua hay aproximadamente unos 500 textos, con unas 200.000 palabras. Este corpus es parcialmente paralelo, dado que la mayor parte de los textos son traducciones directas de uno a otro. En el caso del alemán, el corpus contiene además, reportajes de encuentros extraídos de la revista *on line* de fútbol, Kicker[24], y una hora de comentarios sobre fútbol extraídos de la radio alemana. Estos corpus añaden 1.200 textos y 750.000 palabras, para el caso de la revista; y 10.000 palabras para el caso de la radio.

[22] http://www.kicktionary.de Consultado en marzo 2008
[23] http://www.uefa.com/ Consultado en marzo 2008
[24] http://www.kicker.de Consultado en febrero 2008

Actualmente contiene 2.000 unidades léxicas, organizadas en 104 marcos semánticos.

SemFrame. SemFrame (Green, 2004; Green & Dorr, 2005) es un sistema desarrollado con el objetivo de generar marcos semánticos de forma automática. Dicho proceso se compone de dos fases: identificar clases verbales que evoquen un marco semántico común, y generar marcos semánticos, lo que supone identificar el nombre del *frame* y el conjunto de elementos de *frame*. Para un ejemplo de un marco semántico generado consultar el cuadro 2.20.

Frame **CONTEMPT**: Falta de respeto a una persona o cosa	
Rol	Descripción
Communication	Menospreciar
Person-1/Agent	
Person2/Recipient or Patient	

Cuadro 2.20. Ejemplo de los participantes de un *frameset* de SemFrame versión 2.0

MSFA: *Multilayered Semantic Frame Analysis of Text*. El análisis de marcos semánticos multinivel (en inglés, *Multilayered Semantic Frame Analysis* -MSFA-) (Kuroda *et al.*, 2006) es un marco de trabajo para anotación y análisis semántico inspirado en la aproximación FrameNet. MSFA extiende FrameNet en un intento de superar las deficiencias encontradas en FrameNet.

En el proceso de anotación destaca que el análisis semántico de una oración y la identificación de los *frames* necesitados para ello se realizan al mismo tiempo, a diferencia de FrameNet que parte de un conjunto predefinido de *frames*.

Como resultado se obtiene una pequeña colección de textos en japonés anotados semánticamente, tomados del corpus Kyodai (Kurohashi & Nagao, 2003). En concreto se han anotado 3 artículos de periódicos, con un total de 63 oraciones. Este proceso permitió identificar 700 *frames*.

2.2.3 Otros recursos lingüísticos

- El proyecto **VerbNet**[25] (Kipper, 2005) surgió con el objetivo de cubrir algunas de las limitaciones de otros recursos léxicos de verbos existentes como WordNet (Miller *et al.*, 1990) o las clases de verbos de Levin (Levin, 1993).

  VerbNet es un léxico verbal de gran cobertura, independiente del dominio, no ligado a ningún corpus, y que es compatible con otros recursos existentes, pero con información sintáctica y semántica explícitamente establecidas, utilizando las clases de verbos de Levin para sistemáticamente construir las entradas léxicas.

  Las clases, que están jerárquicamente organizadas, contienen información sobre el conjunto de verbos miembro, los roles temáticos para las estructuras predicado-argumento de los verbos de la clase, y un conjunto de ejemplos, entre otros. La lista de roles considerada incluye 23 roles semánticos (ver cuadro 2.17).

  Aunque las clases incluidas son suficientes para algunas tareas de PLN, no son completas. Por ello, en Kipper *et al.* (2006a) se propone integrar en VerbNet la extensión realizada por Korhonen y Briscoe (2004) a la clasificación de Levin. Dicha extensión añade 57 clases nuevas para tipos de verbos no contemplados en la clasificación inicial de Levin.

  Actualmente, VerbNet tiene 5.257 sentidos de verbos y 274 clases verbales de primer nivel (Kipper *et al.*, 2006b).

- El corpus **Sinica Treebank**[26] (Huang *et al.*, 2000) es uno de los primeros corpus para el Chino Mandarín, anotados estructuralmente e incluyendo información sobre roles semánticos.

  Los textos de Sinica Treebank han sido extraído del corpus Sinica[27] (Chen *et al.*, 1996) el cual incluye textos sobre diversos

[25] http://verbs.colorado.edu/ mpalmer/projects/verbnet.html Consultado en marzo 2008

[26] http://godel.iis.sinica.edu.tw/CKIP/engversion/treebank.htm Consultado en marzo 2008

[27] http://www.sinica.edu.tw/ftms-bin/kiwi.sh Consultado en febrero 2008

Universitat d'Alacant
Universidad de Alicante

temas, como por ejemplo filosofía, ciencia, sociedad, arte, o literatura.

En el proceso de anotación no sólo se anotan las relaciones semánticas de un predicado verbal, sino que también se anotan los argumentos de nombres. Dicho proceso de anotación utiliza la anotación morfológica revisada, y la anotación sintáctica, obtenida de forma automática el corpus Sinica,

La anotación utiliza un conjunto de 74 roles semánticos incluyendo roles temáticos tales como *agent*, *theme*, o *instrument*; adjuntos como *location*, *time* o *manner*; y modificadores específicos para nombres como *predication*, *possessor* o *quantifier*. El detalle del conjunto de roles utilizado se muestra en la figura 2.4.

Sinica TreeBank 3.0 contiene 61.087 árboles sintácticos y 361.834 palabras.

- El **Prague Dependency Treebank** (PDT) [28] (Hajič *et al.*, 2006), diseñado y elaborado en el Instituto de Lingüística Aplicada y Formal de la Universidad Charles de Praga, es un corpus electrónico que contiene una gran cantidad de textos checos, los cuales han sido extraídos del Czech National Corpus[29] (Čermák, 1995). Tales textos consisten en un 40 % de artículos de periódico generales, un 20 % de noticias de economía y análisis, un 20 % de revistas de ciencia popular y otro 20 % de textos de tecnología de la información.

  Una vez más, la idea es crear un corpus que contenga tanta información sintáctico-semántica como sea posible, de manera que sea un recurso útil para tareas de PLN. El corpus es anotado con información de valencia basándose en la teoría *Functional Generative Description* (FDG), comentada anteriormente en la sección 2.1.3, en la que los roles semánticos reciben el nombre de *functors*.

[28] http://ufal.mff.cuni.cz/pdt2.0 Consultado en marzo 2008
[29] http://ucnk.ff.cuni.cz Consultado en febrero de 2008

Universitat d'Alacant
Universidad de Alicante

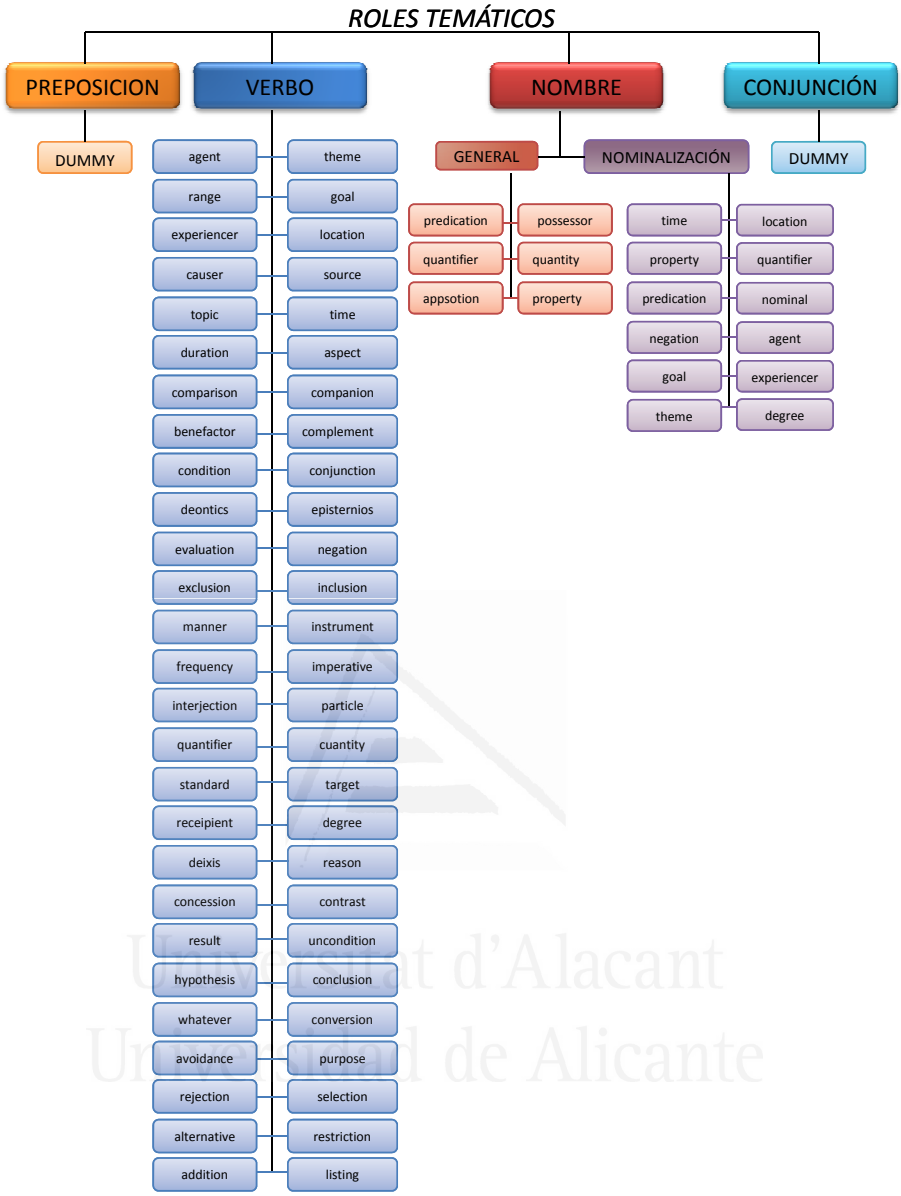


Figura 2.4. Conjunto de roles semánticos utilizados en Sinica Treebank

Universitat d'Alacant
Universidad de Alicante

El PDT tiene una estructura a tres niveles, si bien la cantidad de material anotado decrece conforme el nivel aumenta. En el tercer nivel, denominado nivel tectogramático o de significado lingüístico hay un total de 0,8 millones de palabras. En este nivel a cada oración se le asigna una estructura de árbol con nodos y arcos etiquetados. Los arcos del árbol estableciendo relaciones entre los nodos representan las relaciones entre las palabras de una oración, y el tipo de relación es indicada mediante una etiqueta denominada *functor*.

El proceso de anotación se llevó a cabo de forma semiautomática. A partir de la información en niveles anteriores, y mediante un proceso automático se crearon árboles intermedios, los cuales fueron revisados y corregidos por anotadores.

La siguiente oración (E35) muestra un ejemplo de anotación para el verbo *říkat.1* (en inglés, *to speak*), el cual posee ACT y EFF como *functors* obligatorios, y PAT y ADDR como opcionales.

(E35) Petr.ACT mu.ADDR říkal o Marii.PAT pravdulže je chytrá.EFF

Como resultado de este proceso de anotación se generó una lista de marcos de valencia, denominada PDT-VALLEX. Cada marco contiene el conjunto de roles semánticos y un conjunto de ejemplos de uso.

A partir de la información proporcionada por PDT-VALLEX, se han construido otros recursos, como por ejemplo VerbaLex (Hlaváčková & Horák, 2006), un léxico de valencias de verbos para el checo que muestra información sobre sus roles semánticos. Debido al hecho de que los roles semánticos utilizados en PDT-VALLEX son demasiado genéricos, VerbaLex plantea una anotación semántica a dos niveles. Para ello utiliza elementos seleccionados de la TopOntology y del conjunto de Conceptos Base, ambos de EuroWordNet (Vossen, 1998). Por ejemplo, el elemento *object* de la TopOntology, incluye los elementos *na-*

tural object, *structure*, *container*, *furniture* o *tableware*, en el segundo nivel.

Tras el trabajo realizado para la construcción del corpus PDT surgió la idea de crear un corpus paralelo checo-inglés basado en dependencias con anotación a nivel tectogramático, el Prague Czech-English Dependency Treebank (PCEDT)[30], que diera apoyo a tareas de traducción automática. Este corpus paralelo, en el que se realizaron traducciones manuales para el checo, consta de 21.600 oraciones extraídas de la sección del Wall Street Journal del Penn Treebank. Mientras que para la parte del checo la anotación se lleva a cabo de forma automática, la anotación del inglés (Prague English Dependency Treebank -PEDT-) se realiza de forma manual.

Con el objetivo de facilitar la anotación del corpus PEDT al nivel tectogramático y de aumentar el acuerdo entre anotadores, se ha desarrollado un léxico de verbos en inglés construido dentro del marco de la teoría FDG, el denominado EngValLex (Semecký & Cinková, 2006).

- El proyecto **CESS-ECE**[31] (Taulé *et al.*, 2006), con el objetivo de facilitar a los investigadores un recurso que permita realizar estudios comparativos sobre el comportamiento sintáctico y semántico de las lenguas del proyecto 3LB[32] (Palomar *et al.*, 2004), lleva a cabo la creación de tres corpus, uno para el español (CESS-Esp), otro para el catalán (CESS-Cat) y otro para el euskera (CESS-Eus), de 500.000 palabras los dos primeros y 350.000 el de euskera. Estos corpus estarán etiquetados sintáctica y semánticamente.

  El proceso semiautomático de anotación de información semántica, asigna a cada predicado una estructura léxico-semántica que determina la estructura argumental, las relaciones temáticas de los argumentos respecto al verbo y las alternancias de diátesis que admite.

[30] http://ufal.mff.cuni.cz/pcedt Consultado en marzo 2008
[31] http://www.lsi.upc.edu/ mbertran/cess-ece Consultado en marzo 2008
[32] http://www.dlsi.ua.es/projectes/3lb Consultado en febrero de 2008

Universitat d'Alacant
Universidad de Alicante

La construcción de tales estructuras siguen la línea utilizada en la construcción de VerbNet, adoptando la propuesta de PropBank en la caracterización de los argumentos, donde se distingue entre los argumentos obligatorios, arg0, arg1, etc., y argumentos opcionales o adjuntos, argM.

El número total de formas verbales para castellano es de 7.127 y 1.070 el número de verbos distintos. Para el catalán, se dispone de 7.033 formas verbales y 834 verbos distintos, mientras que para euskera hay 13.261 formas verbales y 375 verbos distintos.

- El principal objetivo del proyecto *Sentence Semantics* (**SenSem**)[33] (Castellón *et al.*, 2006) es la construcción de una base de datos léxica que ilustre el comportamiento sintáctico y semántico de cada uno de los 250 sentidos de verbos más frecuentes para el español. Dicho banco de datos se compone de un léxico donde cada sentido verbal está asociado a un conjunto de ejemplos del corpus analizados y anotados a diferentes niveles. El corpus anotado, el cual está formado por oraciones extraídas de la versión electrónica del periódico "El Periódico de Catalunya", contiene aproximadamente 750.000 palabras con 100 ejemplos para cada verbo, obtenidos de forma aleatoria.

  A fin de incrementar la consistencia en el proceso de anotación manual se ha creado una base de datos léxica de verbos[34] proporcionando información de roles semánticos para cada sentido. En la descripción de los verbos se incluye información sobre el conjunto de papeles semánticos asociado (ver cuadro 2.21) y ejemplos del corpus anotados. Se han descrito un total de 1.122 sentidos de verbos con una media de 4,5 sentidos por lema.

- El principal objetivo del proyecto Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español (**ADESSE**)[35] (García-Miguel & Albertuz, 2005) es proporcionar definiciones manuales e información acerca de roles semánticos y miembros de clases semánticas para todos los verbos de una base de da-

[33] http://grial.uab.es/search Consultado en marzo 2008
[34] http://grial.uab.es/adquisicio Consultado en marzo 2008
[35] http://webs.uvigo.es/adesse/ Consultado en marzo 2008

Rol	Subrol
Initiators	*agent, causal theme*
Themes	*holistic, incremental beneficiary, victims, creation, destruction, consequence*
Localizations	*spatial, temporal, abstract, source, position, direction, trajectory*
Quantity	
Accompaniement	
Instrument	
Identification	

Cuadro 2.21. Lista de roles utilizados en el proyecto SenSem

tos sintáctica de aproximadamente 160.000 cláusulas. Estas han sido extraídas de un corpus para el español de 1.5 millones de palabras, el corpus Arthus[36]. Este corpus contiene textos de diferentes ámbitos, como narrativo, ensayo, teatral, periodístico, o textos orales transcritos, y tanto textos en español de Europa como en español de América.

El principal resultado del proyecto será una base de datos sintáctico-semántica basada en corpus que incluya, entre otros, información sobre roles semánticos. Esta información se organiza en clases y subclases verbales de manera que cada nivel hereda la información semántica por defecto establecida en niveles superiores. Es decir, no se anota cada cláusula del corpus, si no los esquemas sintácticos que ellas instancian. Los esquemas sintácticos apuntan a los roles que se definen para cada sentido verbal.

El conjunto de roles utilizado se sitúa en un nivel intermedio de abstracción. No se han utilizado conjuntos de roles generales, como los protoroles, ni conjuntos de roles específicos para cada verbo como PropBank.

- La base de datos **LCS**[37] (Dorr *et al.*, 2001) contiene estructuras conceptuales léxicas (en inglés, *Lexical Conceptual Structure* -

[36] http://www.bds.usc.es/corpus.html Consultado en febrero de 2008
[37] http://www.umiacs.umd.edu/∼bonnie/LCS_Database_Documentation.html Consultado en marzo 2008

LCS-) construidas a mano y organizadas en clases semánticas. Estas clases fueron derivadas de las clases de verbos de Levin y posteriormente modificadas, principalmente por división de algunas clases en subclases y por adición de clases nuevas. La base de datos aumentada contiene 4.432 verbos en 492 clases, con un total de 11.000 entradas de verbos.

Cada clase está asociada con un *grid* temático (Θ-grid) que resume el comportamiento de los verbos especificando su estructura predicado-argumento. Por ejemplo, la clase de verbos *Roll*, está asociada con el Θ-grid formado por los roles *theme* y *goal* (Green *et al.*, 2001). La pertenencia de un rol particular a una clase de verbos se basa en propiedades perfiladas en Jackendoff (ver sección 2.1.7). El cuadro 2.22 muestra en detalle el conjunto de posibles roles.

Rol	Descripción
Agent	Agente
Experiencer	Experimentador
Information	Información
Theme	Tema
Source	Preposición indicando origen, por ejemplo, *from* o *away from*
	Origen del movimiento
Goal	Preposición indicando objetivo, por ejemplo, *at*, *to*, *toward*
	Punto final del movimiento
Identificational Predicate	Preposición indicando objetivo en el campo de identificación
	Cosa o propiedad
Perceived	Entidad que puede ser percibida
	Preposición indicando camino de la percepción
Location	Preposiciones precediendo lugares estáticos
	Lugar estático
Possessional	Entidad poseída
Time	Preposiciones precediendo tiempo
	Argumento temporal
Proposition	Evento o estado

Cuadro 2.22. Conjunto de roles utilizados en LCS

Universitat d'Alacant
Universidad de Alicante

Múltiples entradas de un verbo representan en la base de datos diferentes sentidos del verbo. Por ejemplo, el cuadro 2.23 muestra los conjuntos de roles para cada una de las clases a las que pertenecen diferentes sentidos del verbo *drop* (Green *et al.*, 2001).

Clase/Sentido	Roles requeridos	Roles opcionales
Drop	*Agent* *Theme* *Goal*	
Putting down	*Agent* *Theme*	Loc *Source* *Goal*
Calibratable changes of state	*Theme*	
Meander (to/from)	*Theme* *Source (from)* *Goal (to)*	
Meander (path)	*Theme* *Goal*	
Roll 1	*Theme*	
Roll 2	*Theme*	*Source* *Goal*
Roll down	*Theme* *Particle*	*Source* *Goal*

Cuadro 2.23. Sentidos del verbo *drop* en LCS

- El proyecto *Interlingual Annotation of Multilingual Corpora* (**IAMTC**)[38] (Reeder *et al.*, 2004), con el objetivo de obtener un recurso útil para diferentes tareas de PLN como traducción automática o recuperación de información, se centra en la creación de un sistema de representación semántica, conocido como representación interlingua, y en el desarrollo de corpus bilingües anotados semánticamente que relacionen textos en inglés con sus correspondientes textos en Japonés, Español, Árabe, Francés, Hindi y Coreano.

[38] http://aitc.aitcnet.org/nsf/iamtc Consultado en marzo 2008

El conjunto de datos consta de 6 corpus bilingües con unas aproximadamente 150.000 palabras cada uno. Cada corpus está formado por 125 artículos de periódicos, cada uno de los cuales contiene normalmente entre 300 y 400 palabras. Los artículos extraídos para cada corpus individual son diferentes. Los corpus para el español, francés y japonés, se basan en los datos de evaluación de la iniciativa *Defense Advanced Research Projects Agency* (DARPA) *Machine Translation* (MT) (White & O'Connell, 1994); el corpus en árabe se basa en el corpus *Multiple-Translation Arabic* (MTA) Parte 1 producido por el *Linguistic Data Consortium* (LDC) (Walker *et al.*, 2003).

La representación semántica abarca tres niveles, denominados IL0 (información sintáctica), IL1 (información semántica) e IL2, e incorpora conocimiento de fuentes tales como la ontología Omega (Philpot *et al.*, 2005) y los $\Theta-grid$ de LCS (Dorr *et al.*, 2001) comentados anteriormente en este mismo apartado. El proceso de anotación manual es incremental, añadiendo en cada nivel nuevas características semánticas y eliminando las sintácticas.

El conjunto de roles utilizado corresponde a una versión simplificada de los $\Theta-grid$ de la base de datos LCS, cuyo detalle puede verse en (Farwell *et al.*, 2004).

- El proyecto **VOLEM** (Verbos: Organización Léxica Multilingüe)[39] es una base de datos léxica multilingüe de un subconjunto de verbos del español, el catalán, el francés y el vasco.

  Para cada verbo se especifica diferente información en las diferentes lenguas, como por ejemplo, los roles semánticos de los argumentos y ejemplos de uso.

- **Pirapides** (Vázquez *et al.*, 2000) es un proyecto centrado en el estudio de predicados verbales para español, inglés y catalán. Dentro de este proyecto se llevó a cabo la construcción manual del léxico LEXPIR (Atserias *et al.*, 2000), un léxico jerárquico de verbos para el español. Partiendo de la clasificación de verbos de

[39] http://grial.uab.es/fproj.php?id=4 Consultado en marzo de 2008

Levin, la agrupación jerárquica de verbos se realizó teniendo en cuenta el significado de los componentes semánticos así como sus alternancias. Cada grupo, a su vez, es clasificado de acuerdo al número de componentes que se pueden realizar explícitamente. Por ejemplo, dentro de la clase trayectoria hay cuatro subclases: movimiento no autónomo, movimiento autónomo, comunicación y transferencia.

Para cada sentido de cada verbo se dispone de diferente información, como por ejemplo los roles semánticos. Esta información se propaga dentro de la jerarquía, de forma que cada verbo hereda los elementos de su grupo y cada grupo de su clase, si bien cada verbo puede imponer sus propias restricciones. Por ejemplo, el verbo charlar, a diferencia del resto de verbos de comunicación no acepta un sintagma nominal en el componente entidad.

El cuadro 2.24 muestra un ejemplo del modelo básico para verbos de trayectoria. Cuando un valor no se indica, por ejemplo, la sintaxis para el componente entidad, es porque dicho valor varía dependiendo de la subclase. En este cuadro, la columna componente representa los roles semánticos de la clase; la columna semántica toma sus valores de la TopOntology de EuroWordNet; la columna acuerdo indica si los elementos deben de estar o no de acuerdo con el verbo de la oración; y La columna opcional indica cuando un elemento es opcional en la oración. El patrón de las oraciones de esta clase se muestra en (E36).

Universitat d'Alacant
Universidad de Alicante

(E36) Alguien <iniciador> traslada algo <entidad> por X <camino> de Y <origen> a Z <destino>

Sintaxis	Preposición	Componente	Semántica	Acuerdo	Opcional
NP	de inicio	iniciador	Human	yes	yes
		entidad	Top	no	yes
PP	de ruta	camino	Top	no	yes
PP	de origen	origen	Top	no	yes
PP		destino	Top	no	yes

Cuadro 2.24. Modelo básico para verbos de trayectoria

Como resultado del proyecto se obtuvo un corpus de 4.006 oraciones y una base de datos de 3.600 formas verbales.

- **HowNet**[40] (Dong & Dong, n.d.) es una base de conocimiento bilingüe inglés-chino construida manualmente, que codifica relaciones entre conceptos y relaciones entre atributos. En HowNet hay 65.000 conceptos en chino y 75.000 equivalentes para inglés.

  Cada concepto contiene 7 campos incluyendo una definición del concepto, su categoría y relaciones semánticas. Respecto a las relaciones semánticas, cada concepto se asocia con una especificación de roles temáticos. HowNet utiliza un conjunto de 69 roles, de los cuales los diez más utilizados corresponden a: *agent*, *patient*, *content*, *experiencer*, *target*, *direction*, *locationFin*, *locationIni*, *locationThru* y *cause*. Por ejemplo, el verbo *cure*, se asocia a los roles *agent*, *patient*, *content* y *tool*.

- Los **corpus del MUC** definen *frames* de caso pertenecientes al dominio terrorista. Estos *frames* contienen *slots* para los roles temáticos que están asociados con cada evento. Si bien inicialmente el conjunto de *frames* fue construido manualmente, se realizaron propuestas para llevar a cabo construcciones semiautomáticas con una cantidad limitada de labor manual. Para obtener más detalle sobre estas propuestas consultar Riloff y Schmelzenbach (1998).

  El conjunto de roles utilizado es específico para el dominio del terrorismo.

- El léxico semántico **SIMPLE**[41] (Lenci *et al.*, 2000; Ruimy *et al.*, 2002), ha sido desarrollado como una continuación al proyecto PAROLE[42] (Ruimy *et al.*, 1998). Su objetivo es añadir información semántica al léxico con información morfológica y sintáctica para 12 lenguas europeas (catalán, danés, alemán, inglés, finlandés, francés, holandés, griego, portugués, español, sueco e italiano) construido en PAROLE.

[40] http://www.keenage.com Consultado en marzo 2008
[41] http://www.ub.es/gilcub/SIMPLE/simple.html Consultado en marzo 2008
[42] http://www.elda.fr/catalogue/en/text/doc/parole.html Consultado en abril 2008

Universitat d'Alacant
Universidad de Alicante

El léxico, que contiene alrededor de 10.000 significados de palabras, contiene para cada predicado, los argumentos y sus roles semánticos, de un conjunto de 15 roles semánticos universales (*agent*, *patient*, *cause*, *topic*, *experiencer*, *beneficiary*, etc.).

- El corpus *Global Document Annotation* (GDA)[43] incluye información semántica para textos en japonés. En concreto contiene información sobre las relaciones representadas por los roles semánticos agente, paciente, instrumento, material, medida, etc. El corpus ha sido construido de forma manual.

- **(Bethard *et al.*, 2004)** propone extender los conjuntos de roles de FrameNet y PropBank con los roles *propositional opinion* y *opinion-holder*, con el fin de poder responder de forma automática preguntas de opinión del tipo "How does X feel about Y?".

  En el caso de FrameNet se selecciona un conjunto de oraciones que contenían argumentos con el rol proposition. Cada una de estas oraciones fu de nuevo anotada con los nuevos roles. Esto produjo un conjunto de 3.041 oraciones.

  Para PropBank se selecciona un conjunto de oraciones conteniendo uno de los 27 verbos considerado altamente probable a ocurrir con oraciones de opinión. El resultado dio lugar a un conjunto de 2.098 oraciones.

  En un primer paso de anotación manual las oraciones se anotaron con etiquetas: *non-opinion*, *opinion-propositional* y *opinion-sentence*. En un segundo paso automático con revisión manual, se anotaron *opinion-holders*.

- **Gomez** propone la construcción de forma semi-automática de una jerarquía de clases verbales a partir de procesos de redefinición y reorganización de las clases verbales de WordNet (Gomez, 1998). Cada clase verbal contiene información sobre los predicados incluidos en la clase, el conjunto de roles semánticos específicos para esa clase, y las restricciones de selección y realizaciones sintácticas de dichos roles. En la jerarquía de clases,

[43] http://i-content.org/GDA/tagset.html#intro

Universitat d'Alacant
Universidad de Alicante

los roles semánticos de las superclases son heredados por clases en niveles inferiores, si bien es posible definir roles semánticos específicos para las subclases.

La jerarquía está formada por unas 3.017 predicados (Gomez, 2004), cubriendo el 98 % de las clases de verbos de WordNet.

Además, con el fin de probar la utilidad de esta jerarquía se obtiene un corpus de 2.000 oraciones anotadas con los roles semánticos que las clases verbales definen (Gomez, 2007). Las oraciones fueron extraídas de la enciclopedia *World Book*[44].

- El proyecto *Dutch Language Corpus Initiative* (D-Coi)[45] tiene por objetivo la creación de un corpus de 500 millones de palabras de alemán escrito contemporáneo con anotación a nivel semántico. La anotación de roles semánticos sigue el esquema de anotación de PropBank.

  El proceso de anotación parte de las oraciones anotadas sintácticamente de forma automática. A continuación, mediante un conjunto de reglas que permite utilizar la traducción al alemán de la información disponible en PropBank, se asignan los roles semánticos a las oraciónes. Finalmente, se realiza un proceso de revisión y corrección manual.

2.3 Relaciones entre recursos

Universitat d'Alacant
Universidad de Alicante

El análisis realizado hasta el momento de los diferentes recursos disponibles, pone de manifiesto, por un lado, que cada uno de ellos codifica una clase de conocimiento diferente, y por otro, que cada uno de ellos posee sus propios puntos fuertes y débiles, a pesar de los muchos años de laborioso esfuerzo humano que su desarrollo ha requerido. Combinarlos juntos puede resultar en una base de conocimiento más rica que pueda posibilitar análisis semántico más robusto y preciso (Shi & Mihalcea, 2005).

[44] http://www.worldbook.com Consultado en marzo 2008
[45] http://lands.let.ru.nl/projects/d-coi/ Consultado en abril 2008

Por ello, diferentes investigadores han propuesto diferentes procesos para establecer relaciones entre los recursos. Actualmente existen procesos para relacionar:

- **PropBank** con
  - LCS (Dorr *et al.*, 2001; Hajičová & Kučerová, 2002; Rambow *et al.*, 2003; Kwon & Hovy, 2006)
  - VerbNet (Rambow *et al.*, 2003; Kipper, 2005; Pazienza *et al.*, 2006; Giuglea & Moschitti, 2006c; Loper *et al.*, 2007).
  - FrameNet (Giuglea & Moschitti, 2006c)
  - EngValLex (Cinková, 2006)
  - Conjunto de roles específico para sistemas de búsqueda de respuestas (Navarro *et al.*, 2004)
- **FrameNet** con
  - VerbNet (Shi & Mihalcea, 2005; Kipper, 2005)
  - LCS (Kwon & Hovy, 2006)
  - Conjunto de 18 roles semánticos (Gildea & Jurafsky, 2002)
  - HowNet (Fung & Chen, 2004)
- **LCS** con
  - PDT (Hajičová & Kučerová, 2002)
  - HowNet (Dorr *et al.*, 2002)

Universitat d'Alacant
Universidad de Alicante

El cuadro 2.25 muestra un resumen de tales procesos, donde el número en la matriz indica la cantidad de procedimientos diferentes para relacionar los recursos implicados. Por ejemplo, entre FrameNet y PropBank se han realizado dos propuestas diferentes que establezcan relaciones entre los conjuntos de roles utilizados en FrameNet y los utilizados en PropBank.

Respecto al conjunto de roles semánticos específico para sistemas de búsqueda de respuestas propuesto en este trabajo (ver

LCS		4	1		1	1	
PropBank	4		1	5		1	1
FrameNet	1	1		1	1		
VerbNet		5	1				
HowNet	1		1				
PDT	1	1					
PCEDT		1					
	LCS	PropBank	FrameNet	VerbNet	HowNet	PDT	PCEDT

Cuadro 2.25. Procedimientos para relacionar recursos

apartado 2.1.11), se ha establecido una relación con el conjunto de roles propuesto en el proyecto PropBank, tal y como se muestra en el cuadro 2.26. Las relaciones establecidas han sido construidas manualmente, y definidas a partir de la observación de ejemplos anotados con ambos conjuntos de roles semánticos.

PropBank	Cjto. roles para BR
A0	Proto-agent
A1	Proto-patient: T-P
A2	Proto-patient: B-R
	Proto-agent: Instrument
	Location
A3	Proto-patient
	Proto-agent: Instrument
	Location
A4	Location: Goal
AM-LOC	Location
AM-MNR	Mode
AM-TMP	Temporal

Cuadro 2.26. Correspondencia entre PropBank y la propuesta de Moreda *et al.* (2007)

Puesto que el conjunto de roles semánticos propuesto en este trabajo presenta una estructura jerárquica, siempre que ha sido posible se ha elegido establecer la relación con los roles de los niveles inferiores. Por ejemplo, el rol A4 se relaciona con el rol *goal*, subrol de *location*. Sin embargo, para casos generales ha sido preferible establecer las relaciones con niveles superiores. Por

Universitat d'Alacant
Universidad de Alicante

ejemplo, A0 y A1 se relacionan con los roles proto-agent y proto-patient, respectivamente. Esto se debe a que A0, generalmente, es un proto-agent: *agent*. Sin embargo, hay oraciones como (E37), en las que el rol A0, "*the wind*", tiene el rol proto-agent: *cause*. Por ello, con el fin de resolver esta ambiguedad, se prefiere relacionar A0 con proto-agent.

(E37) The wind closed the door

Por otra parte, hay roles como A1 y A3 que se relacionan con varios roles del conjunto propuesto. En este caso, la relación se establece por defecto con proto-patient, pero para el caso de determinados verbos si alguno de los roles numerados tuviera ese rol, entonces la relación se establece con el siguiente en la lista. Por ejemplo, si un verbo tiene un argumento con el rol A2 con el sentido de proto-agent y otro argumento con rol AM-LOC, entonces el argumento con rol A3 tendrá sentido de instrumento.

Además, hay que recordar que el conjunto de roles propuesto para sistemas de BR no intenta cubrir el espectro completo de posibles roles semánticos, por lo que no todos los roles de PropBank tendrán una correspondencia en dicho conjunto.

En definitiva se establece, que si bien no se pueden hacer generalizaciones a través de los verbos y sus roles en PropBank, la definición de un conjunto de roles semánticos jerárquico ayuda a realizar ese tipo de generalizaciones y dotar, así, de mayor significado a los roles numerados de PropBank.

Universitat d'Alacant
Universidad de Alicante

3. Enfoques para el tratamiento de Roles Semánticos

La construcción de los recursos lingüísticos presentados en el capítulo 2 se lleva a cabo con el objetivo principal de facilitar el desarrollo de herramientas que permitan identificar o anotar los roles semánticos para su posterior uso, tanto en tareas de PLN, como pueden ser desambiguación del sentido de las palabras o análisis contextual; como en sistemas de PLN, como por ejemplo, búsqueda de respuestas o implicación textual.

Como cualquier otra tarea de PLN, y como se verá en el capítulo 4, las herramientas desarrolladas para la anotación de roles semánticos han seguido principalmente, dos enfoques, ya sea de manera individual o como combinación de ambos: i) hacer uso de conocimiento lingüístico previamente adquirido, o ii) utilizar corpus anotados previamente construidos.

En este capítulo se analizarán ambos enfoques. En primer lugar se presentarán tanto las principales características de los sistemas que hacen uso de corpus anotados, los denominados sistemas basados en corpus, como los principales algoritmos desarrollados hasta el momento (apartado 3.1). Este estudio, prestará especial atención a los procesos de selección de características, tan importantes en esta clase de sistemas, así como a los métodos que implementan dichos procesos de selección (apartado 3.1.4). Finalmente, se mostrarán las ideas básicas sobre los sistemas que hacen uso de conocimiento lingüístico previamente adquirido, los sistemas basados en conocimiento (apartado 3.2).

Universitat d'Alacant
Universidad de Alicante

3.1 Enfoques basados en corpus

El enfoque basado en corpus, también denominado aprendizaje automático (en inglés, *machine learning* -ML-) es una rama de la Inteligencia Artificial relacionada con el desarrollo de algoritmos y técnicas que permitan a los ordenadores aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento.

La información no estructurada que ha de suministrarse al ordenador tendrá que ser traducida o representada en algún formato legible computacionalmente. La representación se hace en forma de atributos. Un atributo, también denominado característica o rasgo, se define como la descripción de alguna medida de una muestra o entidad tratada en el problema de aprendizaje automático en estudio. Los atributos tienen un dominio definido por el tipo de atributo. El dominio determina los valores que puede tomar un atributo. Además, cada muestra o entidad pertenece a una clase o categoría.

El objetivo de ML es obtener una función que asigne una etiqueta de clase a una nueva muestra no etiquetada, es decir, etiquetar o clasificar una serie de muestras utilizando una de entre varias categorías. Por esta razón, estos métodos se llaman a veces clasificadores.

En el caso que nos ocupa, el conjunto de clases a las que pertenecen las muestras será el conjunto de roles elegido para la anotación. Un ejemplo de una entidad será un argumento de un predicado anotado con el rol agente. Una característica podría ser, por ejemplo, las palabras que forman dicho argumento, o también, la categoría gramatical de las palabras que forman el argumento. De esta manera, ante una muestra no etiquetada, es decir, un argumento para el cual se desconoce su rol semántico, el algoritmo de aprendizaje haciendo uso de las características que se hayan definido, determinará la clase a la que dicho argumento pertenece. O

lo que es lo mismo, asignará un rol a dicho argumento de entre el conjunto de roles semánticos que se esté considerando.

Desde el punto de vista de la forma del aprendizaje las técnicas de ML para abordar una tarea se pueden clasificar en aprendizaje supervisado o no supervisado[1].

En el aprendizaje supervisado (apartado 3.1.1) el algoritmo determina la clase a la que pertenece una muestra nueva a partir de un conjunto de muestras etiquetadas con la clase a la que cada una de ellas pertenece. A este conjunto de muestras etiquetadas se le denomina conjunto de aprendizaje o entrenamiento. Al conjunto de muestras nuevas que se pretenden clasificar se le denomina conjunto de test.

El elevado coste de preparación de conjuntos de aprendizaje representativos ha producido la necesidad de realizar el aprendizaje a partir de corpus pequeños, dando lugar a los sistemas semi-supervisados (apartado 3.1.2), o incluso de corpus no anotados, caso de sistemas no supervisados (apartado 3.1.3).

3.1.1 Aprendizaje automático supervisado

La tarea en el aprendizaje automático supervisado es predecir el valor de la función que asigna la etiqueta de clase para cualquier objeto de entrada válido, después de haber visto un número de ejemplos de entrenamiento. Para alcanzar este objetivo el proceso de aprendizaje tiene que generalizar los datos presentados a situaciones no vistas de un modo razonable, vía inducción.

A fin de resolver el aprendizaje automático supervisado se han de considerar varios pasos[2]:

[1] Existen múltiples taxonomías de las técnicas de ML, por ejemplo, dependiendo del tipo de conocimiento a adquirir se puede hablar de aprendizaje simbólico o subsimbólico; o dependiendo de las técnicas empleadas, se puede hablar de sistemas basados en técnicas estadísticas y sistemas basados en razonamiento inductivo. Todas estas dimensiones se entrecruzan de forma que es difícil establecer clases bien definidas y disjuntas (Forsyth, 1988; Michalski, 1983; Kodratoff, 1988; Langley, 1996; Màrquez, 2002).

[2] http://en.wikipedia.org/wiki/Supervised_learning Consultado en marzo 2008

1. Reunir un conjunto de entrenamiento, es decir, obtener un corpus anotado.
2. Determinar el conjunto de características que van a describir a las muestras de cada una de las clases. Normalmente, cada muestra se transforma en un vector de características que describen al objeto.
3. Determinar el correspondiente algoritmo de aprendizaje.
4. Ejecutar el algoritmo de aprendizaje sobre los datos de entrenamiento, ajustando los parámetros del algoritmo de aprendizaje a fin de optimizar el rendimiento sobre un subconjunto de los datos de entrenamiento.
5. Medir el rendimiento del algoritmo sobre un conjunto de test diferente al conjunto de entrenamiento.

Aunque la cantidad de algoritmos de aprendizaje automático supervisado utilizados en PLN es muy variada (ver cuadro 3.1, y para obtener mayor detalle sobre los algoritmos indicados consultar (Màrquez, 2002)), destacan principalmente:

- los algoritmos basados en el principio de máxima entropía, especialmente adaptados al área de PLN;
- los algoritmos basados en memoria[3], destacables por su no necesidad de generalización;
- algoritmos como las máquinas de soporte vectorial (en inglés, *support vector machines -SVM-*)
- los denominados clasificadores o separadores lineales, como Winnow, que han mostrado buenos resultados para la tarea de anotación de roles semánticos (más información sobre esta tarea en capítulo 4).

Independientemente del algoritmo concreto utilizado, dentro del aprendizaje automático supervisado, existen diferentes aproximaciones que intentan estimar cómo de bien ejecuta el modelo

[3] También denominados algoritmos basados en ejemplos, basados en similitud, basados en casos, etc.

Algoritmo	Sistemas desarrollados
Aprendizaje de reglas	CN2 (Clark & Niblett, 1989) Ripper (Cohen & Singer, 1996) FOIL (Quinlan, 1990) PFoil-CNF y PFoil-DNF (Mooney, 1995)
Reglas de transformación	TBL (Brill, 1995) LazyTBL (Samuel, 1998)
Estadístico	Bayes (Duda & Hart, 1973) Modelos ocultos de Markov (Rabiner, 1990)
Árboles de decisión	C4.5 (Quinlan, 1993)
Lineal	Perceptron (Rosenblatt, 1959)
Boosting	AdaBoost y sus variantes (Freund & Schapire, 1997; Freund & Schapire, 1996; Schapire & Singer, 1999)

Cuadro 3.1. Otros algoritmos de aprendizaje supervisado utilizados en PLN

que se ha aprendido a partir de los datos de entrenamiento, sobre datos aún no vistos. Entre estos algoritmos destacan, por ejemplo, los algoritmos de validación cruzada.

A continuación se presentarán los algoritmos de aprendizaje más utilizados en PLN: máxima entropía, aprendizaje basado en memoria, máquinas de soporte vectorial, y clasificadores lineales. Además, se analizarán las principales características de los algoritmos de validación cruzada.

Máxima entropía. Máxima Entropía (ME) proporciona un marco de trabajo para clasificación capaz de integrar muchas fuentes de información heterogéneas (Manning & Schütze, 1999). Los modelos de probabilidad de ME han sido aplicados con éxito a diferentes tareas de PLN, como el análisis morfológico o la detección de límites de oraciones (Ratnaparkhi, 1998).

Un clasificador obtenido por medio de una técnica de ME consta de un conjunto de parámetros o coeficientes los cuales son estimados durante el procedimiento de optimización. Cada coeficiente está asociado a una característica observada en los datos de entrenamiento. El principal propósito es obtener la distribución de probabilidad que maximiza la entropía, esto es, se asume máxima ignorancia y no se considera nada aparte de los datos de entrena-

miento de tal manera que no se induce ningún conocimiento que no esté propiamente en los datos. Algunas ventajas de utilizar ME son: i) aún con características pobres se puede aplicar con precisión, ii) permite representar sin restricciones el conocimiento del problema específico en forma de características (Ratnaparkhi, 1998)

Supongamos un conjunto de contextos X y un conjunto de clases C. En los modelos de ME basados en probabilidad condicional, la función $cl : X \rightarrow C$ elige la clase c con la probabilidad condicional más alta en el contexto x: $cl(x) = \arg\text{máx}_c\, p(c|x)$. Cada característica se calcula mediante una función que está asociada con la clase específica c', y toma la forma de la ecuación (3.1), donde $cp(x)$ es alguna característica observable en el contexto[4]. La probabilidad condicional $p(c|x)$ se define en la ecuación (3.2), donde α_i, representando a los coeficientes objetivo del proceso de optimización y, por tanto, el resultado, finalmente, del aprendizaje, es el parámetro o peso de la característica i; K es el número de características definidas; y $Z(x)$ es una constante para asegurar que la suma de todas las probabilidades condicionales para este contexto es igual a 1.

$$f(x,c) = \begin{cases} 1 \text{ if } c' = c \text{ and } cp(x) = true \\ 0 \text{ otherwise} \end{cases} \tag{3.1}$$

$$p(c|x) = \frac{1}{Z(x)} \prod_{i=1}^{K} \alpha_i^{f_i(x,c)} \tag{3.2}$$

Aprendizaje basado en memoria. Históricamente, los algoritmos basados en memoria o en ejemplos, son descendientes del algoritmo del k-vecino más cercano (en inglés, k-*Nearest Neighbour* -k-NN-).

El aprendizaje basado en ejemplos es un tipo de aprendizaje automático supervisado a partir de ejemplos que se basa en

[4] La aproximación de ME no está limitada a funciones binarias, pero uno de los procedimientos de optimización más conocido utilizado para la estimación de los parámetros, denominado *Generalized Iterative Scaling*, utiliza esta característica.

Universitat d'Alacant
Universidad de Alicante

memorizar todos los ejemplos de aprendizaje tal cual están, sin necesidad de intentar generalizar ninguna regla ni representación más concisa. Para clasificar un nuevo ejemplo, el procedimiento a seguir será obtener de la memoria de ejemplos el conjunto de ejemplos más parecidos al que estamos intentando clasificar y asignar la clase que sea la mayoritaria en ese grupo (Màrquez, 2002).

En esta línea, el grupo ILK de la Universidad de Tilburg ha desarrollado un importante trabajo con el sistema TiMBL. TiMBL (*Tilburg Memory-based Learning Environment*) (Daelemans *et al.*, 2003) es un programa que implementa varios algoritmos de aprendizaje basados en ejemplos con múltiples variantes y opciones. Todos los algoritmos implementados tienen en común que almacenan alguna representación del conjunto de entrenamiento explícitamente en memoria. Durante la fase de test, los nuevos casos son clasificados por extrapolación a partir de los casos almacenados más similares. De esta manera, se puede decir que en este método la generalización se pospone hasta el mismo momento de hacer la clasificación. Para hacer esto, el ejemplo a clasificar $x = \{x_1, ..., x_m\}$ debe compararse con cada uno de los ejemplos guardados $y = \{y_1, ..., y_m\}$ para calcular la distancia entre ellos. Al no generalizar no se prescinde de los ejemplos que representan las excepciones del dominio a tratar (Màrquez, 2002).

Como cualquier sistema basado en ejemplos, TiMBL está formado por dos componentes:

- El componente de aprendizaje (en inglés, *learning component*), es el encargado guardar los ejemplos de entrenamiento en memoria. Un ejemplo de entrenamiento consta de un vector de longitud fija de n pares característica-valor, y un campo en el que se indica la clasificación del vector característica-valor concreto.
- El componente de interpretación (en inglés, *performance component*), está basado en similitud. El resultado del componente de aprendizaje se utiliza como base para mapear la entrada a la salida, lo que generalmente se traduce en una clasificación.

Universitat d'Alacant
Universidad de Alicante

Durante la clasificación, ejemplos de test no vistos se presentan al sistema. La similitud entre el ejemplo nuevo y todos los ejemplos en memoria se calcula utilizando la distancia métrica. La extrapolación se hace asignando la categoría más frecuente dentro del conjunto de ejemplos más similares (k-NN), como la categoría del nuevo ejemplo. En el caso de empate entre categorías, es necesario aplicar algún método de resolución.

La distancia métrica $\triangle(X, Y)$ (ecuación 3.3) es la distancia entre las instancias X e Y, representada por n características, y δ es la distancia por característica. La distancia entre dos patrones es simplemente la suma de la diferencias entre las características. El algoritmo k-NN con esta métrica es denominado IB1.

$$\triangle(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{3.3}$$

donde

$$\delta(x_i, y_i) = \begin{cases} |\frac{x_i - y_i}{max_i - min_i}| & \text{si numérico, si no} \\ 0 & \text{si } x_i = y_i \\ 1 & \text{si } x_i \neq y_i \end{cases} \tag{3.4}$$

A fin de romper los casos de empate en situaciones en las que más de una categoría sea la más frecuente dentro del conjunto de ejemplos, se incrementa el valor del parámetro k en uno, y los nuevos vecinos más cercanos en esta nueva distancia $k + 1$ se añaden al actual conjunto de vecinos más cercanos. Si el empate persiste, se selecciona la etiqueta de clase con el mayor conjunto de ocurrencias total en los datos de entrenamiento. Por defecto el número de vecinos utilizado en la extrapolación es 1.

En un intento de aliviar el coste de la clasificación, TiMBL ofrece la posibilidad de utilizar otros algoritmos de clasificación diferentes de IB1, como la optimización basada en árboles de decisión, denominada IGTREE; o un híbrido entre IB1 e IGTREE2, entre otros.

Universitat d'Alacant
Universidad de Alicante

También es posible utilizar distancias métricas diferentes y más complejas como MVDM o *Jeffrey*, entre otras. Incluso es posible hacer uso de información sobre el comportamiento de las características asignando pesos a las mismas mediante el uso del peso *Information Gain*, o de la estadística *chi-squared*, por ejemplo.

Otros sistemas que hacen uso de algoritmos basados en ejemplos son KENMORE (Cardie, 1996) o los trabajos de Cardie (1993).

Máquinas de soporte vectorial. Las máquinas de soporte vectorial (en inglés, *Support Vector Machines* -SVM-) son una propuesta desarrollada por Vladimir Vapnik y su equipo en los laboratorios AT&T (Cortes & Vapnik, 1995), cuya idea principal es separar las clases con un hiperplano que maximize el margen entre ellas.

En su forma más básica, dos dimensiones y clases linealmente separables, se denominan SVM lineales (LSVM). Los LSVM llevan a cabo una clasificación entre dos clases de puntos, encontrando para ello una superficie de decisión determinada por ciertos puntos del conjunto de entrenamiento, denominados vectores de soporte. Para ello, aprenden el hiperplano lineal que separa con un margen máximo el conjunto de ejemplos positivos del conjunto de ejemplos negativos. El margen se define como la distancia del hiperplano a los ejemplos positivos y negativos más próximos a él. Los vectores de soporte son los elementos críticos de los datos de entrenamiento puesto que son los puntos de datos ubicados más cerca de la superficie de decisión, es decir, los puntos que determinan el margen. Mover los vectores de soporte supone mover la superficie de decisión.

De manera más simple, el objetivo es establecer la ecuación de un hiperplano que divida los datos de entrenamiento dejando todos los datos de la misma clase en el mismo lado mientras se maximiza la distancia mínima entre cualquiera de las dos clases y el hiperplano.

De forma gráfica, la figura 3.1 muestra un ejemplo en un espacio de dos dimensiones y clases linealmente separables. Los ele-

mentos de las clases son representados por círculos y cuadrados, marcando con color más oscuro los correspondientes a vectores de soporte. Los diferentes hiperplanos representados en la figura como líneas continuas separan en un espacio de dos dimensiones. El hiperplano óptimo, representado con línea discontinua, es el correspondiente al margen óptimo.

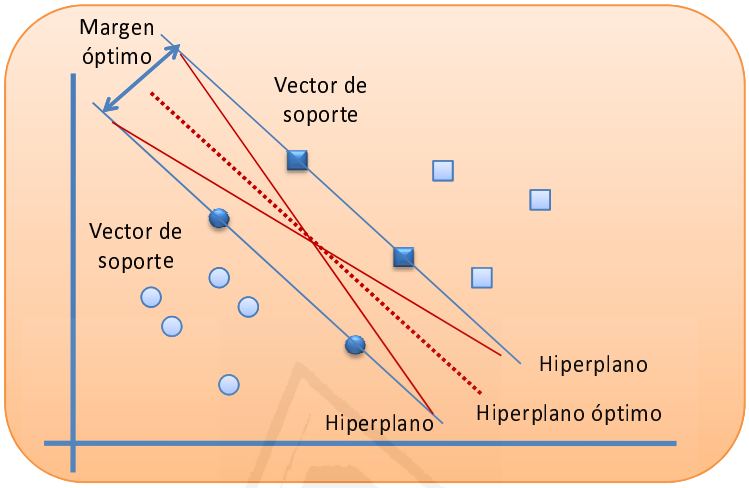


Figura 3.1. Ejemplo de un problema linealmente separable en un espacio de dos dimensiones.

De manera formal (Cortes & Vapnik, 1995), dado el conjunto de entrenamiento de la ecuación (3.5)

$$(y_1, x_1), ..., (y_n, x_n), y_i \epsilon \{-1, 1\} \tag{3.5}$$

se dice que dicho conjunto es linealmente separable si existe un vector w y un escalar b tales que las desigualdades de la ecuación (3.6) son válidas para todos los elementos del conjunto de entrenamiento.

$$w.x_i + b \geq 1 \text{ si } y_i = 1$$
$$w.x_i + b \leq -1 \text{ si } y_i = -1$$

o lo que es lo mismo

$$y_i(w.x + b) \geq 1, i = 1, ..., n \tag{3.6}$$

El hiperplano óptimo

$$w_0.x + b_0 = 0 \tag{3.7}$$

es el único que separa los datos de entrenamiento con un margen máximo: lo que determina la dirección $w/|w|$, donde la distancia entre las proyecciones de los vectores de entrenamiento de las dos clases es máxima. La distancia $\rho(w, b)$ viene dada por la ecuación (3.8)

$$\rho(w, b) = \min_{x:y=1} \frac{x.w}{|w|} - \max_{x:y=-1} \frac{x.w}{|w|} \tag{3.8}$$

El hiperplano óptimo (w_0, b_0) es el argumento que maximiza la distancia. Por lo que de las ecuaciones (3.6) y (3.8) se deduce

$$\rho(w_0, b_0) = \frac{2}{|w|} = \frac{2}{\sqrt{w_0.w_0}} \tag{3.9}$$

Los vectores x_i para los cuales $y_i(w.x_i + b) = 1$, son los denominados vectores de soporte.

A pesar de que el algoritmo básico induce separadores lineales, los datos de entrenamiento no siempre son linealmente separables. En ese caso, dicho algoritmo puede ser extendido a superficies de separación no lineales mediante las denominadas *kernel functions*. Estas funciones permiten transformar el espacio de atributos de entrada en un espacio de características de dimensionalidad mucho mayor. Para ello se mapean los puntos de entrada en puntos de características y se busca el hiperplano de separación óptimo en el correspondiente espacio de características. Los vectores de soporte pasan a ser denominados vectores de margen. En este espacio de alta dimensionalidad se construye un clasificador lineal, que una vez devuelto al espacio original proporciona, por ejemplo, clasificadores polinómicos.

Gráficamente la figura 3.2 muestra cómo la función ϕ mapea los puntos de entrada de forma que sea posible la construcción de un clasificador lineal.

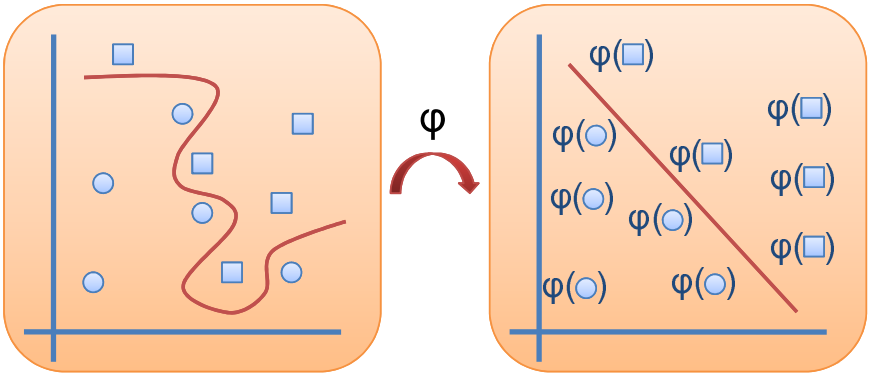


Figura 3.2. Ejemplo de un problema linealmente no separable en un espacio de dos dimensiones.

Una vez que los SVM han sido entrenados, la fase de test consiste simplemente en determinar en qué lado de la superficie de decisión se ubica un determinado dato de test y asignarle la etiqueta de clase correspondiente (Burges, 1998).

Clasificadores lineales. En este tipo de modelos, como por ejemplo Winnow, el clasificador, normalmente binario, aprende un conjunto de pesos para los atributos del espacio de entrada, que son indicadores de su grado de importancia. Para clasificar realizan una combinación lineal del conjunto de atributos (típicamente una suma ponderada de los pesos asociados a los atributos que aparecen en el ejemplo a clasificar) y asignan la clase positiva si el resultado supera un determinado umbral de confianza. En caso contrario asignan la negativa.

Los algoritmos de aprendizaje basados en una actualización multiplicativa de los pesos han demostrado tener propiedades muy buenas cuando son aplicados en dominios de gran dimensionalidad, con presencia de errores en el conjunto de aprendizaje, y

especialmente, cuando los propios ejemplos se distribuyen de manera muy dispersa en el espacio de atributos, es decir, cuando dependen solamente de un conjunto muy pequeño de atributos (Màrquez, 2002).

Su característica fundamental es que el número de ejemplos que requiere para aprender la función objetivo crece linealmente con el número de atributos relevantes y sólo logarítmicamente con el número total de atributos.

Dentro de este tipo de clasificadores destaca la arquitectura SNoW (Roth, 1998), la cual consiste en una red de separadores lineales en el espacio de atributos que usa el algoritmo Winnow en cada nodo y que permite un tipo de aprendizaje robusto, eficiente, incremental y adaptativo (Màrquez, 2002).

La arquitectura de SNoW es una red de umbrales en la que los nodos en la primera capa de la red son asignados a características de entrada, dadas las sentencias de entrada. Los nodos objetivo son representados por nodos en un segundo nivel. Los enlaces del primer al segundo nivel tienen pesos. Cada nodo objetivo es en realidad definido como una función lineal de los nodos de niveles más bajos.

Un nodo objetivo puede verse como una red autónoma, aunque todos los nodos objetivos se alimentan de la misma entrada. La red es dispersa en el sentido de que un nodo objetivo no necesita estar conectado a todos los nodos del nivel de entrada. Por ejemplo, no necesita estar conectado a nodos de entrada que nunca estaban activos con él en la misma oración, o puede decidir, durante el entrenamiento desconectarse por si mismo de alguna de las entradas no relevantes.

El aprendizaje en SNoW se hace de modo *online*. Cada ejemplo es tratado autónomamente por cada subred objetivo. Cada ejemplo etiquetado es tratado como positivo para el nodo objetivo correspondiente a su etiqueta, y negativo para el resto. En realidad, cada ejemplo etiquetado es utilizado una vez por todos los nodos para refinar su definición en término de los otros y es entonces descartado.

Universitat d'Alacant
Universidad de Alicante

En tiempo de predicción, dada una sentencia de entrada la cual activa un subconjunto de los nodos de entrada, la información propaga a través de todas las subredes, determinando la predicción aquella que produzca la actividad más alta.

Validación cruzada. Independientemente del algoritmo concreto utilizado, dentro del aprendizaje supervisado, existen diferentes aproximaciones que intentan estimar cómo de bien ejecuta el modelo que se ha aprendido a partir de unos datos de entrenamiento, sobre datos aún no vistos. Entre estos algoritmos se encuentran los algoritmos de validación cruzada (en inglés, *cross validation*).

La versión más simple de validación cruzada separa los datos en dos conjuntos, llamados conjuntos de entrenamiento y conjuntos de test. El problema de este método es la alta varianza, puesto que la evaluación puede ser significativamente diferente dependiendo de cómo se haga la división.

A fin de mejorar el método anterior se propone el método denominado *k-fold cross validation*. En *k-fold cross validation* el conjunto de aprendizaje se divide en k particiones mutuamente excluyentes, conteniendo todas aproximadamente el mismo número de muestras, de manera que el proceso de entrenamiento y test se repite k veces. A partir de las k particiones se obtienen k clasificadores, utilizando como conjunto de aprendizaje para el clasificador i-ésimo todas las particiones menos la partición i-ésima, y el error se estima sobre las muestras de la partición no utilizada en el aprendizaje. Por último, el error se obtiene como la media de los errores de los k clasificadores. La ventaja de este método es que ya no importa tanto cómo estén divididos los datos. Su gran desventaja es, por un lado, el elevado coste computacional que supone la ejecución del proceso k veces, y por otro, la dificultad de determinar el mejor k, lo cual depende de la cantidad de ejemplos y de la distribución de clases.

Cuando k es igual al tamaño de las muestras se habla de *leave-one-out cross validation* (LOOCV). LOOCV supone utilizar una única muestra del conjunto original para validación de los datos, y las muestras restantes como datos de entrenamiento. Este pro-

ceso se repite de manera que cada muestra es utilizada una vez como datos de validación. Al igual que antes, el error final se medirá atendiendo a la media del error de todas las ejecuciones. La desventaja de este método es su alto coste computacional, superior al de *k-fold cross validation*.

3.1.2 Aprendizaje automático semi-supervisado

El aprendizaje automático semi-supervisado hace uso tanto de datos de entrenamiento etiquetados como no etiquetados para generar una función o clasificador apropiado. Normalmente, la cantidad de datos etiquetados es pequeña frente a los no etiquetados.

Conceptualmente, estos métodos se basan en la gran redundancia del lenguaje natural para poder aprender iterativamente patrones de comportamiento parecidos a los que se dispone en el pequeño conjunto de aprendizaje a partir de un corpus no etiquetado y comparativamente mucho mayor (Màrquez, 2002).

Ejemplos de aprendizaje automático semi-supervisado son los algoritmos de expectación máxima, *co-training*, *bootstrapping* o máquinas de soporte vectorial transductivo (en inglés, *transductive support vector machines*).

A continuación se analizarán las principales características de los algoritmos de *bootstrapping* y *co-training*.

Bootstrapping. Las técnicas de estimación basadas en *bootstrap* se proponen para reducir la alta variabilidad que exhibe la validación cruzada en muestras pequeñas, consiguiendo un aumento de eficiencia comparable a un aumento en el tamaño de la muestra, si bien esta mejora se paga en un coste computacional aún mayor.

Se toman r muestras del conjunto de aprendizaje con reemplazamiento, las cuales conformarán el conjunto de aprendizaje. Todas aquellas muestras que no formen parte del conjunto de aprendizaje se tomarán como conjunto de prueba. Al ser tomadas las muestras con reemplazamiento, se sigue una distribución binomial con lo que la probabilidad de que una muestra no sea elegida

después de N intentos es $(1 - \frac{1}{N})^N \approx e^{-1} \approx 0,368$. Por tanto, el número esperado de muestras diferentes que pertenezcan al conjunto de prueba será de $0,368N$, y al conjunto de aprendizaje $0,632N$.

Si $X_1 ... X_n$ son datos de una distribución con función de distribución $F(x)$, el reemplazamiento se refiere a que el algoritmo de *bootstrap* reemplaza ejemplos de F con ejemplos de la función de distribución empírica que es una estimación de F. La figura 3.3 muestra la función de distribución como la línea continua y con línea discontinua la función de distribución empírica. Lo ideal sería poder ampliar el conjunto de aprendizaje con nuevas muestras de F, pero dado que tales muestras se desconocen, los nuevos ejemplos se toman de la función de distribución que es una aproximación de F.

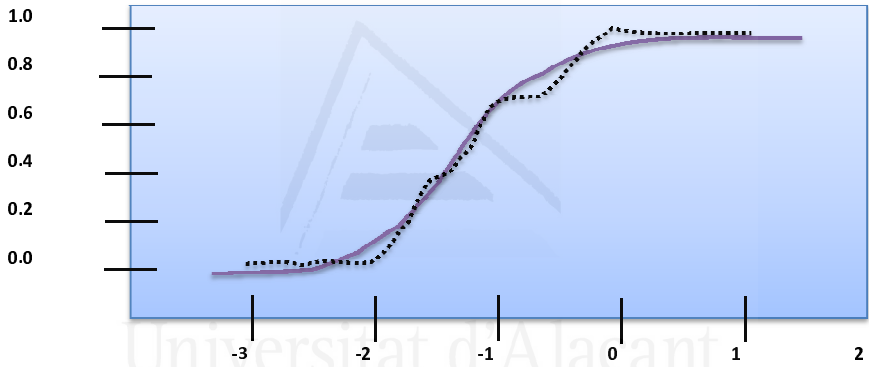


Figura 3.3. Función de distribución empírica como una estimación de la función de distribución verdadera.

Universitat d'Alacant
Universidad de Alicante

La creación del conjunto de prueba y aprendizaje según el esquema anterior se repite b veces, dando lugar a b conjuntos de prueba y aprendizaje. Igual que en la validación cruzada, se obtienen b clasificadores a partir de los conjuntos de aprendizaje y para cada uno un error estimado sobre el conjunto de validación correspondiente.

Co-training. El coentrenamiento (en inglés, *co-training*) es un paradigma de aprendizaje que puede aprender iterativamente a partir de ejemplos etiquetados y ejemplos no etiquetados. En el coentrenamiento, dos clasificadores sencillos se entrenan a partir de un pequeño conjunto anotado, denominado semilla, y clasifican un conjunto no anotado. De esas dos clasificaciones cada clasificador elige n nuevos ejemplos y los incorpora al conjunto anotado para volver a entrenar y clasificar en un proceso iterativo que termina según criterios establecidos de antemano, por ejemplo, hasta que no queden más ejemplos por clasificar (Suárez, 2004; Suárez *et al.*, 2005).

A medida que se ejecutan las iteraciones, el conjunto anotado se va haciendo mayor con las contribuciones de cada clasificador. Así, el clasificador que llamaremos h_1 utiliza en la siguiente iteración los ejemplos que ha clasificado el clasificador, h_2, y viceversa. De esta forma se espera que se reduzca el error cometido por cada clasificador en una tasa significativa.

Los clasificadores son diferentes porque utilizan dos vistas distintas de los mismos datos para aprender. El término vista se puede asimilar a una selección de atributos, es decir cada clasificador entrena con conjuntos distintos de atributos pero sobre los mismos ejemplos.

El problema principal que presenta este algoritmo es la rápida degradación de la precisión a partir de un determinado número de iteraciones (dependiendo del problema de clasificación y de los datos procesados). Algunos investigadores criticaron algunas de sus restricciones, como la necesidad de que las dos vistas sean totalmente independientes, al tiempo que proponen modificaciones y mejoras (Abney, 2002; Collins & Singer, 1999; Nigam & Ghani, 2000).

El coentrenamiento sólo puede aplicarse a problemas de clasificación que cumplan unas determinadas condiciones. De ahí la dificultad de su uso. Dichas condiciones son las siguientes:

1. Cada vista de los datos debe ser suficiente por si misma para realizar la tarea.
2. Los ejemplos anotados por coentrenamiento obtienen esa misma clase con cualquiera de las dos vistas.
3. Las vistas son condicionalmente independientes dada la clase.

3.1.3 Aprendizaje automático no supervisado

En el aprendizaje automático no supervisado todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos. Puesto que las clases se desconocen a priori, el algoritmo debe ser capaz de descubrirlas a base de agrupar ejemplos similares en categorías. Estos métodos de agrupación de datos no etiquetados reciben el nombre de algoritmos de agrupación (en inglés, *clustering*). Un ejemplo es Random Forest (Breiman, 2001).

Clustering es la clasificación de objetos en diferentes grupos, o más exactamente, la partición de los datos en subconjuntos o *clusters*, de manera que los datos en cada subconjunto comparte algún trato común, generalmente una proximidad para alguna medida de distancia definida.

Cualquier proceso de *clustering* supone los siguientes pasos (Jain *et al.*, 1999):

1. Representación de los datos mediante vectores de características multidimensionales. Opcionalmente incluyendo extracción de características y/o selección. La extracción de características se refiere al cálculo de nuevas características de los datos originales. La selección de características es el proceso de identificación del subconjunto de características más efectivo a utilizar en el proceso de agrupación.
2. Definición de una medida de proximidad o similitud entre dos vectores de características, apropiada al dominio de los datos.

Universitat d'Alacant
Universidad de Alicante

Aunque se utilizan una gran variedad de medidas de distancia, la métrica más popular es la distancia euclídea.

3. Agrupación o *clustering*. Existen diferentes aproximaciones de clustering de datos. Según (Jain *et al.*, 1999) tales aproximaciones se pueden clasificar en:

   - Algoritmos jerárquicos. Estos algoritmos producen un conjunto anidado de *clusters*, en el cual cada par de *clusters* es anidado progresivamente a un *cluster* mayor hasta que sólo queda un *cluster*. La representación jerárquica suele ser en forma de árbol.

     Un algoritmo jerárquico construye o deshace una jerarquía de *clusters*, por lo que, dependiendo de la estructura algorítmica se pueden clasificar en:

     - Algoritmos aglomerativos, que empiezan con cada vector de características en un *cluster* diferente y sucesivamente mezclan *clusters* juntos hasta que se satisface un criterio de parada.
     - Algoritmos divisivos, que empiezan con todos los vectores en un único *cluster* y realizan particiones hasta que se alcanza un criterio de parada.

   - Algoritmos particionales. Producen una partición única determinando todos los *clusters* a la vez.

   Por otra parte, dependiendo de la salida del proceso de agrupación, se pueden considerar:

   - Algoritmos de *clustering* duros, los cuales sitúan cada vector de características en un único *cluster*.
   - Algoritmos de *clustering* difusos, los cuales asignan grados de permanencia a varios *clusters* para cada vector de características de entrada.

4. Abstracción de los datos, si es necesario.
5. Evaluación de la salida, si es necesario.

Universitat d'Alacant
Universidad de Alicante

Información detallada sobre diferentes algoritmos de aprendizaje no supervisado se puede encontrar en (Machine Learning, 2002).

3.1.4 Selección de características

En cualquier algoritmo de ML la calidad del conocimiento inducido depende fuertemente de la calidad de las medidas utilizadas en lo que a representatividad del concepto se refiere. Ello es debido a que, por un lado, no se puede utilizar un número infinito de medidas, y por otro algunos de estos algoritmos disminuyen su rendimiento cuando la calidad de estas medidas no es suficientemente buena. Esta disminución de rendimiento se debe principalmente: i) a que el ruido en los ejemplos de entrenamiento se exagera al añadir atributos que aportan ruido, puesto que al clasificador le cuesta distinguir entre señal y ruido; ii) a la existencia de características no relevantes que no aportan información al proceso; y iii) a características redundantes que incluyen la misma información que otras.

Por ello un problema importante es la selección de los atributos o características[5] más relevantes para la inducción del conocimiento. La selección de características opera sobre los datos de manera que normalmente reduce su dimensión (reduce el número de características), eliminando información no relevante o redundante, y transformándolo a una forma más apropiada para clasificaciones siguientes. De esta manera, antes de que un algoritmo de aprendizaje se ejecute para hacer predicciones sobre casos nuevos, se debe decidir qué atributos utilizar en esas predicciones y cuáles ignorar.

El problema de la selección se basa en el concepto de relevancia, entendiendo como atributos relevantes en un problema dado, a los que mejor definen el concepto o clase y por tanto, los que deben ser seleccionados. Su motivación es triple: simplificar el clasificador,

[5] También denominado selección de variables o selección de subconjuntos

mejorar la precisión del clasificador y reducir la dimensionalidad de los datos para el clasificador.

La selección de atributos se puede considerar como un problema de búsqueda en un cierto espacio de estados, donde cada estado se corresponde con un cierto atributo o subconjunto de atributos, y el espacio engloba todas los posibles subconjuntos de atributos que se pueden generar. El proceso de selección de atributos puede entenderse como el recorrido de dicho espacio hasta encontrar un estado (atributo o combinación de atributos) que optimice alguna función definida sobre un conjunto de atributos (Lorenzo, 2001). La figura 3.4 muestra el espacio de estados para un conjunto inicial de cuatro características.

Una vez definido el espacio de búsqueda será necesario establecer: i) un punto de inicio para empezar la búsqueda, ii) una estrategia para recorrer el espacio de búsqueda, iii) una función de evaluación de cada subconjunto de atributos, es decir, de cada elemento en el espacio de búsqueda, y iv) un criterio de parada del proceso de búsqueda.

A continuación se verán en detalle cada uno de los aspectos anteriores (para un resumen de las principales opciones ver cuadro 3.2), así como un resumen de los principales métodos desarrollados atendiendo a tales criterios.

Punto de inicio. En primer lugar, es necesario establecer un punto de inicio para empezar la búsqueda. Los dos puntos obvios son, o bien comenzar con el `conjunto completo`, formado por todos los atributos e ir eliminando a medida que avanza el proceso, o bien comenzar por el `conjunto vacío` sin ningún atributo e ir añadiéndolos. Otra opción podría ser elegir un `conjunto aleatorio` y comenzar la búsqueda desde ese punto.

Estrategia de búsqueda. En segundo lugar, habrá que establecer una estrategia para recorrer el espacio. Una posibilidad es la solución `exhaustiva`, que consiste en recorrer todo el espacio de búsqueda. En esta estrategia, si bien se asegura la obtención del conjunto óptimo de características, su complejidad la hace una

Punto de inicio	Conjunto completo
	Conjunto vacío
	Conjunto aleatorio
Estrategia de búsqueda	Exhaustiva
	Secuencial: FS, BE, FSS, BSE
	Aleatoria
Función de evaluación	Filtro
	Envolvente
	Embebida
	Evaluación de atributos individuales
	Evaluación de subconjuntos de atributos
Criterio de parada	Saturación de la función de evaluación
	Umbral
	Límite en el número de atributos
	Límite en el número de iteraciones

Cuadro 3.2. Aspectos a establecer en cualquier proceso de selección de características

opción impracticable si el número de atributos es elevado. Para un conjunto de n atributos el espacio de búsqueda es $2^n - 1$; para la selección del subconjunto de m atributos de forma exhaustiva es necesario comprobar los

$$\binom{m}{n} = \frac{n!}{m!(n-m)!}$$

subconjuntos. En el caso del espacio de estados mostrado en la figura 3.4, la búsqueda exhaustiva supondría recorrer los 15 subconjuntos mostrados.

Para evitar el recorrido de todo el espacio se han definido estrategias que permiten obtener un subconjunto de atributos que no aseguran el óptimo, pero que tienen un valor próximo con respecto a la función de evaluación utilizada. De entre las más utilizadas se encuentran las secuenciales y las aleatorias (Dash & Liu, 1997; Doak, 1994).

Las estrategias `aleatorias` se basan en visitar diferentes regiones del espacio de búsqueda sin un orden predefinido, evitando de esta forma que se pueda obtener un óptimo local de la fun-

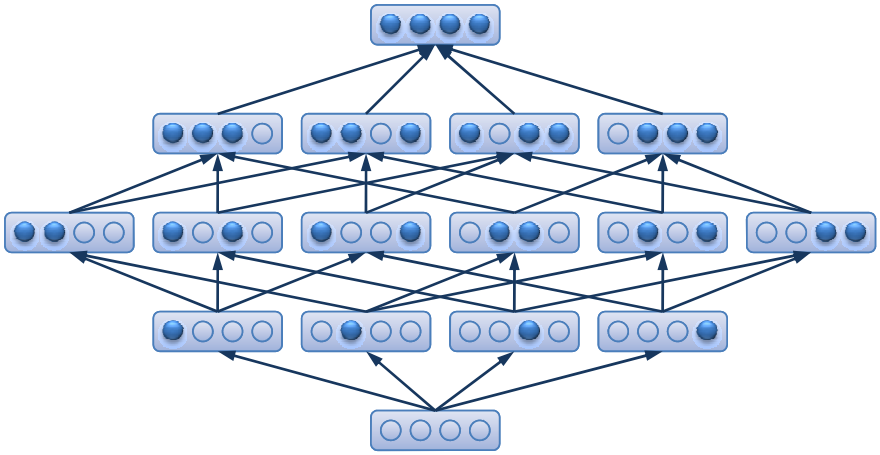


Figura 3.4. Espacio de búsqueda para un conjunto de cuatro características.

ción de evaluación de un determinado subconjunto de atributos (Lorenzo, 2001).

En las estrategias `secuenciales`, también denominadas heurísticas, o bien se van añadiendo iterativamente nuevos atributos a los ya seleccionados, o bien se van eliminando del conjunto inicial. La primera estrategia se denomina selección secuencial hacia adelante (en inglés, *forward selection* (FS) o *forward hill climbing*), y la segunda eliminación secuencial hacia atrás (en inglés, *backward elimination* (BE) o *backward hill climbing*). Los algoritmos correspondientes a ambas estrategias se muestran en las figuras 3.5 y 3.6, respectivamente. FS implica un menor coste computacional que BS al comenzar con conjuntos de menor número de atributos. Por otro lado, BS en principio, puede obtener mejores resultados que FS puesto que comenzar con el conjunto completo de atributos supone considerar todos los atributos, cosa que no se llega a hacer en FS. Un inconveniente de ambos métodos es que no es posible la vuelta atrás ya que una vez se ha añadido un atributo se conserva hasta el final de la búsqueda.

Existen extensiones de estos algoritmos que permiten inclusión y eliminación de atributos en cada paso. Para el caso de

1. Calcular todos los valores parciales para cada atributo independiente
2. Seleccionar el atributo que mejor valor parcial obtiene
3. Calcular todos los valores parciales para todas las combinaciones con los atributos restantes
4. Volver al paso 2

Figura 3.5. Algoritmo de búsqueda *forward selection.*

1. Calcular todos los valores parciales para cada combinación de n-1 atributos
2. Eliminar el atributo que menor valor parcial obtiene
3. Calcular todos los valores parciales para todas las combinaciones de n-1 atributos con los restantes
4. Volver al paso 2

Figura 3.6. Algoritmo de búsqueda *backward elimination.*

la selección secuencial hacia adelante (en inglés, *forward stepwise selection* -FSS-), el procedimiento consiste en añadir en cada paso el atributo que provoque un mayor incremento de la función de evaluación y luego comenzar un proceso de eliminación condicional. Esta eliminación supone ir extrayendo atributos de forma que la cardinalidad del conjunto se vaya reduciendo, siempre que la función de evaluación para cada dimensión obtenida por eliminación de un atributo sea mayor que la que se obtuvo por adición de un atributo. El procedimiento para la eliminación secuencial hacia atrás (en inglés, *backward stepwise elimination* -BSE-) es simétrico. La descripción de estos algoritmos se puede encontrar en diferentes libros sobre estadística (Draper & Smith, 1981; Neter *et al.*, 1990).

Función de evaluación. En tercer lugar, es imprescindible establecer una función de evaluación de cada subconjunto de atributos. Esta medida de evaluación estará definida para un conjunto de atributos y deberá medir la capacidad discriminante del conjunto de atributos para distinguir entre las diferentes clases definidas en el problema.

Existen diferentes taxonomías de estas medidas de evaluación. Tradicionalmente parece que todas ellas coinciden en dos algoritmos principales: filtro y envolvente (en inglés, *wrapper*) (John *et al.*, 1994; Kohavi & John, 1997). En el primer tipo de algoritmos, los denominados `filtro`, el proceso de selección se realiza como un preproceso independiente al proceso de clasificación. En función de características generales del conjunto de entrenamiento se seleccionan o filtran unas características y se excluyen otras, de ahí su nombre. Ver figura 3.7.

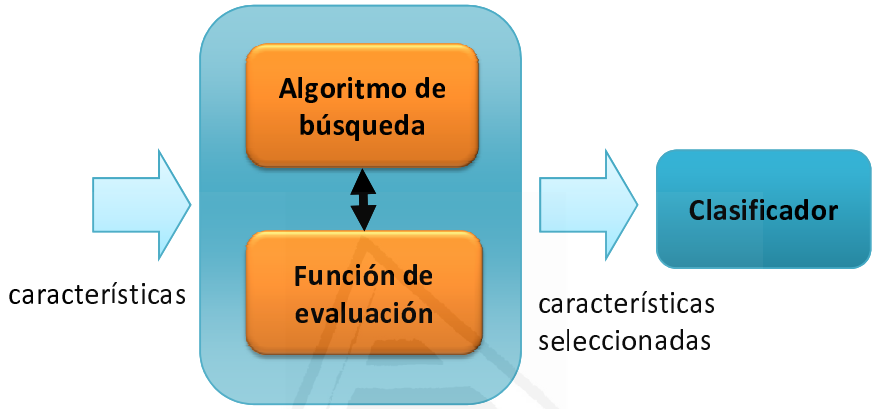


Figura 3.7. Modelo filtro para selección de características.

Los algoritmos `envolventes` hacen uso del proceso de clasificación para evaluar la calidad de cada conjunto de atributos seleccionados en cada momento. En este caso, el algoritmo de aprendizaje se ejecuta sobre los datos de entrenamiento y se utiliza la precisión, o cualquier otra medida, de la clasificación resultante para evaluar el conjunto de características. Ver figura 3.8.

La desventaja que los algoritmos *wrapper* presentan frente a los algoritmos tipo filtro, es el coste computacional que supone ejecutar el algoritmo de aprendizaje para cada subconjunto de características. Ahora bien, el hecho de que las aproximaciones de tipo filtro ignoren por completo los efectos del subconjunto de características seleccionado en el algoritmo de aprendizaje, su-

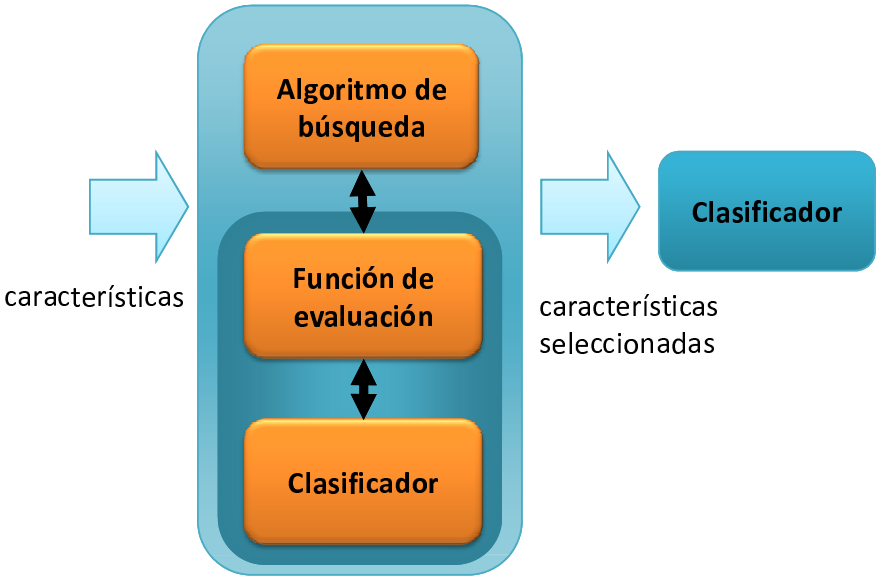


Figura 3.8. Modelo *wrapper* para selección de características.

pone una gran desventaja para estas aproximaciones. De hecho varios investigadores han demostrado la mejora en utilización de métodos *wrapper* frente a filtro (Vafaie & Jong, 1993; John *et al.*, 1994).

Actualmente, la situación ha cambiado desde los sistemas iniciales con tan sólo unas 40 características a sistemas con cientos de miles de características. Esto ha dado lugar a la aparición de una nueva clase especial de algoritmos de selección de tipo *wrapper*, los algoritmos `embebidos`. Estos algoritmos embebidos realizan selección de características durante el proceso de entrenamiento por lo que generalmente, si bien son específicos para un algoritmo de aprendizaje automático dado (Guyon & Wlisseeff, 2003), suponen una mejora respecto al coste computacional.

Desde otro punto de vista puede interesar clasificar los algoritmos en aquellos que evalúan atributos `individuales` y en los que evalúan `subconjuntos de atributos` (Hall & Holmes, 2000). La evaluación individual es menos cara, computacionalmente hablan-

do, puesto que el proceso de evaluación valora características individuales y les asigna pesos de acuerdo a su grado de importancia en la clase. Sin embargo, esta aproximación es incapaz de detectar características redundantes puesto que es probable que estas características tengan pesos similares. Por otra parte, aunque la evaluación de subconjuntos puede manejar tanto redundancia como relevancia de características, en esta aproximación las medidas de evaluación se definen para un subconjunto de características mostrando, por tanto, un alto coste computacional (Lee *et al.*, 2006).

Ejemplos de funciones de evaluación son la distancia euclídea; la distancia de Mantarás (de Mantarás, 1991); la entropía; la ganancia normalizada, la cual determina la ganancia de información de una característica eligiendo aquellas características con mayor ganancia (Duda *et al.*, 2001); medidas de dependencia, las cuales cuantifican la capacidad de predecir el valor de una variable a partir del valor de otra variable; medidas de consistencia (Dash *et al.*, 2000), las cuales calculan el ratio de inconsistencia sobre el conjunto de datos para un conjunto de características dado; medida GD (Lorenzo *et al.*, 1997), la cual intenta recoger las posibles interdependencias que existan entre los atributos; la medida propuesta por (Liu *et al.*, 1998), que garantiza que la búsqueda de características relevantes es completa (garantiza subconjunto óptimo) pero no exhaustiva; Gini (Breiman, 2001); o las medidas comparadas en (Forman, 2003) como *Chi-Squared*, *odss ratio*, *probability ratio*, *random*, F1-medida, *odds ratio numerator*, precisión, o *Bi-Normal Separation* (BNS).

Universitat d'Alacant
Universidad de Alicante

Criterio de parada. Por último, hay que establecer un criterio de parada que permita determinar cuándo se ha encontrado el conjunto de atributos para los que la función de evaluación da el valor óptimo. No fijar un criterio de parada supone que para la búsqueda exhaustiva se recorra todo el espacio de búsqueda, con el consiguiente coste computacional que ello supone, aunque con ello se pueda obtener el subconjunto para el cual la función de evaluación utilizada da el valor óptimo. Una opción para fijar este criterio es considerar que la función utilizada para medir la cali-

dad de los atributos seleccionados sufre un proceso de saturación cuando encuentra el mejor conjunto de atributos. En este caso se detiene la búsqueda cuando a partir de un determinado estado, el valor de la función utilizada no mejora sustancialmente para los estados sucesores. Otros criterios podrían ser por ejemplo, establecer un umbral, o simplemente limitar el número de atributos o el número de iteraciones.

Métodos de selección de características. Atendiendo al esqueleto básico que se acaba de presentar, y que cualquier modelo de selección de características debería definir, se han desarrollado un gran número de métodos diferentes. A continuación se detallan las características básicas de los métodos más destacados (ver cuadro 3.3), y un resumen de otros muchos se muestra en los cuadros 3.4, 3.5 y 3.6.

Método	Punto de inicio	Estrategia de búsqueda	Función de evaluación	Criterio de parada
FOCUS	Vacío	Exhaustiva	Filtro	
Relief	Aleatorio	Secuencial	Filtro	Umbral
DTM	Completo	Secuencial	Filtro	Saturación
BSE-SLASH	Completo	Secuencial	Wrapper	Saturación
BEAM	Aleatorio	Secuencial	Wrapper	Iteraciones
B&B	Completo	Secuencial	Wrapper	Umbral
CFS	Vacío	Secuencial	Subconjunto	Saturación

Cuadro 3.3. Características de los principales métodos de selección de características

Universitat d'Alacant
Universidad de Alicante

- **FOCUS** (Almuallim & Dietterich, 1994) es un método de tipo filtro que busca de manera exhaustiva el subconjunto de atributos de mínimo tamaño suficiente para mantener consistencia con los datos de entrenamiento. Examina todos los subconjuntos de un tamaño dado antes de pasar al siguiente tamaño. Dicho de otra manera, FOCUS examina todos los subconjuntos de características seleccionando el mínimo subconjunto que es suficiente para determinar la clase. Existe una segunda versión denomi-

nada FOCUS-2 (Almuallim & Dietterich, 1994) la cual realiza una búsqueda heurística.

Los métodos FOCUS son una implementación de MIN- FEATURES. Para los casos en los que MIN-FEATURES es inabordable, (Almuallim & Dietterich, 1994) propone la utilización de otros sistemas como *Mutual-Information-Greedy* (MIG), *Simple-Greedy* (SG) y *Wighted-Greedy* (WG). Estos sistemas proporcionan una buena solución, aunque no la óptima. Cada uno de estos algoritmos implementa un procedimiento iterativo donde, en cada iteración, la característica que parece más prometedora se añade a la solución parcial. Esto continua hasta que se encuentra un conjunto suficiente de características. La única diferencia entre los tres algoritmos es el criterio utilizado para seleccionar la mejor característica en cada iteración. MIG selecciona la característica que conlleva la mínima entropía. SIG, partiendo del conjunto de todos los conflictos, elige la característica que cubre el mayor número de conflictos que no están cubiertos todavía. Los conflictos cubiertos por la característica se eliminan de la lista. El proceso se repite hasta que se eliminan todos los conflictos. En SG cada conflicto contribuye en una unidad a la puntuación de cada característica que lo cubre. En WG el incremento depende del número total de características que lo cubran.

- **Relief** (Kira & Rendell, 1992) es también un método de tipo filtro que asocia a cada atributo un peso indicando la relevancia relativa del atributo, con el fin de hacer la distinción de clases. Es un algoritmo aleatorio. Toma muestras aleatoriamente del conjunto de entrenamiento y actualiza los valores de relevancia basándose en la diferencia entre los ejemplos seleccionados y los dos ejemplos más cercanos de la misma clase y de la clase opuesta, seleccionando aquellas características que superen un cierto umbral especificado por el usuario. El proceso es repetido para un número de instancias especificadas por el usuario. Está diseñado para algoritmos booleanos. Extensiones de este sistema son Relief-E y Relief-F (Kononenko, 1994).

- **DTM** (Cardie, 1993) es otro método filtro que utiliza selección de subconjunto para eliminar características no relevantes de un conjunto de datos para ser utilizados con el algoritmo del vecino más cercano. Como una métrica de la utilidad de un atributo, se utilizó C4.5 para obtener un árbol de decisión de un conjunto de entrenamiento y eliminar las características que no aparecieran en el árbol resultante. En Cardie y Howe (1977) se propone una extensión a DTM.
- **BSE-SLASH** (Caruana & Freitag, 1994) es un método de tipo *wrapper*, que realiza una búsqueda heurísticas comenzando por el conjunto completo de características. En cada paso elimina cualquier atributo no utilizado por el proceso de aprendizaje en ese paso. Esto le permite saltar rápidamente a regiones del espacio de búsqueda donde todos los atributos en el subconjunto activo juegan un papel en lo que se está aprendiendo.
- **BEAM** (Aha & R.L.Bankert, 1994) es un método aleatorio, el cual consiste en una extensión de BSS que permite trabajar con un mayor número de características. Surge con el objetivo de reducir el coste computacional de BSS puesto que éste comienza con el conjunto completo de características. BEAM aleatoriamente muestra el espacio de características para un número fijo de iteraciones y empieza con el subconjunto de características que mejor se comporte en estas iteraciones. Mantiene una cola de tamaño fijo de los estados que mejor se comportan ordenados en orden decreciente de precisión. La cola es actualizada cada vez que un estado es seleccionado y evaluado. La evaluación puede ser completa, si se evalúan todos los subconjuntos de un número más pequeño de características, o gradual si sólo evalúa un subconjunto de características. Devuelve la mejor evaluación.
- **B&B**, *Branch and Bound* (Narendra & Fukunaga, 1977), es un método heurístico que comienza con el conjunto completo de características eliminando una característica cada vez. Sin restricciones en expansión de nodos en el espacio de búsqueda podría producir una búsqueda exhaustiva. Sin embargo, si cada nodo es evaluado por una medida, y se establece un límite superior para valores aceptables de tal medida, entonces B&B

retrocede cuando se descubre un nodo no factible. ABB, automático B&B, (Liu *et al.*, 1998) es una extensión en la que la medida para evaluación de los nodos se determina automáticamente.

- **CFS**, *Correlation-based feature selection* (Hall & Holmes, 2000), es el primero de los métodos que evalúa subconjuntos de atributos más que atributos individuales. El corazón del algoritmo es una heurística de evaluación de subconjuntos que tiene en cuenta la utilidad de las características individuales para predecir la clase, junto con el nivel de intercorrelación entre ellas. La heurística asigna puntuaciones altas a subconjuntos que contienen atributos que están altamente correlacionados con la clase y tienen baja intercorrelación unos con otros.

3.2 Enfoques basados en conocimiento

Un sistema basado en conocimiento se puede definir como un sistema que resuelve problemas utilizando una representación simbólica del conocimiento humano.

La arquitectura de un sistema basado en conocimiento de alguna manera refleja la estructura cognitiva y los procesos humanos (ver figura 3.9). Por ello, entre sus componentes fundamentales se encuentra la base de conocimiento, la cual encapsula en algún formalismo de representación, el conocimiento del dominio que debe ser puesto en juego por el sistema para resolver el problema dado.

Entre los diferentes formalismos de representación del conocimiento destacan las reglas, las redes semánticas, los marcos o *frames*, y los basados en lenguajes lógicos, como la lógica de predicados y sus extensiones, la lógica modal o el lambda cálculo (Moreno *et al.*, 1999a).

La principal desventaja de los sistemas basados en conocimiento es el gran trabajo que supone desarrollar y mantener los recursos necesarios y el hecho de que éstos tienen el problema de ser dependientes del dominio y del idioma.

Universitat d'Alacant
Universidad de Alicante

Tipo	Método	Observaciones
Filtro	(Modrzejewski, 1993)	
	POE&ACC (Mucciardi & Gose, 1971)	
	(Dumais *et al.*, 1998)	
	(Caruana & de Sa, 2003)	
	(Stoppiglia *et al.*, 2003)	
	PC (Guyon & Wlisseeff, 2003)	
	(Leite & Rino, 2006)	
	(Lee *et al.*, 2006)	
	(Molina *et al.*, 2002)	
	(Piramuthu, 1998)	
	(Zhang *et al.*, 2006a)	
	(Hermes & Buhmann, 2000)	
	(Duda *et al.*, 2001)	
Wrapper	(Moore & Lee, 1994)	
	Oblivion (Langley & Sage, 1994)	
	(Embrechts *et al.*, 2001)	
	VS-SSVM (Bi *et al.*, 2003)	
	(Hguyen *et al.*, 2006)	
	(Torkkola *et al.*, 2004)	
	(Wu & Zhang, 2004)	
	(Maejía-Lavalle & Arroyo-Figueroa, 2006)	
	RFE-SVM (an dJ. Weston *et al.*, 2002)	
	(Bedo *et al.*, 2006)	Extensión
	(Huang & Kecman, 2005)	Extensión
	(Koppel & Schler, 2004)	Extensión
	R^2W^2 (Weston *et al.*, 2001)	Extensión
Embebidos	Grafting (Perkins *et al.*, 2003)	
	LPSVM (Fung & Mangasarian, 2002)	
	AROM (Cortes & Vapnik, 1995)	
	(Weston *et al.*, 2003)	
	(Li *et al.*, 2004)	Específico SVM
	(Li & Liu, 2006)	Específico SVM
	(Neuman *et al.*, 2005)	Específico SVM
	(Zhou *et al.*, 2003)	Específico ME
	(Malouf, 2002)	Específico ME
	(Zhang *et al.*, 2006b)	Específico ME
	PFS (Zhang *et al.*, 2006b)	
	(Zhou *et al.*, 2003)	
	(Jebara & Jaakkola, 2000)	
	(Zhu *et al.*, 2004)	

Cuadro 3.4. Otros métodos de selección de características (1/3)

Universitat d'Alacant
Universidad de Alicante

Tipo	Método	Observaciones
Aleatorios	(Skalak, 1994)	
	(Doak, 1994)	
	LVF (Liu & Setiono, 1996b)	
	LVS (Liu & Setiono, 1998b)	Variante de LVF
	LVI (Liu & Setiono, 1998a)	Variante de LVF
	LVW (Liu & Setiono, 1996a)	Variante de LVF
	SetCover (Dash *et al.*, 2000)	
	Las Vegas (Brassard & Bratley, 1996)	
	GADistAI (Yang & Honavar, 1998)	
	(Brill *et al.*, 1992)	
	SLAVE (González & Pérez, 1997)	
	FSS-EBNA (Inza *et al.*, 2000)	
	RMHC (Skalak, 1994)	
	(Siedlecki & Skalansky, 1989)	
	(Vafaie & Jong, 1993)	
	(Guerra-Salcedo *et al.*, 1999)	
Exhaustivo	MDML (Sheinvald *et al.*, 1990)	
	AMB&B (Foroutan, 1987)	
	(Davies & Russell, 1994)	
Heurístico	Chi2 (Liu & Setiono, 1995)	
	MIFS (Battiti, 1994)	
	(Lallich & Rakotomalala, 2000)	
	EUBAFES (Scherf & Brauer, 1997)	
	IS (Vafaie & Imam, 1994)	
	C-SEP (Fayyad & Irani, 1992)	
	(Setiono & Liu, 1996)	
	(Setiono & Liu, 1997)	
	(Yang & Moody, 1999)	
	(Koller & Sahami, 1996)	
	CR (Wang *et al.*, 1999)	
	1-R (Holte, 1993)	
	(Kohavi & Frasca, 1994)	
	(Kohavi & John, 1997)	
	RC (Domingos, 1997)	
	RACE (Moore & Lee, 1994)	
	FSV (Bradley & Mangasarian, 1998)	
	(Lorenzo, 2001)	

Cuadro 3.5. Otros métodos de selección de características (2/3)

Universitat d'Alacant
Universidad de Alicante

Tipo	Método	Observaciones
Híbridos	QBB (Dash *et al.*, 2000)	LVF y ABB
	(Legrand & Nicolayannis, 2005)	Filtro y wrapper
	(Frohlich *et al.*, 2003)	Filtro y wrapper
	(Chen, 2003)	Filtro y wrapper
	(Xing *et al.*, 2001)	Filtro y wrapper
	(Ng, 1998)	Filtro y wrapper
	(Das, 2001)	Filtro y wrapper
Otros sistemas	(Law *et al.*, 2004)	Espect. Max.
	Ripper (Cohen & Singer, 1996)	
	(Forman, 2003)	
	WR (an dJ. Weston *et al.*, 2002)	
	SVM-RFE (an dJ. Weston *et al.*, 2002)	
	(Neal, 1998)	
	(Bengio & Chapados, 2003)	
	(Rakotomamonjy, 2003)	
	(Blum & Langley, 1997)	
	(Zhang *et al.*, 2003)	
	GAZBFDR (Yan *et al.*, 2006)	
	(Quixtiano-Xicohténcatl *et al.*, 2006)	
	(Bo & Jonassen, 2002)	
	(Li & Hu, 2006)	
	(Tsamardinos *et al.*, 2006)	

Cuadro 3.6. Otros métodos de selección de características (3/3)

Universitat d'Alacant
Universidad de Alicante

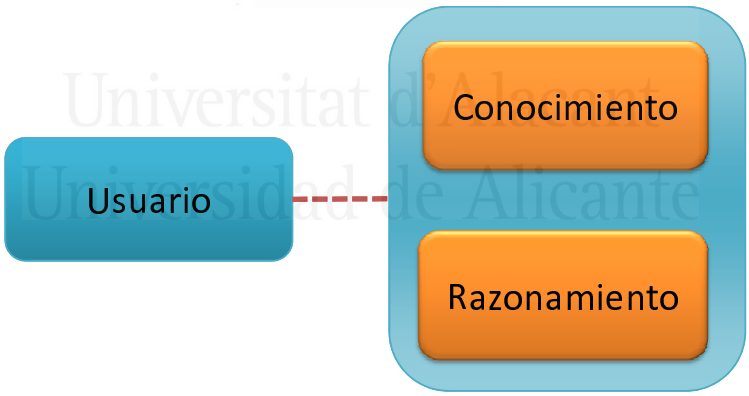


Figura 3.9. Arquitectura básica de un sistema basado en conocimiento.

A continuación se presentarán, de forma breve puesto que tales formalismos no constituyen el objeto central de este trabajo, las características más importantes de los principales formalismos de representación: reglas (apartado 3.2.1), lógica de predicados (apartado 3.2.2) y *frames* (apartado 3.2.3).

3.2.1 Representación basada en reglas

En los sistemas de conocimiento basados en reglas, el conocimiento se puede representar en estructuras como la mostrada a continuación (Moreno *et al.*, 1999b):

`SI` <condición> `ENTONCES` <acción>

Por ejemplo,

`SI` <argumento comienza por "in"> `ENTONCES` <asignar rol de lugar>

3.2.2 Lógica de predicados

La lógica de predicados es un lenguaje formal donde las oraciones bien formadas son representadas mediante un conjunto de variables, constantes, símbolos de predicado, símbolos de función, conectivas, cuantificador existencial y cuantificador universal (Moreno *et al.*, 1999b).

Por ejemplo, considerar la oración (E38)

(E38) John cross the Atlantic ocean

su representación mediante lógica de predicados sería la mostrada en el ejemplo (E39)

(E39) existe(X, tema(atlántico(X)&oceano(X)), past(cross(agente(john),X)))

Universitat d'Alacant
Universidad de Alicante

3.2.3 *Frames*

Uno de los formalismos de representación más utilizados en el análisis semántico, y más concretamente, en la anotación de roles semánticos son los marcos o *frames*.

Un *frame* se define como un conjunto de atributos, llamados huecos o *slots*, con valores asociados y posibles restricciones sobre esos valores. Cada *frame* puede describir alguna entidad del dominio, o puede describir algún objeto particular (Moreno *et al.*, 1999b).

Uno de los ejemplos más representativos dentro de la anotación de roles lo constituye el léxico desarrollado en el proyecto FrameNet comentado en el capítulo 2.

Universitat d'Alacant
Universidad de Alicante

4. Sistemas de Anotación Automática de Roles Semánticos

Los sistemas de anotación de roles semánticos (en inglés, *semantic role labeling* -SRL-) persiguen resolver la ambigüedad que se produce al determinar la agrupación y relaciones entre palabras de una oración. Estos sistemas proponen mecanismos que ayudan a identificar qué conjuntos de palabras de una oración forman cada constituyente, y qué relaciones existen entre esos constituyentes y el predicado, expresando tales relaciones en forma de roles semánticos.

El origen de los sistemas de anotación de roles semánticos se encuentra en sistemas de extracción de información, los cuales empezaron como complejos sistemas basados en reglas diseñados a mano (Hirst, 1998; Richardson *et al.*, 1998), hasta llegar a sistemas más simples estadísticos (Gildea & Palmer, 2002).

Los primeros sistemas propuestos consideraban sólo oraciones sencillas o un vocabulario restringido. Sin embargo, los sistemas actuales realizan la asignación de roles en diferentes lenguas y prácticamente sin limitaciones.

Cada uno de estos sistemas se caracteriza por el corpus utilizado, y por tanto, la lengua para la que han sido definidos, el conjunto de roles utilizado en la anotación, la información requerida para llevar a cabo el proceso de anotación y la estrategia o enfoque seguido en dicha anotación.

En este capítulo se analizarán los sistemas de SRL desarrollados hasta el momento, atendiendo por un lado, a las características comentadas (roles, información utilizada, y corpus), y por otro, al enfoque utilizado para la resolución del problema (ver

Universitat d'Alacant
Universidad de Alicante

capítulo 3), ya sea basado en corpus (apartado 4.1) en cualquiera de sus formas de aprendizaje, supervisado (apartado 4.1.1), semi-supervisado (apartado 4.1.2) o no supervisado (apartado 4.1.3); o basado en conocimiento (apartado 4.2). Además, se realizará un repaso de las principales conclusiones extraídas de las conferencias y competiciones celebradas con el objetivo de medir la capacidad de estos sistemas de anotación de roles (apartado 4.3). En concreto se analizarán las conferencias CoNLL (apartado 4.3.1) y las competiciones Senseval (apartado 4.3.2).

4.1 Enfoques basados en corpus

4.1.1 Aprendizaje automático supervisado

Las primeras propuestas que se realizaron en esta línea desarrollaron varios sistemas cuyo objetivo era asignar roles de caso al estilo de Fillmore, a los constituyentes de una oración. Para ello utilizaron algoritmos de aprendizaje supervisado basados en redes neuronales y la información proporcionada por el análisis sintáctico. Estos sistemas se caracterizaron por poseer varias limitaciones como el hecho de considerar sólo oraciones sencillas con un número limitado de constituyentes y sin cláusulas anidadas, y con un vocabulario restringido. Tal es el caso de propuestas como las de (McClelland & Kawamoto, 1986; John & McClelland, 1990; Miikkulainen & Dyer, 1991; Rosa, 2001). Aunque se realizaron extensiones posteriores que permitieron el tratamiento de oraciones más complejas con múltiples cláusulas, como en (Miikkulainen, 1990; Jain, 1990; Miikkulainen, 1996), la generalización a nuevas oraciones y estructuras seguía siendo limitada.

Otro de los primeros sistemas que intentó asignar de forma automática roles semánticos a los textos del corpus Penn Treebank fue (Blaheta & Charniak, 2000)[1]. En este sistema, la tarea es

[1] En realidad, este trabajo persigue añadir información a los constituyentes de una oración utilizando para ello 20 etiquetas que aportan, tanto información sintáctica como semántica. Para el trabajo que nos ocupa, nos hemos centrado únicamente en las correspondientes a la información semántica.

afrontada de forma parcial puesto que las etiquetas de función del Treebank no incluyen todos los argumentos de la mayoría de los predicados. Los resultados obtenidos utilizando un algoritmo basado en máxima entropía son 80,425 % de precisión, 77,595 % de cobertura, y 78,980 % de medida F.

Tras estos intentos, se desarrolló el `primer sistema` que tuvo como objetivo aprender automáticamente a identificar todos los roles semánticos de una extensa variedad de predicados en textos no restringidos para inglés (Gildea & Jurafsky, 2002). El sistema requiere identificar previo a la tarea y de forma manual, para cada oración: i) una palabra que evoque el *frame* y, ii) el propio *frame*. Este sistema se caracteriza por:

- *Corpus*. Las oraciones del British National Corpus incluidas en la base de datos de FrameNet. Se excluyeron palabras con menos de 10 ejemplos en el corpus.
- *Roles*. Conjunto de roles definidos en FrameNet.
- *Información*. Partiendo del árbol de análisis sintáctico, obtenido de forma automática (Collins, 1997), se extraen características léxicas y sintácticas, incluyendo:
  - El tipo de sintagma de cada constituyente.
  - La función gramatical de cada constituyente. Esta característica tiene dos posibles valores correspondientes al sujeto y objeto del verbo. Además, su aplicación se restringe únicamente a sintagmas nominales.
  - La posición del constituyente en la oración, indicando si se encuentra antes o después del predicado.
  - El camino en el árbol de análisis desde la palabra que evoca el *frame* hasta el constituyente. Este valor se representa como una cadena en la que el primer elemento corresponde a la etiqueta de *PoS* de la palabra que evoca el *frame*, y el último al tipo de sintagma o categoría sintáctica del constituyente marcado como elemento de *frame*. Las etiquetas correspondientes a los verbos se generalizan a VB. La cadena contiene

Universitat d'Alacant
Universidad de Alicante

además indicación de los movimientos hacia arriba y hacia abajo en el árbol. Un ejemplo de esta característica sería la cadena VB↑VP↑VP↑S↓NP.

- Voz de la oración, indicando si se encuentra en activa o pasiva.
- El núcleo del constituyente.

- *Estrategia de anotación.* La tarea se afronta como dos subproblemas. En primer lugar, determinar los constituyentes de una oración y después, asignarles las etiquetas semánticas correctas.
- *Algoritmo de aprendizaje.* Clasificador estadístico que combina probabilidades de distribuciones condicionadas sobre una variedad de subconjuntos de las características. Debido a la dispersión de los datos, no es posible estimar la probabilidad de cada rol paras las características comentadas. Por ello, las probabilidades se calculan para varios subconjuntos de características y se interpolan como una combinación lineal de las distribuciones resultantes. La interpolación se realiza sobre la distribución más específica para la cual los datos están disponibles.
- *Resultados.* Determinar el rol semánticos de constituyentes ya identificados, 82 % de precisión. Identificar constituyentes y sus roles, 64,6 % de precisión y 61 % de cobertura. Si sólo se utiliza la información de análisis sintáctico parcial los resultados bajan a 27,6 % de precision y 22 % de cobertura.

En (Gildea & Palmer, 2002) se utiliza el sistema anterior sobre una versión preliminar del corpus PropBank. Con el fin de obtener resultados comparables a los obtenidos con FrameNet, las oraciones del corpus se analizaron con el analizador sintáctico de Collins y se excluyeron predicados con menos de 10 ejemplos. Los resultados obtenidos utilizando la información del análisis sintáctico automático fueron 79,20 % de precisión en la identificación de roles; y 57,70 % de precisión y 50,00 % de cobertura en el caso de detección de argumentos e identificación de roles. Haciendo uso de la información de análisis sintáctico manual se obtuvo 82,80 % de precisión en la identificación de roles, y 71,10 % de precisión y 64,40 % de cobertura en la detección de argumento y roles. No

excluir predicados con menos de 10 ejemplos supone una ligero empeoramiento de los resultados.

En (Gildea & Hockenmaier, 2003) utilizan también el mismo sistema pero en lugar de partir de oraciones analizadas por el analizador de Collins, utilizan un analizador sintáctico basado en *Combinatory Categorial Grammar* (CCG) (Hochenmaier & Steedman, 2002). Las características fueron adaptadas al nuevo formato de la información. Los resultados obtenidos fueron un 71 % de precisión y 63,1 % de cobertura. Cabe destacar que estos resultados se ven afectados por el el hecho de que en muchos casos no se pueden establecer correspondencias entre las estructuras de constituyentes de CCG y del TreeBank.

Extensiones a la propuesta de Gildea. Utilizando la propuesta de Gildea como base se han desarrollado multitud de sistemas caracterizados por utilizar corpus de entrenamiento diferentes, conjuntos de roles diferentes, añadir más características, utilizar otros algoritmos de aprendizaje, utilizar sólo información sintáctica parcial o por el contrario, añadir algún otro tipo de información, centrados sólo en la tarea de desambiguación del rol, para lenguas diferentes del inglés, para tareas específicas, como desambiguación de preposiciones o roles de nombres, o incluso para dominios específicos como el biomédico.

Respecto al `corpus utilizado` la mayoría de los sistemas han utilizado o bien PropBank, o bien FrameNet, aunque existen variaciones atendiendo principalmente a la lengua utilizada. Destacan:

Universitat d'Alacant
Universidad de Alicante

- Sistemas desarrollados para inglés que han utilizado:
  - Propbank (Chen & Rambow, 2003; Surdeanu *et al.*, 2003; Blunsom, 2004; Punyakanok *et al.*, 2005b; Xue & Palmer, 2004; Ping, 2005; Girju *et al.*, 2004; Toutanova *et al.*, 2005; Ye & Baldwin, 2005; Arcuri, 2006; Liu & Sarkar, 2006; Musillo & Merlo, 2006; Che *et al.*, 2006; Surdeanu *et al.*, 2007; Moschitti *et al.*, 2008).

- FrameNet (Fleischman *et al.*, 2003a; Ohara & Wiebe, 2003; Pado & Boleda, 2004a; Pradhan *et al.*, 2004a).
- Penn Treebank (Ohara & Wiebe, 2002; Ye & Baldwin, 2005).
- Sistemas que han realizado pruebas con diferentes corpus, como los trabajos de (Hacioglu & Ward, 2003; Moschitti, 2004) que muestran resultados tanto con Propbank como con FrameNet; o diferentes trabajos de Moschitti con Pighin y Basili que muestran resultados para PropBank (Moschitti *et al.*, 2008), FrameNet (Moschitti, 2006a) y VerbNet (Pighin & Moschitti, 2007).
- Una combinación de la información de FrameNet con VerbNet y Propbank a fin de que el sistema sea lo más independiente posible del corpus utilizado (Giuglea & Moschitti, 2004; Giuglea & Moschitti, 2006c; Giuglea & Moschitti, 2006a; Giuglea & Moschitti, 2006b).
- Una selección de 37 verbos del corpus Reuters[2] de acuerdo a su frecuencia relativa en el corpus BNC (Busser & Moens, 2003).

- Para el checo, con el sistema (Sgall *et al.*, 2002) que utiliza el corpus nacional checo.
- El chino (You & Chen, 2004) con el corpus Sinica Treebank, (Sun & Jurafsky, 2004), o con una selección de 10 verbos del corpus Penn Chinesse Propbank (Pradhan *et al.*, 2004a).
- El sueco con oraciones extraídas de FrameNet y traducidas (Johansson & Nugues, 2006a).
- El portugués, con un conjunto de oraciones generadas automáticamente mediante un generador de oraciones (Rosa, 2007).
- Sistemas que combinan inglés y alemán, como en (Erk & Padó, 2006), en el que se utilizan los corpus FrameNet para el inglés y SALSA/TIGER para el alemán.

[2] http://about.reuters.com/researchandstandards/corpus/ Consultado marzo 2008

Universitat d'Alacant
Universidad de Alicante

- Sistemas para español y catalán (Surdeanu & Turmo, 2008) que han hecho uso del corpus CESS-ECE.
- O incluso sistemas que han utilizado corpus para dominios específicos, como el corpus BioProp para el dominio biomédico utilizado en (Tsai *et al.*, 2006).

Respecto al `conjunto de roles utilizado`, la mayoría de los sistemas han utilizado los roles propios de cada corpus, si bien algunos de ellos llevan a cabo alguna variación como en el caso de (Hacioglu & Ward, 2003) los cuales realizan un mapeo de los roles de Propbank a un conjunto de 22 roles; (Moschitti, 2004) que reduce a 18 los roles de FrameNet atendiendo a la frecuencia de aparición; (Pado & Boleda, 2004a) que ignora la mayoría de los roles no núcleo de FrameNet; (Tsai *et al.*, 2006) que sólo considera los adjuntos de Propbank; (Busser & Moens, 2003) que utiliza patrones de roles funcionales; o (Rosa, 2007) que utiliza un conjunto de 7 roles: *agent, patient, experiencer, theme, location, cause, value.*

Respecto a la `información utilizada` casi todos los sistemas utilizan la información proporcionada por analizadores sintácticos completos, ya sea manualmente anotado u obtenido con analizadores sintácticos estadísticos, si bien existen algunas excepciones. Además esta información sintáctica suele complementarse con información a otro niveles, como por ejemplo información sobre entidades nombradas o sobre frecuencias de aparición. En concreto,

- Análisis sintáctico completo (Fleischman *et al.*, 2003a; Surdeanu *et al.*, 2003; Moschitti, 2004; Pado & Boleda, 2004a; Xue & Palmer, 2004; Ping, 2005; Pradhan *et al.*, 2004a; You & Chen, 2004; Sun & Jurafsky, 2004; Punyakanok *et al.*, 2005b; Toutanova *et al.*, 2005; Arcuri, 2006; Erk & Padó, 2006; Giuglea & Moschitti, 2006a; Tsai *et al.*, 2006; Che *et al.*, 2006; Surdeanu *et al.*, 2007; Moschitti *et al.*, 2008; Surdeanu & Turmo, 2008).
- Análisis sintáctico parcial (Busser & Moens, 2003; Blunsom, 2004; Punyakanok *et al.*, 2005b; Ye & Baldwin, 2005; Johansson & Nugues, 2006a; Surdeanu *et al.*, 2007).

- *Tree Adjoining Grammar* (Chen & Rambow, 2003; Liu & Sarkar, 2006).
- La información proporcionada por el nivel analítico del corpus en (Sgall *et al.*, 2002).
- La información proporcionada por el análisis morfológico, la información sobre cláusulas y resolución de anáfora (Rosa, 2007).

Además, algunos sistemas añaden información de carácter diferente al sintáctico, como por ejemplo,

- Entidades con nombre (Surdeanu *et al.*, 2003; Surdeanu *et al.*, 2007).
- Sentido del verbo (Pradhan *et al.*, 2005c).
- Información sobre etiquetas de roles ya asignadas (Fleischman *et al.*, 2003a; Hacioglu & Ward, 2003; Ping, 2005; Surdeanu *et al.*, 2007).
- Información sobre frecuencia de roles para cada verbo como en (Blunsom, 2004).
- Información sobre los pares rol-función gramatical de cada predicado (Pado & Boleda, 2004a).
- Información proporcionada por el mapeo entre los corpus FrameNet, PropBank y VerbNet, a través de las clases de verbos de Levin (Giuglea & Moschitti, 2006a).
- Información específica del dominio biomédico (Tsai *et al.*, 2006).
- Información sobre categoría gramatical (Ohara & Wiebe, 2002).

En este aspecto de la información utilizada destaca el sistema de Hacioglu el cual lleva a cabo diferentes pruebas utilizando tanto información sintáctica total como parcial (Hacioglu & Ward, 2003; Hacioglu *et al.*, 2003; Hacioglu, 2004a; Hacioglu, 2004b), y diferentes extensiones a este sistema propuestas en (Pradhan *et al.*, 2003; Pradhan *et al.*, 2004c), el cual utiliza información de análisis sintáctico combinada de diferentes fuentes (Pradhan *et al.*, 2004b; Pradhan *et al.*, 2005b).

Universitat d'Alacant
Universidad de Alicante

Por último, cabe también destacar en algunos de los sistemas el uso de otros recursos diferentes al corpus de entrenamiento, como la TopOntology de EuroWordNet en (Sgall *et al.*, 2002), ComLex y VerbLex en (Girju *et al.*, 2004), o WordNet en (Ohara & Wiebe, 2002).

Respecto al `algoritmo de aprendizaje supervisado` utilizado, destacan como más utilizados máxima entropía y máquinas de soporte vectorial. Si bien, en menor medida, otros muchos algoritmos han sido también utilizados.

- Máxima Entropía (Fleischman *et al.*, 2003a; Busser & Moens, 2003; Pado & Boleda, 2004a; Xue & Palmer, 2004; Toutanova *et al.*, 2005; Ye & Baldwin, 2005; Arcuri, 2006; Erk & Padó, 2006; Tsai *et al.*, 2006).
- Máquinas de Soporte Vectorial (Sgall *et al.*, 2002; Busser & Moens, 2003; Hacioglu & Ward, 2003; Moschitti, 2004; Ping, 2005; Pradhan *et al.*, 2004a; Girju *et al.*, 2004; Sun & Jurafsky, 2004; Giuglea & Moschitti, 2006a; Johansson & Nugues, 2006a; Moschitti *et al.*, 2008).
- Árboles de decisión C4.5 (Chen & Rambow, 2003; Busser & Moens, 2003), y C5 (Sgall *et al.*, 2002; Surdeanu *et al.*, 2003).
- Redes de Bayes (Ohara & Wiebe, 2002; Busser & Moens, 2003).
- Vecino más cercano (Busser & Moens, 2003).
- Combinación de Máxima Entropía con Modelos de Markov (Blunsom, 2004).
- TiMBL (Pado & Boleda, 2004a).
- Experimentos con Ripper (Pado & Boleda, 2004b).
- SNoW (Punyakanok *et al.*, 2005b).
- Clasificador de lista de decisión discriminativa (Liu & Sarkar, 2006).
- Redes neuronales (Musillo & Merlo, 2006).

Universitat d'Alacant
Universidad de Alicante

- Modelos probabilísticos basados en ejemplos (You & Chen, 2004).
- El algoritmo J48 de WEKA (Ye & Baldwin, 2005).
- Perceptron (Che *et al.*, 2006).
- AdaBoost, perceptron y máquinas de soporte vectorial (Surdeanu *et al.*, 2007).
- Varios algoritmos de clasificación, AdaBoost, TiMBL, perceptron y máquinas de soporte vectorial, según la subtarea a realizar (Surdeanu & Turmo, 2008).
- Algoritmo basado en recirculación (Rosa, 2007).

En algunos de estos sistemas, a fin de reducir el efecto de la variabilidad de los datos, se especifica el uso de técnicas de *k-fold cross validation*, como en (Sgall *et al.*, 2002; Busser & Moens, 2003; Ohara & Wiebe, 2002; Pado & Boleda, 2004a; Ye & Baldwin, 2005), todos ellos con $k = 10$, excepto (Surdeanu & Turmo, 2008; Surdeanu & Turmo, 2008) que utilizan $k = 5$.

En otros, se complementa el algoritmo de aprendizaje automático con el uso de algoritmos de re-*ranking* (Fleischman *et al.*, 2003a; Surdeanu & Turmo, 2008), a fin de encontrar la secuencia de etiquetas más probable; o con restricciones lingüísticas y estructurales, codificadas como procedimientos de inferencia (Punyakanok *et al.*, 2005b).

Algunos llevan a cabo procesos de selección de características como en los que a partir de las características de Gildea comentadas anteriormente llevan a cabo extensiones mediante el algoritmo FS (Pradhan *et al.*, 2003; Pradhan *et al.*, 2004c; Pradhan *et al.*, 2005c), o añaden al conjunto final todas las características que de forma individual han aportado alguna mejora (Pradhan *et al.*, 2004a); (Busser & Moens, 2003) que lleva a cabo un proceso de selección aleatorio; (You & Chen, 2004) que evalúa varias combinaciones de características; (Ohara & Wiebe, 2002) que utiliza el algoritmo FS; (Ping, 2005) que lleva a cabo un proceso ascendente, de forma manual; o (Surdeanu & Turmo, 2008) que realiza

Universitat d'Alacant
Universidad de Alicante

pruebas acumulativas con grupos de características. En este sentido destacan los sistemas de (Che *et al.*, 2006; Moschitti *et al.*, 2008) que plantean una alternativa al diseño normal de características que hacen uso de la información facilitada por el árbol de análisis sintáctico. Este nuevo diseño da lugar a las denominadas características estructuradas o *tree kernel*, las cuales definen implícitamente un espacio de características basado en todos los posibles subárboles obtenidos bajo algún criterio, por ejemplo, el conjunto de subárboles formados por el predicado y cada uno de sus argumentos. las llamadas *tree kernel functions* permiten medir la similitud entre árboles sin definir características explícitamente. La idea parte del hecho de que a veces el problema no es tanto la selección de características, si no la generación de las mismas, y se sustenta en que utilizar el árbol de análisis en sí mismo, más que cualquiera de sus representaciones mediante características, hace posible que el clasificador se centre únicamente en las propiedades útiles para decidir. Son importantes los trabajos realizados sobre diferentes características estructuradas realizados por (Moschitti *et al.*, 2005; Moschitti, 2006a; Moschitti, 2006b; Moschitti *et al.*, 2006a; Moschitti *et al.*, 2006b; Pighin & Moschitti, 2007).

Respecto a la `estrategia de anotación` lo más habitual es llevar a cabo dos fases, una para identificar los argumentos de un predicado dado y otra para determinar el rol semántico de cada uno de los argumentos identificados. Sin embargo, hay sistemas que realizan los dos procesos en un único paso. En concreto

- Dos pasos (Chen & Rambow, 2003; Fleischman *et al.*, 2003a; Hacioglu & Ward, 2003; Surdeanu *et al.*, 2003; Busser & Moens, 2003; Xue & Palmer, 2004; Ping, 2005; Pradhan *et al.*, 2004a; Punyakanok *et al.*, 2005b; Toutanova *et al.*, 2005; Ye & Baldwin, 2005; Arcuri, 2006; Erk & Padó, 2006; Johansson & Nugues, 2006a; Liu & Sarkar, 2006; Che *et al.*, 2006; Moschitti *et al.*, 2008).
- Un único paso (Hacioglu & Ward, 2003; Pradhan *et al.*, 2004a; Sun & Jurafsky, 2004).

- Junto con el análisis sintáctico, como en el caso de (Musillo & Merlo, 2006), utilizando un analizador estadístico *Simple Synchrony Network* (Henderson, 2003).

La identificación de los argumentos normalmente se afronta constituyente a constituyente, considerando nodos del árbol sintáctico y determinando si el nodo o constituyente en cuestión es o no argumento. Existen excepciones como en el caso de (Punyakanok *et al.*, 2005b), en el que la tarea se hace determinando el principio y el fin de cada argumento; (Hacioglu & Ward, 2003) que realiza pruebas de identificación palabra a palabra y sintagma a sintagma (Hacioglu, 2004a) en el caso de disponer de información sintáctica parcial, y relación a relación (Hacioglu, 2004b) en el caso de utilizar árboles de dependencias; (Toutanova *et al.*, 2005) que realiza la identificación de todos los nodos del árbol de manera conjunta a fin de evitar el solapamiento de los argumentos; o (Moschitti *et al.*, 2008) que decide si una estructura completa de argumentos es o no correcta, en lugar de mirar argumentos de forma independiente. Para ello primero selecciona el conjunto de argumentos potenciales, en un segundo paso construye todos los posibles subárboles formados por dichos argumentos, y finalmente selecciona uno de los subárboles como correcto.

La clasificación, anota los argumentos detectados en el paso anterior con la etiqueta de rol correspondiente. Para ello, o bien se define un clasificador para cada tipo de rol o bien un único clasificador para todos los roles. Destacan sistemas que asignan una secuencia o patrón de etiquetas en lugar de etiquetas individuales a fin de evitar inconsistencias en la anotación (Fleischman *et al.*, 2003a; Pado & Boleda, 2004a); o sistemas como (Busser & Moens, 2003) que lleva a cabo una clasificación verbo a verbo, (Pradhan *et al.*, 2005b) que desarrolla un clasificador diferente para cada clase de argumento, o como (Pado & Boleda, 2004a) que desarrolla un clasificador para cada *frame*.

En este sentido (Gordon & Swanson, 2007) propone un método para generalizar el proceso de clasificación para roles no presentes en los datos de entrenamiento. Este método se basa en la idea de

que verbos que aparecen en contextos sintácticos similares tienen comportamientos similares respecto a sus argumentos. El método, que no contempla adjuntos, realiza un proceso de alineación de roles semánticos entre verbos con el mismo número de argumentos y caminos en el árbol de análisis similares.

Algunos sistemas llevan a cabo un pre-proceso. En la mayoría de los casos el objetivo es considerar sólo como posibles argumentos los nodos del árbol sintáctico que sean hermanos del predicado en estudio, filtrando así constituyentes que claramente no son argumentos de un verbo (Xue & Palmer, 2004; Punyakanok *et al.*, 2005b; Arcuri, 2006; Erk & Padó, 2006; Liu & Sarkar, 2006; Che *et al.*, 2006). De esta manera se consigue reducir el tiempo de proceso. En otros casos, el objetivo es identificar el *frame* al que pertenece el predicado (Erk & Padó, 2006).

También son necesarios a veces post-procesos. Bien para corregir inconsistencias, como solapamiento, etc. (Punyakanok *et al.*, 2005b; Arcuri, 2006; Moschitti *et al.*, 2008), bien para etiquetar los adjuntos del tipo modal y de negación (Blunsom, 2004; Che *et al.*, 2006), bien para resolución de casos complejos (You & Chen, 2004), bien para reclasificación de las posibles estructuras obtenidas (Moschitti *et al.*, 2008).

Trabajos recientes proponen nuevos tipos de estrategias combinando la información facilitada por otros sistemas de SRL. Este es el caso del trabajo de (Surdeanu *et al.*, 2007; Surdeanu & Turmo, 2008). En el primero se analizan diferentes maneras de combinar las salidas de los sistemas utilizados. Atendiendo a los resultados obtenidos en combinación, en la propuesta de (Surdeanu & Turmo, 2008) las salidas de dos sistemas, una adaptación del inglés al español y catalán de uno de los sistemas utilizados en (Surdeanu *et al.*, 2007), y el sistema (Morante & Busser, 2007), son reclasificadas atendiendo a la redundancia de las estructuras proporcionadas por ambos sistemas, y luego combinadas atendiendo a la información proporcionada por los candidatos propuestos por los sistemas.

Por otro lado, cabe destacar que algunos sistemas consideran los argumentos identificados y sólo se centran en la tarea de clasificación de roles (Sgall *et al.*, 2002; Ohara & Wiebe, 2002; Moschitti, 2004; Pado & Boleda, 2004a; You & Chen, 2004; Giuglea & Moschitti, 2006a; Ye & Baldwin, 2005; Tsai *et al.*, 2006; Rosa, 2007), o bien realizan el proceso de identificación de argumentos de forma manual (Busser & Moens, 2003), o mediante el uso de expresiones regulares (Ye & Baldwin, 2005).

Otros sistemas se centran en la asignación de roles semánticos a sintagmas preposicionales, afrontando la tarea de anotar roles semánticos de preposiciones como una tarea de desambigüación del sentido de las palabras. Para ello, los roles semánticos de los sintagmas preposicionales se tratan como sentidos de palabras para las preposiciones asociadas. Este es el caso de los sistemas (Ohara & Wiebe, 2002; Ye & Baldwin, 2005).

Finalmente destacar sistemas que afrontan la tarea no sólo para predicados verbales si no también para predicados nominales (Pradhan *et al.*, 2004a; Girju *et al.*, 2004).

Evaluación. Realizar una comparación entre los sistemas presentados no siempre es posible, debido principalmente a dos razones. Por un lado, no siempre los autores facilitan medidas comparables, y por otro, no todos los sistemas son equiparables en cuanto a objetivos que se persiguen, información utilizada, corpus, etc.

Por lo general los sistemas son evaluados respecto a precisión, cobertura y medida $F_{\beta=1}$. Precisión (P) es la porción de roles predichos por el sistema que son correctos ($P = correctos/predichos$). Cobertura (C) es la porción de roles correctos que son predichos por el sistema ($R = correctos/(predichos + no_predichos)$). Media $F_{\beta=1}$ calcula la media armónica entre precisión y recall ($F_{\beta=1}$=(2pr)/(p+r)). En otros casos se habla del ratio de error, entendido como el número de roles asignados incorrectamente o no asignados, dividido entre el número total de roles a ser asignados.

El cuadro 4.2 muestra los resultados de los sistemas que llevan a cabo el proceso de anotación en dos fases: identificación de argumentos y asignación de roles. El cuadro 4.3 muestra los

resultados de los sistemas sobre la fase de identificación de argumentos, cuando dicha información es facilitada por los autores. Finalmente, el cuadro 4.4 muestra los resultados de los sistemas que, o bien sólo realizan la asignación de roles, o bien facilitan información sobre la tarea de asignación de manera independiente suponiendo argumentos detectados. La descripción de las siglas utilizadas en la columna de observaciones (OBS) de los cuadros de resultados anteriores se muestra en el cuadro 4.1.

Sigla	Descripción
PB	Corpus Propbank
PT	Penn Treebank
FN	Corpus FrameNet
TI	Corpus SALSA/TIGER
CC	Tratamiento constituyente a constituyente
PP	Tratamiento palabra a palabra
SS	Tratamiento sintagma a sintagma
RR	Tratamiento relación a relación
2P	Proceso en dos pasos
1P	Proceso en un único paso
ST	Análisis sintáctico total
STA	Análisis sintáctico total automático
STM	Análisis sintáctico total manual
STC	Análisis sintáctico total combinado Collins-Charniak
SP	Análisis sintáctico parcial
SPA	Análisis sintáctico parcial automático
SPM	Análisis sintáctico parcial manual
ME	Máxima Entropía
TiMBL5	TiMBL considerando sólo cinco *frames* aleatorios
1C	Un clasificador único
NC	Un clasificador para cada preposición
SR	Combinación por satisfacción de restricciones
CL	Combinación de clasificadores locales
CG	Combinación de clasificadores globales
ca	Catalán
es	Español

Cuadro 4.1. Detalle de las siglas utilizadas en la columna OBS en los cuadros de resultados 4.2, 4.3, 4.4

Universitat d'Alacant
Universidad de Alicante

4.1.2 Aprendizaje automático semi-supervisado

Entre los sistemas que han utilizado estrategias de anotación semi-supervisadas, destaca el sistema de (Thompson *et al.*, 2003), el cual extiende la propuesta de Gildea en dos aspectos. Por un lado, a partir de la palabra que evoca un *frame* y de una lista de constituyentes, determina de forma automática el *frame* y los roles de los constituyentes en ese *frame*. Y por otro, permite identificar roles instanciados al valor nulo.

Este sistema además se caracteriza por:

- *Corpus* FrameNet.
- *Roles*. FrameNet.
- *Información utilizada*. Análisis sintáctico.
- *Algoritmo de aprendizaje*. Modelos ocultos de Markov.
- *Estrategia de anotación*. Sólo realiza la asignación de roles semánticos.
- *Resultados*. 86,10 % de precisión en entrenamiento y 79,3 % en test. El *frame* es elegido correctamente en un 98,10 % en entrenamiento y 97,50 % en test.

Universitat d'Alacant
Universidad de Alicante

4.1.3 Aprendizaje automático no supervisado

Varios sistemas han sido propuestos a fin de llevar a cabo la tarea SRL de forma no supervisada, si bien todos ellos se han cen-

Sistema	OBS	Evaluación		
		P (%)	C (%)	$F_{\beta=1}$ (%)
(Gildea & Jurafsky, 2002)	ST	64,60	61,00	62,75
	SP	27,60	22,00	24,48
(Gildea & Palmer, 2002)	STA	57,70	50,00	53,58
	STM	71,10	64,10	67,42
(Gildea & Hockenmaier, 2003)		71,00	63,10	66,82
(Chen & Rambow, 2003)		60,28	52,13	55,91
(Fleischman *et al.*, 2003b)		66,70	57,40	61,50
(Hacioglu & Ward, 2003)	PB-CC-2P	80,00	74,00	77,00
	PB-CC-1P	77,00	73,00	75,00
(Hacioglu *et al.*, 2003)	PB-PP-ST	79,00	71,00	74,79
	PB-PP-SP	59,00	44,00	50,41
(Hacioglu, 2004a)	PB-SS-SPA	63,00	56,00	59,00
	PB-SS-SPM	66,00	59,00	62,00
(Hacioglu, 2004b)	PB-RR	85,50	83,60	84,40
(Blunsom, 2004)		71,29	50,45	59,09
(Pado & Boleda, 2004a)	ME			47,50
	TiMBL5			53,40
(Xue & Palmer, 2004)	STM			88,51
(Pradhan *et al.*, 2004a)	1P			57,80
(Sun & Jurafsky, 2004)	STM	81,60	82,10	81,10
	STA	86,00	69,20	76,70
(Pradhan *et al.*, 2005c)		84,00	75,00	79,25
(Punyakanok *et al.*, 2005b)	STM	88,81	89,35	89,08
	STA	77,09	75,51	76,29
	SPM	75,34	75,28	75,31
	SPA	75,48	67,13	71,06
	STC	80,53	76,94	78,69
(Ping, 2005)				85,22
(Toutanova *et al.*, 2005)	STM			91,20
(Arcuri, 2006)				69,85
(Erk & Padó, 2006)	FN	78,40		
	TI	67,30		
(Liu & Sarkar, 2006)	STM	83,47	81,82	82,64
(Musillo & Merlo, 2006)				82,80
(Johansson & Nugues, 2006a)		67,00	47,00	55,25
(Che *et al.*, 2006)		82,46	70,65	76,10
(Surdeanu *et al.*, 2007)	SR	84,65	75,51	79,82
	CL	87,47	74,67	80,56
	CG	84,84	76,30	80,34
(Moschitti *et al.*, 2008)		74,36	73,87	74,11
(Surdeanu & Turmo, 2008)	ca	92,16	85,83	88,88
	es	89,75	83,46	86,49
		Error (%)		
(Pado & Boleda, 2004b)	Ripper	22,80		

Cuadro 4.2. Datos sobre la evaluación de sistemas de SRL supervisados

Universitat d'Alacant
Universidad de Alicante

Sistema	OBS	Evaluación		
		P (%)	C (%)	$F_{\beta=1}$ (%)
(Hacioglu & Ward, 2003)	FN	66,10	61,00	63,45
(Surdeanu *et al.*, 2003)		91,62	85,06	88,22
(Xue & Palmer, 2004)	STM			93,80
	STA			81,30
(Pradhan *et al.*, 2004a)	FN			76,30
		Precisión (%)		
(Ping, 2005)		90,50		

Cuadro 4.3. Datos sobre identificación de argumentos de sistemas de SRL supervisados

tra únicamente en la subtarea de asignación de roles semánticos, suponiendo argumentos ya identificados[3].

Para llevar a cabo la tarea han utilizado diferentes *corpus*, como por ejemplo

- 69 artículos tomados de la asociación de tenis profesional[4] en el sistema de (Dennis *et al.*, 2003)
- Propbank en (Dennis *et al.*, 2003; Nielsen & Pradhan, 2004; Padó *et al.*, 2006)
- FrameNet en (Padó *et al.*, 2006)
- BNC en (Swier & Stevenson, 2004)

Diferentes corpus dan lugar a diferentes *conjuntos de roles*:

- El sistema de (Dennis *et al.*, 2003) considera sólo los roles ganador y perdedor, en el caso del dominio del tenis, y los roles protoagente y protopaciente en el caso de Propbank.
- El sistema de (Swier & Stevenson, 2004) reduce a 13 los roles definidos en VerbNet.

[3] El sistema propuesto por (Swier & Stevenson, 2004) en realidad lleva a cabo el paso previo de identificación de los argumentos. Este paso es afrontado mediante un proceso de mapeo con VerbNet y no mediante técnicas de aprendizaje automático no supervisado.

[4] Association of Tennis Professionals (ATP) http://www.atptennis.com/1/en/home/ Consultado en abril 2008.

Universitat d'Alacant
Universidad de Alicante

Sistema	OBS	Evaluación		
		P (%)	C (%)	$F_{\beta=1}$ (%)
(Chen & Rambow, 2003)		82,57	71,41	76,59
(Hacioglu & Ward, 2003)	FN	67,60	55,90	61,20
(Giuglea & Moschitti, 2006a)	FN			84,45
(Tsai *et al.*, 2006)		89,13	86,07	87,57
(Johansson & Nugues, 2006a)		75,00	75,00	75,00
(Rosa, 2007)		94,00	94,00	
		Precisión (%)		
(Gildea & Jurafsky, 2002)		82,00		
(Gildea & Palmer, 2002)	STA	79,20		
	STM	82,80		
(Ohara & Wiebe, 2002)	PT-NC	78,50		
	PT-1C	85,80		
(Ohara & Wiebe, 2003)	FN-NC	70,30		
	FN-1C	49,40		
(Fleischman *et al.*, 2003a)	STM	85,70		
	STA	76,30		
(Surdeanu *et al.*, 2003)		83,05		
(Busser & Moens, 2003)		82,00		
(Moschitti, 2004)	PB	93,20		
	FN	85,20		
(Xue & Palmer, 2004)	STM	92,95		
(Pradhan *et al.*, 2004a)	FN	80,90		
(You & Chen, 2004)		92,71		
(Punyakanok *et al.*, 2005b)	SCA	90,93		
	SCM	91,32		
	SPA	90,62		
	SPM	91,00		
(Ping, 2005)		90,62		
(Toutanova *et al.*, 2005)	STA	94,90		
(Ye & Baldwin, 2005)	PT	99,00		
	PB	58,68		
(Giuglea & Moschitti, 2006a)	PB	81,00		
		Error (%)		
(Sgall *et al.*, 2002)		17,70		

Cuadro 4.4. Datos sobre asignación de roles de sistemas de SRL supervisados

Universitat d'Alacant
Universidad de Alicante

Por otro lado, respecto a la *información utilizada*, todos ellos han hecho uso de análisis sintáctico total, a excepción del sistema de (Swier & Stevenson, 2004) que utiliza análisis sintáctico parcial.

Evaluación. Una vez más los resultados de la evaluación de tales sistemas no es comparable entre unos y otros, por lo que el cuadro 4.5 muestra únicamente la información facilitada por los autores[5].

Sistema	Evaluación
(Dennis *et al.*, 2003)	67,00 % de las ocasiones asigna correctamente el rol ganador y el 74,00 % el de perdedor
(Dennis *et al.*, 2003)	75,00 % de las ocasiones los roles fueron correctamente asignados
(Nielsen & Pradhan, 2004)	88,30 % utilizando análisis sintáctico revisado manualmente
(Swier & Stevenson, 2004)	87,20 % de precisión
(Padó *et al.*, 2006)	80,50 % de medida F entrenando con FrameNet, y 98,60 % entrenando con PropBank, y siempre seleccionando verbos vistos en entrenamiento

Cuadro 4.5. Datos sobre la evaluación de sistemas de SRL no supervisados

4.2 Enfoques basados en conocimiento

4.2.1 Representación basada en reglas

Se han propuesto diferentes sistemas para anotación de roles que hacen uso de un conjunto de reglas para llevar a cabo dicha tarea.

Respecto al `corpus utilizado`, la mayoría de los sistemas propuestos se han desarrollado para inglés utilizando diferentes

[5] No se ha mostrado la información relativa a la identificación de argumentos del sistema de (Swier & Stevenson, 2004) por ser realizada esta fase mediante algoritmos de mapeo. En cualquier caso, el sistema presenta un 90,10 % de precisión para dicha tarea.

corpus, como por ejemplo el corpus Peen Treebank utilizado en (Palmer *et al.*, 2001), los corpus Reuters[6] y *Air Accident Investigation Unit* (AAAI)[7] utilizados en (Hensman & Dunnion, 2004), FrameNet en (Shi & Mihalcea, 2004), EuroWordNet en (Wagner, 2005), o la enciclopedia World Book[8] en (Gomez, 2007).

También ha habido propuestas para otras lenguas como para el checo, con el Czech National Corpus utilizado en (Žabokrtský, 2000), el alemán, como (Fliedner, 2003) con el FrameNet alemán y (Stevens, 2007), el español (Atserias, 2006) con la información facilitada por LEXPIR, o el chino con el Chinese Sinica Treebank utilizado en (Chan, 2006).

Respecto al `conjunto de roles utilizado`, casi todos han hecho uso de los roles propios del corpus, como los roles específicos para cada verbo de PropBank en (Palmer *et al.*, 2001; Stevens, 2007), FrameNet en (Shi & Mihalcea, 2004; Fliedner, 2003), y Lexpir en (Atserias, 2006), los *functors* del Czech National Corpus en (Žabokrtský, 2000), o los 74 roles semánticos del Chinese Sinica Treebank en (Chan, 2006). Otros sistemas han definido conjuntos de roles específicos para la tarea, como en (Gomez, 2007) o en (Wagner, 2005) en el que sólo se han anotado los roles *agent, patient, instrument* y *location*. Y otros han utilizado conjuntos de roles definidos en otros recursos, como el sistema de (Hensman & Dunnion, 2004) que utiliza el conjunto de roles definido en VerbNet.

Respecto a la `información requerida` para llevar a cabo el proceso de anotación, la mayor parte de los sistemas hacen uso de análisis sintáctico completo, como los sistemas de (Palmer *et al.*, 2001; Hensman & Dunnion, 2004; Shi & Mihalcea, 2004; Gomez, 2007; Stevens, 2007). Existen excepciones, como los sistemas que sólo requieren de análisis sintáctico parcial (Chan, 2006; Atserias, 2006), de información sobre el orden de las palabras en función del

[6] http://about.reuters.com/researchandstandards/corpus/ Consultado marzo 2008
[7] http://www.aaiu.ie/ Consultado marzo 2008
[8] http://www.worldbook.com/ Consultado marzo 2008

Universitat d'Alacant
Universidad de Alicante

tipo de oración[9] (Fliedner, 2003), o de la información proporcionada por el nivel analítico del Czech National Corpus (Žabokrtský, 2000).

Algunos de los sistemas han hecho uso de recursos como WordNet, para considerar restricciones sobre los roles en el proceso de mapeo (Hensman & Dunnion, 2004; Gomez, 2007), para utilizar las clases semánticas de la Top Ontology (Atserias, 2006), o para buscar sinónimos de verbos (Shi & Mihalcea, 2004); VerbNet, para disponer de marcos semánticos (Hensman & Dunnion, 2004); o diccionarios para la construcción de reglas (Žabokrtský, 2000).

Finalmente, respecto a la `estrategia de anotación`, muchos de los sistemas sólo afrontan la asignación de roles semánticos suponiendo argumentos detectados. Tal es el caso de los sistemas (Hensman & Dunnion, 2004; Žabokrtský, 2000; Shi & Mihalcea, 2004; Gomez, 2007; Stevens, 2007). El resto de sistemas lleva a cabo el proceso de anotación en dos pasos, de manera que en el primero se determinan los argumentos y en el segundo el rol que juega cada uno de los argumentos identificados. Dentro de estos sistemas, destacan entre otros, el sistema utilizado en el proceso de anotación del corpus Propbank (Palmer *et al.*, 2001), o el sistema desarrollado para el chino (Chan, 2006). La única excepción la constituyen la propuesta de (Fliedner, 2003) en la cual un conjunto de reglas manualmente definidas ayudan a, en un único paso, encontrar el *frame* que evoca una palabra y asignar los elementos de *frame*; y la propuesta de (Atserias, 2006) que afronta la tarea como un problema de optimización, transformando el conocimiento y los procesos en un conjunto de restricciones y tratando de encontrar una solución que las satisfaga para el máximo grado posible.

Para el proceso concreto de la anotación algunos sistemas han desarrollado procesos de mapeo con patrones verbales (Hensman & Dunnion, 2004), con plantillas de árboles (Palmer *et al.*, 2001), con reglas extraídas de FrameNet para cada palabra objetivo (Shi & Mihalcea, 2004), o atendiendo a las restricciones de selección

[9] Basado en la noción de los tipos de oraciones para el alemán (Satzfeld)

asociadas a cada rol (Gomez, 2007); y otros han desarrollado reglas específicas para cada *functor* (Žabokrtský, 2000) o cada rol (Wagner, 2005).

Cabe también destacar que en el caso del sistema de (Žabokrtský, 2000) se lleva a cabo un proceso previo encargado de transformar las estructuras de árbol del nivel analítico en listas de arcos.

Evaluación. La información sobre la evaluación de los sistemas anteriormente comentados se muestra en el cuadro 4.6. Como puede observarse la evaluación facilitada por los autores no siempre permite realizar comparaciones entre los sistemas. La tercera columna indica si la evaluación mostrada se refiere sólo a la anotación de roles semánticos (cl) o también a la identificación de argumentos (id+cl).

Sistema	Evaluación	Fases
(Žabokrtský, 2000)	78,2 % tanto de precisión como de cobertura	cl
(Palmer *et al.*, 2001)	83,7 % de precisión tras la revisión manual	id+cl
(Fliedner, 2003)	Para la mayoría de las oraciones los elementos de *frame* núcleo son asignados correctamente	id+cl
(Hensman & Dunnion, 2004)	60 % de las veces en corpus Reuters, y 70 % en corpus AAIU, identifica correctamente el rol semántico	cl
(Shi & Mihalcea, 2004)	74,5 % de precisión	cl
(Wagner, 2005)	84 % de precisión	cl
(Chan, 2006)	84 % de precisión, 92 % de cobertura y 87,8 de F-medida	id+cl
(Atserias, 2006)	72 % de precisión, 70 % de cobertura	id+cl
(Gomez, 2007)	91 % de precisión para roles y 82 % para adjuntos	cl
(Stevens, 2007)	53,80 % medida F	cl

Cuadro 4.6. Datos sobre la evaluación de sistemas de SRL basados en conocimiento

Universitat d'Alacant
Universidad de Alicante

4.2.2 Representación basada en *frames*

El trabajo de (Mohit & Narayanan, 2003) propone un sistema de anotación de roles semánticos que hace uso del sistema GATE[10] (Cunningham *et al.*, 2002) y de un conjunto de patrones al estilo de los utilizados en tareas de Extracción de Información, construidos a partir de la información contenida en FrameNet. El sistema obtiene una precisión de 68,80 % y una cobertura del 55 %, para un conjunto de oraciones relacionadas con la investigación criminal, extraídas de *Yahoo News Services*[11].

4.3 Campañas internacionales de evaluación de SRL

Ante el auge en el estudio de los roles semánticos y de los sistemas de anotación de los mismos, se han propuesto varias conferencias con el único objetivo de crear un foro específico donde discutir y contrastar resultados y experiencias. Entre estas conferencias destacan *Conference on Computational Natural Language Learning* (CoNLL) y Senseval.

4.3.1 CoNLL shared task

Las *shared tasks*[12] del CoNLL-2004 (Carreras & Màrquez, 2004) y CoNLL-2005 (Carreras & Màrquez, 2005), como en convocatorias previas, tienen el objetivo general de avanzar en estrategias de aprendizaje automático. En concreto, estas ediciones se centraron en el reconocimiento de roles semánticos para inglés, utilizando para ello las estructuras predicado-argumento del corpus PropBank. Dada una oración, la tarea consistía en analizar las proposiciones (un verbo y su conjunto de argumentos) expresadas por los verbos objetivo de la oración. De esta

[10] http://gate.ac.uk/ Consultado en abril 2008
[11] http://news.yahoo.com/ Consultado en abril 2008
[12] http://www.lsi.upc.es/ srlconll/ Consultado en abril 2008

Universitat d'Alacant
Universidad de Alicante

manera, para cada verbo objetivo debían ser reconocidos todos los constituyentes en la oración que rellenaran un rol semántico. En CoNLL-2004 el objetivo fue desarrollar sistemas que utilizaran información sintáctica parcial. Mientras que en CoNLL-2005, el principal foco de interés fue incrementar la cantidad de información sintáctica y semántica utilizada.

La evaluación se lleva a cabo respecto a precisión, cobertura y medida $F_{\beta=1}$. Precisión, porción de argumentos predichos por un sistema que son correctos. Cobertura, porción de argumentos correctos que son predichos por un sistema. Medida $F_{\beta=1}$, es la media armónica entre precisión y cobertura.

Un argumento será considerado como correctamente reconocido, cuando tanto las palabras que forman el argumento como su rol semántico, sean correctas. Para la evaluación total de un sistema, el argumento verbal queda excluido puesto que normalmente coincide con el verbo objetivo de la proposición, el cual es proporcionado en los datos de entrada.

CoNLL *shared task* 2004. Siguiendo ediciones anteriores, en la edición del 2004[13] la información de entrada contenía varios niveles de anotación: palabras, PoS (Giménez & Màrquez, 2003), *chunks* (Carreras & Màrquez, 2003), cláusulas (Carreras & Màrquez, 2003) y entidades con nombre (Chieu & Ng, 2003).

Se proporcionaron datos de entrenamiento (secciones 15-18), para entrenar los sistemas; de desarrollo (sección 20), para ajustar los parámetros de los sistemas; y de test (sección 21), para evaluar los sistemas, todos ellos extraídos del corpus PropBank.

En esta edición participaron 10 equipos caracterizados por:

- *Información utilizada.* Destacan (Hacioglu *et al.*, 2004; van den Bosch *et al.*, 2004) por el uso de las etiquetas ya asignadas; (Hacioglu *et al.*, 2004; van den Bosch *et al.*, 2004; Kouchnir, 2004) patrones de predicados o roles; (Higgins, 2004) por hacer

[13] http://www.lsi.upc.es/ srlconll/st04/st04.html Consultado en abril 200

uso de la información sobre entidades con nombre; estadísticas sobre secuencias de *chunks* (Baldewein *et al.*, 2004a).

- *Algoritmo de aprendizaje.* Dos de los participantes utilizaron máxima entropía (Baldewein *et al.*, 2004a; Lim *et al.*, 2004); otros dos aprendizaje basado en transformación (Higgins, 2004; Williams *et al.*, 2004); dos más aprendizaje basado en memoria (Kouchnir, 2004; van den Bosch *et al.*, 2004); máquinas de soporte vectorial (Hacioglu *et al.*, 2004; Park *et al.*, 2004); perceptrón (Carreras *et al.*, 2004); y SNoW (Punyakanok *et al.*, 2004).

  Solamente tres equipos (van den Bosch *et al.*, 2004; Kouchnir, 2004; Baldewein *et al.*, 2004a) realizan un proceso de selección de características.

- *Estrategia de anotación.* La mayoría de los participantes afrontaron la tarea como identificación y clasificación (Park *et al.*, 2004; Kouchnir, 2004; Baldewein *et al.*, 2004a; Punyakanok *et al.*, 2004), o como una única tarea (Hacioglu *et al.*, 2004; Higgins, 2004; Williams *et al.*, 2004). Como excepción, (Carreras *et al.*, 2004) destaca por realizar la anotación de todas las proposiciones de forma simultánea, y (Lim *et al.*, 2004) por asignar etiquetas de rol a los constituyentes de la cláusula inmediata y en un segundo paso a los constituyentes de cláusulas superiores.

  Algunos de los sistemas realizan pre-procesos para filtrado de argumentos (Punyakanok *et al.*, 2004; Baldewein *et al.*, 2004a) o post-procesos (van den Bosch *et al.*, 2004; Williams *et al.*, 2004; Higgins, 2004; Kouchnir, 2004; Punyakanok *et al.*, 2004), para corrección de errores. El sistema (Park *et al.*, 2004) realiza un post-proceso a la fase de identificación aplicando un conjunto de reglas, tanto automáticas como manuales.

  Casi todos realizan la clasificación sintagma a sintagma, destacando (Baldewein *et al.*, 2004a) que lleva a cabo una clasificación por *chunks*, y (Punyakanok *et al.*, 2004) y (Higgins, 2004) que la realizan palabra a palabra.

Los cuadros 4.7 y 4.8 muestran los resultados obtenidos por los sistemas sobre los conjuntos de desarrollo y test, respectivamente.

Sistema	Precisión (%)	Cobertura (%)	$F_{\beta=1}$ (%)
(Hacioglu *et al.*, 2004)	74,18	69,43	71,72
(Punyakanok *et al.*, 2004)	71,96	64,93	68,26
(Carreras *et al.*, 2004)	73,40	63,70	68,21
(Lim *et al.*, 2004)	69,78	62,57	65,97
(Park *et al.*, 2004)	67,27	64,36	65,78
(Higgins, 2004)	65,59	60,16	62,76
(van den Bosch *et al.*, 2004)	69,06	57,84	62,95
(Kouchnir, 2004)	44,93	63,12	52,50
(Baldewein *et al.*, 2004a)	64,90	41,61	50,71
(Williams *et al.*, 2004)	53,37	32,43	40,35

Cuadro 4.7. Resultados de la *shared task* del CoNLL-2004 sobre el conjunto de desarrollo

Sistema	Precisión (%)	Cobertura (%)	$F_{\beta=1}$ (%)
(Hacioglu *et al.*, 2004)	72,43	66,77	69,49
(Punyakanok *et al.*, 2004)	70,07	63,07	66,39
(Carreras *et al.*, 2004)	71,81	61,11	66,03
(Lim *et al.*, 2004)	68,42	61,47	64,76
(Park *et al.*, 2004)	65,63	62,43	63,99
(Higgins, 2004)	64,17	57,52	60,66
(van den Bosch *et al.*, 2004)	67,12	54,46	60,13
(Kouchnir, 2004)	56,86	49,95	53,18
(Baldewein *et al.*, 2004a)	65,73	42,60	51,70
(Williams *et al.*, 2004)	58,08	34,75	43,48

Cuadro 4.8. Resultados de la *shared task* del CoNLL-2004 sobre el conjunto de test

Universitat d'Alacant
Universidad de Alicante

El cuadro 4.9 muestra los resultados relativos a la subtarea de la asignación de roles, suponiendo identificación de argumentos correctos.

Sistema	Precisión (%)	Cobertura (%)	$F_{\beta=1}$ (%)
(Hacioglu *et al.*, 2004)	78,61	72,47	75,42
(Punyakanok *et al.*, 2004)	77,82	70,04	73,72
(Carreras *et al.*, 2004)	79,22	67,41	72,84
(Lim *et al.*, 2004)	75,43	67,76	71,39
(Park *et al.*, 2004)	73,64	70,05	71,80
(Higgins, 2004)	70,72	63,40	66,86
(van den Bosch *et al.*, 2004)	75,48	61,23	67,61
(Kouchnir, 2004)	66,52	58,43	62,21
(Baldewein *et al.*, 2004a)	75,13	48,70	59,09
(Williams *et al.*, 2004)	70,62	42,25	52,87

Cuadro 4.9. Resultados de la *shared task* del CoNLL-2004 sobre el conjunto de test. Fase de asignación de roles

CoNLL *shared task* 2005. Comparada con la tarea del 2004, la edición del 2005[14] incluía como novedades:

- Corpus de entrenamiento mayores. Secciones 02-21 para entrenamiento, 24 para desarrollo y 23 para test.
- Árboles de análisis sintáctico proporcionados por diferentes analizadores: el analizador de Collins (Collins & Singer, 1999) y el de Charniak (Charniak, 2000).
- Evaluación de los sistemas en corpus diferentes al utilizado en el entrenamiento. Tres secciones del corpus Brown (ck01-03).
- Posibilidad de utilizar recursos externos, aunque ningún sistema participante los utilizó.

La mayor complejidad de la información utilizada, así como el aumento del tamaño del corpus de entrenamiento conlleva serios problemas en tiempo de proceso y recursos. Así lo destacan (Moschitti *et al.*, 2005; Pradhan *et al.*, 2005a; Màrquez *et al.*, 2005; Che *et al.*, 2005; Johansson & Nugues, 2005a; Mitsumori *et al.*, 2005).

En esta edición participaron 19 equipos caracterizados por:

[14] http://www.lsi.upc.es/ srlconll/st05/st05.html Consultado en abril 2008

- *Información utilizada.* Destacan nuevas características como las de sujeto omitido y camino proyectado de (Haghighi *et al.*, 2005).

  Hay sistemas que utilizan la información facilitada por varios analizadores sintácticos, como (Màrquez *et al.*, 2005) que combina un clasificador que hace uso de información sintáctica parcial con otro que utiliza información sintáctica total; o (Pradhan *et al.*, 2005a) que combina dos analizadores sintácticos y un *chunker*. Este tipo de propuestas resulta cara en tiempo de proceso y recursos, por lo que a fin de mantener unos requerimientos de tiempo y memoria algo más admisibles, algunos sistemas reducen los verbos con los que entrenar, atendiendo a su frecuencia de aparición (Màrquez *et al.*, 2005).

  Otras novedades interesantes son la incorporación del conocimiento semántico al análisis sintáctico (Yi & Palmer, 2005; Sutton & McCallum, 2005); el uso de la clase semántica de los núcleos de los sintagmas (Mitsumori *et al.*, 2005); el uso de una base de datos de patrones construida a partir de los datos de entrenamiento (Lin & Smith, 2005).

- *Algoritmo de aprendizaje.* 6 de los equipos utilizaron máxima entropía (Che *et al.*, 2005; Haghighi *et al.*, 2005; Park *et al.*, 2005; Sutton & McCallum, 2005; Yi & Palmer, 2005; Venkatapathy *et al.*, 2005), 4 de ellos máquinas de soporte vectorial (Mitsumori *et al.*, 2005; Pradhan *et al.*, 2005a; Ozgencil & McCracken, 2005; Moschitti *et al.*, 2005), combinación de máxima entropía y máquinas de soporte vectorial (Tsai *et al.*, 2005), o de máxima entropía, vectores de soporte vectorial y aprendizaje basado en memoria (Sang *et al.*, 2005), SNoW (Punyakanok *et al.*, 2005a), árboles de decisión (Ponzetto & Strube, 2005), el algoritmo AdaBoost (Màrquez *et al.*, 2005; Surdeanu & Turmo, 2005), máquinas de vectores relevantes (Johansson & Nugues, 2005a), *tree conditional random fields* (Cohn & Blunsom, 2005), y *consensus in pattern matching* (Lin & Smith, 2005), el cual, como los propios autores indican, constituye una aproximación novedosa y diferente al problema de la anotación de roles semánticos.

(Pradhan *et al.*, 2005a) destaca por utilizar un conjunto de características diferentes para cada tipo de rol, los cuales se han obtenido mediante un proceso de selección de características FS. También destaca por el uso de técnicas de k-*fold cross validation*, así como por los intentos de reducir los problemas de escalado en grandes conjuntos de entrenamiento, utilizando conjuntos de entrenamiento semilla.

Otros sistemas que llevan a cabo un proceso de selección de características son (Sang *et al.*, 2005; Ozgencil & McCracken, 2005; Park *et al.*, 2005) con una proximación de tipo *hill-climbing*, y (Mitsumori *et al.*, 2005) donde se analiza el efecto de eliminar algunas características.

- *Estrategia de anotación.* La mayoría de los sistemas utilizaron un procedimiento en dos pasos, identificación y clasificación (Haghighi *et al.*, 2005; Punyakanok *et al.*, 2005a; Moschitti *et al.*, 2005; Sang *et al.*, 2005; Yi & Palmer, 2005; Ozgencil & McCracken, 2005; Johansson & Nugues, 2005a; Park *et al.*, 2005; Venkatapathy *et al.*, 2005; Lin & Smith, 2005; Sutton & McCallum, 2005). Otros llevan a cabo la tarea en un único paso (Pradhan *et al.*, 2005a; Surdeanu & Turmo, 2005; Tsai *et al.*, 2005; Che *et al.*, 2005; Cohn & Blunsom, 2005; Ponzetto & Strube, 2005; Màrquez *et al.*, 2005; Mitsumori *et al.*, 2005).

  En la etapa de identificación la mayoría de los sistemas realizan una anotación de los nodos del árbol sintáctico, buscando un mapeo uno a uno entre argumentos y constituyentes de análisis. Las únicas excepciones las constituyen los sistemas de (Pradhan *et al.*, 2005a) y (Mitsumori *et al.*, 2005), los cuales realizan una tokenización secuencial basada en *chunks*. Por otra parte, destaca el sistema de (Sang *et al.*, 2005) que a fin de realizar la identificación de argumentos lleva a cabo dos podas, una basada en palabras y otra basada en sintagmas; y (Venkatapathy *et al.*, 2005) que afronta la identificación, como una clasificación de los argumentos en obligatorios, opcionales y nulos.

  En la etapa de clasificación todos los sistemas afrontan la tarea como un problema de n clases, a excepción de (Moschitti *et al.*,

2005) que en un primer paso anota roles núcleo, adjuntos, referencias y discontinuos, y después el subtipo concreto dentro de cada tipo; (Lin & Smith, 2005) que también clasifica de forma independiente roles núcleo y adjuntos; y (Venkatapathy *et al.*, 2005) que asigna la secuencia de roles más probable haciendo uso de la información de *frame*.

La mayoría sistemas llevan a cabo algún tipo de pre-proceso (Punyakanok *et al.*, 2005a; Màrquez *et al.*, 2005; Surdeanu & Turmo, 2005; Tsai *et al.*, 2005; Moschitti *et al.*, 2005; Sang *et al.*, 2005; Yi & Palmer, 2005; Ozgencil & McCracken, 2005; Johansson & Nugues, 2005a; Cohn & Blunsom, 2005; Park *et al.*, 2005; Venkatapathy *et al.*, 2005; Ponzetto & Strube, 2005; Sutton & McCallum, 2005; Lin & Smith, 2005), y algunos de ellos requieren también de un post-proceso (Sang *et al.*, 2005; Che *et al.*, 2005; Surdeanu & Turmo, 2005; Ponzetto & Strube, 2005; Punyakanok *et al.*, 2005a; Tsai *et al.*, 2005; Che *et al.*, 2005; Yi & Palmer, 2005; Ozgencil & McCracken, 2005).

Los cuadros 4.10 y 4.11 muestran los resultados obtenidos por los sistemas sobre los conjuntos de desarrollo y de test, respectivamente. El cuadro 4.12 muestra los resultados sobre el conjunto de test del corpus Brown.

Universitat d'Alacant
Universidad de Alicante

El cuadro 4.13 muestra los resultados en la etapa de clasificación, suponiendo argumentos correctamente detectados, para los 10 sistemas que obtuvieron los mejores resultados.

Sistema	Precisión (%)	Cobertura (%)	$F_{\beta=1}$ (%)
(Punyakanok *et al.*, 2005a)	80,05	74,83	77,35
(Haghighi *et al.*, 2005)	77,66	75,72	76,68
(Màrquez *et al.*, 2005)	78,39	75,53	76,93
(Pradhan *et al.*, 2005a)	80,90	75,38	78,04
(Surdeanu & Turmo, 2005)	79,14	71,57	75,17
(Tsai *et al.*, 2005)	81,13	72,42	76,53
(Che *et al.*, 2005)	79,65	71,34	75,27
(Moschitti *et al.*, 2005)	74,95	73,10	74,01
(Sang *et al.*, 2005)	76,79	70,01	73,24
(Yi & Palmer, 2005)	75,70	69,99	72,73
(Ozgencil & McCracken, 2005)	73,57	71,87	72,71
(Johansson & Nugues, 2005a)	73,40	70,85	72,10
(Cohn & Blunsom, 2005)	73,51	68,98	71,17
(Park *et al.*, 2005)	72,68	69,16	70,87
(Mitsumori *et al.*, 2005)	71,68	64,93	68,14
(Venkatapathy *et al.*, 2005)	71,88	64,76	68,14
(Ponzetto & Strube, 2005)	71,82	61,60	66,32
(Lin & Smith, 2005)	70,11	61,96	65,78
(Sutton & McCallum, 2005)	70,11	61,96	65,78

Cuadro 4.10. Resultados de la *shared task* del CoNLL-2005 sobre el conjunto de desarrollo

4.3.2 Senseval

Las conferencias Senseval[15] son un foro tradicional para evaluación y comparación de sistemas de desambiguación del sentido de las palabras (en inglés, *Word Sense Disambiguation* -WSD-). Desde su primera organización en 1998, estas conferencias han ido ampliando sus objetivos hasta incluir en los años 2006 y 2007 la anotación de roles semánticos entre sus tareas.

Senseval 2006. La conferencia Senseval-3[16], propone, entre otras, una tarea para el desarrollo de sistemas SRL, denominada "`Anotación automática de roles semánticos`" (Litkowski, 2004). La tarea, inspirada en los estudios de (Gildea & Jurafsky, 2002), consiste en identificar los elementos de *frame* dentro de una oración y ano-

[15] http://www.senseval.org/ Consultado en abril 2008
[16] http://www.senseval.org/senseval3 Consultado en abril 2008

Sistema	Precisión (%)	Cobertura (%)	$F_{\beta=1}$ (%)
(Punyakanok *et al.*, 2005a)	82,28	76,78	79,44
(Haghighi *et al.*, 2005)	79,54	77,39	78,45
(Màrquez *et al.*, 2005)	79,55	76,45	77,97
(Pradhan *et al.*, 2005a)	81,97	73,27	77,37
(Surdeanu & Turmo, 2005)	80,32	72,95	76,46
(Tsai *et al.*, 2005)	82,77	70,90	76,38
(Che *et al.*, 2005)	80,48	72,79	76,44
(Moschitti *et al.*, 2005)	76,55	75,24	75,89
(Sang *et al.*, 2005)	79,03	72,03	75,37
(Yi & Palmer, 2005)	77,51	72,97	75,17
(Ozgencil & McCracken, 2005)	74,66	74,21	74,44
(Johansson & Nugues, 2005a)	75,46	73,18	74,30
(Cohn & Blunsom, 2005)	75,81	70,58	73,10
(Park *et al.*, 2005)	74,69	70,78	72,68
(Mitsumori *et al.*, 2005)	74,15	68,25	71,08
(Venkatapathy *et al.*, 2005)	73,76	65,52	69,40
(Ponzetto & Strube, 2005)	75,05	64,81	69,56
(Lin & Smith, 2005)	71,49	64,67	67,91
(Sutton & McCallum, 2005)	68,57	64,99	66,73

Cuadro 4.11. Resultados de la *shared task* del CoNLL-2005 sobre el conjunto de test

tarlos con el nombre del elemento de *frame* apropiado. Para ello se parte de la oración, una palabra objetivo, que puede ser un nombre, un adjetivo o un verbo, y su *frame*.

Para la tarea y como conjunto de test, se utilizaron aproximadamente 8.000 oraciones de FrameNet seleccionadas de forma aleatoria de 40 *frames* seleccionados también aleatoriamente, que tuvieran al menos 370 anotaciones. Esto se traduce en unas 200 oraciones de test para cada *frame*, pudiendo utilizar el resto de las oraciones en el *frame* como entrenamiento. En consecuencia se dispone como mínimo de 170 oraciones de entrenamiento, siendo la media 614 oraciones por *frame*.

En el test, los participantes podían enviar dos ejecuciones. El caso restrictivo, utilizando solamente los datos disponibles en las oraciones de test y en el *frame* correspondiente, relativos al patrón sintáctico pero no a la forma gramatical ni al tipo de sintagma. Y

Sistema	Precisión (%)	Cobertura (%)	$F_{\beta=1}$ (%)
(Punyakanok *et al.*, 2005a)	73,38	62,93	67,75
(Haghighi *et al.*, 2005)	70,24	65,37	67,71
(Màrquez *et al.*, 2005)	70,79	64,35	67,42
(Pradhan *et al.*, 2005a)	73,73	61,51	67,07
(Surdeanu & Turmo, 2005)	72,41	59,67	65,42
(Tsai *et al.*, 2005)	73,21	59,49	65,64
(Che *et al.*, 2005)	71,13	59,99	65,09
(Moschitti *et al.*, 2005)	65,92	61,83	63,81
(Sang *et al.*, 2005)	70,45	60,13	64,88
(Yi & Palmer, 2005)	67,88	59,03	63,14
(Ozgencil & McCracken, 2005)	65,52	62,93	64,20
(Johansson & Nugues, 2005a)	65,17	60,59	62,79
(Cohn & Blunsom, 2005)	67,63	60,08	63,63
(Park *et al.*, 2005)	64,58	60,31	62,38
(Mitsumori *et al.*, 2005)	63,24	54,20	58,37
(Venkatapathy *et al.*, 2005)	65,25	55,72	60,11
(Ponzetto & Strube, 2005)	66,69	52,14	58,52
(Lin & Smith, 2005)	65,75	52,82	58,58
(Sutton & McCallum, 2005)	62,91	54,85	58,60

Cuadro 4.12. Resultados de la *shared task* del CoNLL-2005 sobre el conjunto de test del corpus Brown

Sistema	Precisión (%)	Cobertura (%)	$F_{\beta=1}$ (%)
(Punyakanok *et al.*, 2005a)	86,78	80,98	83,78
(Haghighi *et al.*, 2005)	83,49	81,24	82,35
(Màrquez *et al.*, 2005)	85,01	81,69	83,32
(Pradhan *et al.*, 2005a)	86,86	77,64	81,99
(Surdeanu & Turmo, 2005)	83,81	76,12	79,78
(Tsai *et al.*, 2005)	87,54	74,98	90,77
(Che *et al.*, 2005)	85,57	77,40	81,28
(Moschitti *et al.*, 2005)	82,23	80,83	81,52
(Sang *et al.*, 2005)	83,90	76,47	80,01
(Yi & Palmer, 2005)	82,41	77,58	79,92

Cuadro 4.13. Resultados de la *shared task* del CoNLL-2005 sobre el conjunto de test. Fase de clasificación. 10 mejores sistemas

Universitat d'Alacant
Universidad de Alicante

el caso no restrictivo, utilizando cualquiera de los datos de FrameNet, a excepción de los nombres de los elementos del *frame*. Esto se traduce en dos subtareas, una de identificación y anotación, y una exclusivamente de anotación, suponiendo los límites de los elementos de *frame* obtenidos por un módulo anterior.

Los sistemas son evaluados respecto a, i) precisión[17] y cobertura[18] de elementos de *frame*, ii) la coincidencia de los elementos de *frame* identificados por los sistemas con los identificados en los datos de FrameNet. Además, debido a la dificultad de la tarea, se proporcionaron también medidas adicionales para subconjuntos de respuestas, superconjuntos y solapamientos, y no se penalizó a los sistemas que identificaron más elementos de *frame* de los identificados en FrameNet.

En esta edición participaron 8 equipos, la mayoría de los cuales participaron en las dos tareas, a excepción de 1 que sólo participó en la no restrictiva, la clasificación de argumentos ya detectados; y 3 que lo hicieron sólo en la restrictiva, la identificación de argumentos y su clasificación.

Los sistemas se caracterizaron por:

- *Información utilizada.* Todos los sistemas hicieron uso de la información sintáctica total y de algún tipo de información extraída del *frame*.

  Destacan (Moldovan *et al.*, 2004) por utilizar información extraída de PropBank, y (Ahn *et al.*, 2004) el cual procesa la salida del analizador con el fin de obtener las estructuras de dependencia, etiquetas funcionales y dependencias no locales. Este último además hace uso de WordNet para de poder utilizar la clase semántica para nombres.

- *Algoritmo de aprendizaje.* A excepción de (Thompson *et al.*, 2004) que hace uso de algoritmos de aprendizaje semi-supervisados, el resto de sistemas utilizan algoritmos supervisados. Dos de

[17] Número de respuestas correctas dividido por el número de intentos

[18] Número de respuestas correctas dividido por el número de elementos de *frame* en el conjunto de test

ellos utilizan máxima entropía (Baldewein *et al.*, 2004b; Kwon *et al.*, 2004), otros dos TiMBL (Ahn *et al.*, 2004; Baldewein *et al.*, 2004b), y otros dos más, máquinas de soporte vectorial (Bejan *et al.*, 2004; Moldovan *et al.*, 2004). Destaca el sistema (Ngai *et al.*, 2004) por realizar pruebas con varios algoritmos de clasificación , tanto de forma individual como combinada. En concreto: *boosting*, máquinas de soporte vectorial, máxima entropía, *Snow* y listas de decisión.

Respecto a la selección de características sólo dos sistemas analizan en alguna medida el impacto de las características. (Baldewein *et al.*, 2004b) quitando una característica cada vez, y (Bejan *et al.*, 2004) que lleva a cabo un proceso aleatorio.

- *Estrategia de anotación.* La mayoría de los sistemas llevan a cabo el proceso en dos pasos (Baldewein *et al.*, 2004b; Bejan *et al.*, 2004; Kwon *et al.*, 2004; Moldovan *et al.*, 2004; Thompson *et al.*, 2004), a excepción de (Ahn *et al.*, 2004) que lo realiza en un único paso, y de (Ngai *et al.*, 2004) que sólo afronta la clasificación suponiendo argumentos identificados.

  También es común a muchos de los sistemas afrontar los procesos mediante un clasificador para cada *frame* (Baldewein *et al.*, 2004b; Bejan *et al.*, 2004; Ngai *et al.*, 2004). Por otro lado, (Kwon *et al.*, 2004) identifica secuencias de constituyentes, (Thompson *et al.*, 2004) anota secuencias de roles, (Ngai *et al.*, 2004) hace pruebas con un clasificador para cada par elemento_de_*frame-frame*, y (Moldovan *et al.*, 2004) realiza la identificación considerando las clases: no argumento, mapeo perfecto, argumento potencial, argumento que contiene sub-árbol, solapamiento parcial y sub-árbol que contiene argumento.

  La forma de afrontar la clasificación por *frames*, plantea problemas de baja disponibilidad de datos de entrenamiento. Como solución algunos sistemas proponen procesos de generalización de manera que los ejemplos de varios *frames* sirven como entrenamiento para un *frame* (Bejan *et al.*, 2004; Baldewein *et al.*, 2004b).

Universitat d'Alacant
Universidad de Alicante

Algunos sistemas requieren de un pre-proceso (Baldewein *et al.*, 2004b; Kwon *et al.*, 2004; Moldovan *et al.*, 2004), y (Bejan *et al.*, 2004) de post-proceso.

Los resultados de las mejores ejecuciones de cada equipo para la tarea restrictiva se muestran en el cuadro 4.14.

Sistema	Precisión	Cobertura	Solap.
(Bejan *et al.*, 2004)	0,899	0,772	0,882
(Ahn *et al.*, 2004)	0,869	0,752	0,847
(Kwon *et al.*, 2004)	0,802	0,654	0,784
(Moldovan *et al.*, 2004)	0,807	0,780	0,777
(Baldewein *et al.*, 2004b)	0,736	0,594	0,675
(Ngai *et al.*, 2004)	0,583	0,111	0,480
(Thompson *et al.*, 2004)	0,387	0,335	0,295

Cuadro 4.14. Resultados de la tarea restrictiva en Senseval-3

Los resultados de las mejores ejecuciones de cada equipo para la tarea de clasificación se muestran en el cuadro 4.15.

Sistema	Precisión	Cobertura
(Bejan *et al.*, 2004)	0,946	0,907
(Ngai *et al.*, 2004)	0,926	0,705
(Moldovan *et al.*, 2004)	0,898	0,839
(Kwon *et al.*, 2004)	0,867	0,858
(Thompson *et al.*, 2004)	0,858	0,849

Cuadro 4.15. Resultados de la tarea no restrictiva en Senseval-3

Universitat d'Alacant
Universidad de Alicante

SemEval 2007. En la conferencia Senseval-4, denominada SemEval[19] se plantean algunas novedades respecto a la edición anterior. En primer lugar se organizan dos nuevas tareas: anotación

[19] http://nlp.cs.swarthmore.edu/semeval/index.shtml Consultado en abril 2008

de roles semánticos para el árabe y anotación de roles semánticos para catalán y español. Por otro lado, respecto a la anotación de roles semánticos para el inglés, en esta edición, no se facilita la información sobre el *frame* relacionado con cada oración, siendo, por tanto necesario un paso previo que identifique dicho *frame*.

- Anotación de roles semánticos para catalán y español (Màrquez *et al.*, 2007a). Esta tarea tiene como objetivo evaluar y comparar sistemas automáticos para la anotación de diversos niveles de información semántica para catalán y español. La tarea se divide a su vez en varias subtareas, como desambigüación de nombres y reconocimiento de entidades nombradas, si bien en este trabajo sólo se analizará la correspondiente a SRL, por quedar el resto fuera de nuestro ámbito de estudio.

  El corpus utilizado es un subconjunto del corpus CESS-ECE, en concreto 3.611 oraciones de la parte del español (CESSESP), y 3.202 oraciones de la parte del catalán (CESS-CAT). Este corpus se dividió en 90 % para entrenamiento y 10 % test. Para entrenamiento se proporcionó información sobre lema, etiquetas de PoS, información sintáctica manualmente corregida e incluyendo funciones sintácticas (objeto directo, objeto indirecto, etc.), los roles semánticos y la clase semántica del verbo.

  A su vez el corpus de test se dividía en dos subconjuntos: *in-domain* y *out-of-domain*. El primero homogéneo con respecto al conjunto de entrenamiento, mientras que el segundo correspondía a una parte del corpus CESS-ECE cuya anotación fue posterior al desarrollo de los recursos.

  Solamente participaron dos equipos en la tarea: ILK2 (Morante & Busser, 2007), de la Universidad de Tilburg (información detallada sobre este sistema en (Morante & van den Bosch, 2007)), y el equipo de la UPC (Màrquez *et al.*, 2007b), la Universidad de Cataluña. Estos sistemas se caracterizaron por:

  - Información utilizada. Los dos sistemas utilizaron la información sintáctica proporcionada en el corpus de entrenamiento, incluyendo la función sintáctica. En el caso de ILK2, además

se utilizó información extraída de WordNet; y en el caso de la UPC, información sobre listas de secuencias de argumentos posibles.

- Algoritmo de aprendizaje. ILK2 utilizó aprendizaje supervisado basado en memoria, en concreto el sistema TiMBL; incluyó un proceso de selección de características basado en métodos *hill-cimbling* y el uso de técnicas de validación cruzada.

  El sistema presentado por la UPC implementa una estrategia de reclasificación variante del perceptron de reclasificación de (Collins & Duffy, 2002).

- Estrategia de anotación. La tarea se afronta, para el caso del sistema ILK2, en dos pasos, identificación y clasificación; mientras que el sistema de la UPC lo hace en un único paso. Además ILK2 realiza post-proceso con el fin de corregir algunos errores en la predicción de argumentos del tipo ArgM.

Los resultados de ambos sistemas, tanto para español como para catalán, y tanto para la parte del corpus homogénea al corpus de entrenamiento (in) como la que no (out), se muestran en el cuadro 4.16.

Sistema	Corpus	Precisión(%)	Cobertura(%)	$F_{\beta=1}$(%)
ILK2	catalán	84,72	82,12	83,40
	español	84,30	83,98	84,14
	in	84,71	84,12	84,41
	out	84,26	81,84	83,03
UPC	catalán	84,49	77,97	81,10
	español	83,88	78,49	81,10
	in	84,17	82,90	83,53
	out	84,19	72,77	78,86

Cuadro 4.16. Resultados de SemEval. Tarea: SRL para catalán y español

Universitat d'Alacant
Universidad de Alicante

- Anotación de roles semánticos para el árabe (Diab *et al.*, 2007b). En esta tarea se utilizó el corpus Arabic PropBank, seleccionando los 95 verbos más frecuentes en el corpus. Los datos para desarrollo constaban de 896 oraciones con 1.710 argumentos, para entrenamiento 8.402 oraciones con 21.194 argumentos, y para test 902 oraciones con 1.657 argumentos.

  La evaluación de la tarea se afronta atendiendo a la precisión, cobertura y medida $F_{\beta=1}$ de los sistemas.

  Los resultados obtenidos, significativamente más altos para test que para desarrollo, ponen de manifiesto que el conjunto de test presenta un menor nivel de dificultad para la tarea que el de desarrollo.

  En esta ocasión sólo se presentó un sistema, CUNIT (Diab *et al.*, 2007a), obteniendo los resultados mostrados en el cuadro 4.17. CUNIT se caracteriza por utilizar: i) aprendizaje automático supervisado para la tarea, en concreto máquinas de soporte vectorial, ii) información sobre análisis sintáctico completo; y también por asumir una estrategia de anotación en dos pasos: identificación de argumentos y determinación de sus roles semánticos.

Corpus	Precisión(%)	Cobertura(%)	$F_{\beta=1}$(%)
desarrollo	81,31	74,67	77,84
test	84,71	78,39	81,43

Cuadro 4.17. Resultados de SemEval. Tarea: SRL para árabe

Universitat d'Alacant
Universidad de Alicante

- Extracción de la estructura semántica de *frames* (Baker *et al.*, 2007). Dada una oración, la tarea consistía en, una vez reconocida la palabra que evoca un *frame*: i) asignarle el *frame* correcto de FrameNet, ii) detectar los elementos de *frame* en la oración, iii) anotar los correspondientes elementos de *frame*[20].

[20] La tarea también requiere determinar las realizaciones sintácticas asociadas a los elementos de *frame*, tales como funciones gramaticales o tipos de sintagmas. Sin

Los datos de entrenamiento están formados por las oraciones de FrameNet (versión 1.3), textos del *American National Corpus* (ANC)[21]. Los datos de test se obtuvieron a partir de textos anotados manualmente por el equipo de FrameNet.

La evaluación se lleva a cabo atendiendo a las medidas precisión, cobertura y medida $F_{\beta=1}$.

A esta tarea se presentaron 3 sistemas, UTD-SRL (Bejan & Hathaway, 2007), LTH[22] (Johansson & Nugues, 2007), y CLR (Litkowski, 2007). Éste último sistema sólo participó en la tarea de detección del *frame* utilizando para ello un conjunto de reglas manualmente construidas. Los otros dos sistemas se caracterizan por:

- Información utilizada. UTD-SRL utiliza la información sintáctica proporcionada por el analizar Collins, y en general, reune características de otras muchas propuestas. LTH hace uso de la información facilitada por un analizador de dependencias y por WordNet.
- Algoritmo de aprendizaje. UTD-SRL utiliza una combinación de máquinas de soporte vectorial y máxima entropía, y lleva a cabo un proceso de selección de características. LTH hace uso de máquinas de soporte vectorial.
- Estrategia de anotación. UTD-SRL afronta el problema en tres pasos: i) identificación del *frame* mediante 566 clasificadores multiclase, uno por cada palabra objetivo en FrameNet que evoquen al menos dos *frames* con al menos cinco oraciones anotadas cada uno, ii) identificación de argumentos mediante un clasificador binario, iii) asignación de los roles de dichos argumentos mediante 489 clasificadores multiclase, uno para cada *frame* en FrameNet. LTH lo afronta en dos pasos, identificación argumentos y posterior asignación del rol.

embargo, esta parte de la tarea queda fuera del alcance del trabajo presentado en esta Tesis.

[21] http://www.americannationalcorpus.org/ Consultado en abril 2008

[22] Los datos de test se presentaron 10 días más tarde de lo permitido.

El cuadro 4.18 muestra los resultados obtenidos por los sistemas participantes.

Sistema	Precisión(%)	Cobertura(%)	$F_{\beta=1}$(%)
UTD-SRL	73,65	87,08	79,80
LTH	52,80	68,80	59,70
CLR	55,30	37,20	44,50

Cuadro 4.18. Resultados de SemEval. Tarea: Estructura semántica

- Anotación de roles semánticos para inglés (Pradhan *et al.*, 2007). En esta ocasión se decidió trabajar con dos conjuntos de roles diferentes, PropBank y VerbNet. El corpus proporcionado era el resultado de un proceso de mapeo automático realizado entre ambos corpus (Loper *et al.*, 2007) para los 50 verbos seleccionados.

  Se propusieron dos niveles de participación: i) cerrado, en el que los sistemas sólo podían utilizar los datos proporcionados por la organización; ii) abierto, en el que los sistemas podían utilizar cualquiera de las secciones 02-21 de PropBank, así como cualquier otro recurso.

  En esta tarea participaron dos equipos, uno de los cuales, el sistema UBC-UPC, participó en los dos niveles, haciendo uso de los corpus utilizados en la tarea del CoNLL-2005 para entrenamiento en el caso del nivel abierto; y el otro, el sistema RTV, sólo en el nivel cerrado. Las características que describen a estos sistemas son:

  - Información utilizada. Ambos sistemas utilizan la información proporcionada por el análisis sintáctico completo. En el caso del sistema RTV, esta información se representó en forma de características lineales y también estructuradas.

    El sistema UBC-UPC utilizó además información sobre preferencias de selección basadas en WordNet.

- Algoritmo de aprendizaje. El sistema RTV utilizó máquinas de soporte, y el sistema UBC-UPC, máxima entropía.
- Estrategia de anotación. Los dos sistemas siguen una estrategia de anotación en dos pasos, añadiendo un post-proceso, y además un pre-proceso en el caso del sistema UBC-UPC.

  El sistema RTV destaca por utilizar un clasificador por verbo para la tarea de identificación de argumentos, y un clasificador por verbo y tipo de rol, para la tarea de determinar el rol de cada argumento.

Los resultados para ambos sistemas se muestran en el cuadro 4.19.

Sistema	Tipo	Precisión(%)	Cobertura(%)	$F_{\beta=1}$(%)
UBC-UPC	Abierto	85,31	82,08	83,66
	Cerrado	85,31	82,08	83,66
RTV	Cerrado	81,58	70,16	75,44

Cuadro 4.19. Resultados de SemEval. Tarea: SRL para inglés

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

5. Aportación a la anotación automática de Roles Semánticos

5.1 Introducción

Un rol semántico es la relación entre un constituyente sintáctico (generalmente, aunque no siempre, argumento del verbo) y un predicado (generalmente, aunque no siempre, un verbo). Un rol identifica el papel de un argumento del verbo en el evento que dicho verbo expresa, por ejemplo, un agente, un paciente, un beneficiario, etc., o también adjuntos, como causa, manera o temporal.

Consideremos, por ejemplo, la oración (E40)

(E40) Mary hit John with a baseball yesterday in the park

Las palabras de esta oración se agrupan formando cuatro constituyentes sintácticos, cada uno de ellos con un rol semántico diferente. El constituyente sintáctico "*Mary*" tiene el rol agente, y los constituyentes, "*John*" y "*with a baseball*" tienen los roles paciente e instrumento, respectivamente. Además, los constituyentes "*in the park*" y "*yesterday*" tienen los roles lugar y tiempo, respectivamente.

Los sistemas desarrollados hasta el momento con el objetivo de anotar de forma automática esta clase de información, han sido muy diversos (ver capítulo 4). Estos sistemas se caracterizan por el corpus utilizado en su construcción, y por tanto, la lengua para la que han sido definidos, el conjunto de roles utilizado en la anotación, la información requerida para llevar a cabo el proceso

de anotación, la estrategia seguida en la anotación y, según el caso concreto, el algoritmo de aprendizaje o la representación del conocimiento utilizada, que se haya elegido para la construcción del sistema.

Sin embargo, a pesar de la gran cantidad de propuestas para anotación automática desarrolladas ninguna ha realizado un análisis exhaustivo y profundo de la influencia de la información utilizada en el proceso de desambiguación o anotación de roles semánticos.

Por ello este trabajo aborda el desarrollo de una herramienta para la anotación automática de roles semánticos, denominada SemRol, la cual posee un fuerte componente de análisis y ajuste de la información utilizada. Este componente de análisis da lugar a que el proceso de anotación de roles se realice desde dos perspectivas diferentes:

- Clasificación por sentidos vs única. Frente a la forma clásica de clasificación que utiliza un único clasificador independientemente del sentido del verbo para el cual los argumentos están siendo clasificados, en este trabajo se propone otorgar un papel relevante al sentido del verbo utilizando, para ello, clasificadores diferentes para sentidos de verbos diferentes.
- Clasificación global vs individual. Generalmente el proceso de selección de características se lleva a cabo sobre un conjunto de roles vistos como un todo. Este trabajo propone, además, analizar de forma individual qué características son las más adecuadas para predecir cada uno de los diferentes roles utilizados.

Las características genéricas, así como la arquitectura de SemRol se presentan en el apartado 5.2, y el detalle de sus dos módulos principales, el módulo de procesamiento *off-line* y el módulo de anotación *on-line*, en los apartados 5.3 y 5.4, respectivamente. Para finalizar se presentan los resultados obtenidos por SemRol en el proceso de anotación y su comparación con otros sistemas (apartado 5.5).

5.2 SemRol: Una herramienta de anotación automática de roles semánticos

SemRol es una herramienta de anotación automática de roles semánticos basada en corpus. Al igual que los sistemas analizados en el capítulo 4, la propuesta aquí presentada se caracteriza por el corpus utilizado en su construcción, el conjunto de roles utilizado, el algoritmo de aprendizaje, la información de la que se nutre dicho algoritmo y la estrategia de anotación que se ha seguido. Un resumen de esta información se muestra en el cuadro 5.1.

Corpus	PropBank (secciones 02-21)
Conjunto de roles	PropBank
Algoritmo de aprendizaje	TiMBL
	ME
Estrategia de anotación	sentidos vs única
	global vs individual
Información utilizada	Análisis sintáctico parcial: cláusulas y sintagmas
	PoS
	Sentido de los verbos
	Entidades nombradas

Cuadro 5.1. Características generales de SemRol

Por otra parte, SemRol ha sido diseñado con el fin de poder llevar a cabo un análisis de la influencia de los diferentes tipos de información utilizada por esta clase de sistemas en el proceso de clasificación de argumentos. Esto hace que SemRol se organice en una arquitectura con un importante componente de procesamiento *off-line*.

Los siguientes apartados analizan en detalle cada una de las características comentadas, corpus (apartado 5.2.1), conjunto de roles semánticos utilizado (apartado 5.2.2), estrategia de anotación seguida (apartado 5.2.3), algoritmo de aprendizaje utilizado (apartado 5.2.4), información utilizada (apartado 5.2.5), e introducen las principales aspectos de la arquitectura de SemRol (apartado 5.2.6).

5.2.1 Corpus

De todos los corpus presentados en el capítulo 2, destacan fundamentalmente dos corpus desarrollados para el inglés: PropBank y FrameNet.

PropBank añade información de predicado-argumento, o roles semánticos, a las estructuras sintácticas de la parte del Wall Street Journal (WSJ) utilizada para el proyecto Penn Treebank II (Marcus, 1994). La anotación se realiza mediante un etiquetador automático basado en reglas (Palmer *et al.*, 2001) cuya salida es corregida de forma manual. Para el proceso de anotación, puesto que se parte del corpus WSJ, se tienen los árboles de análisis sintáctico revisados a mano. Dicho proceso de anotación está basado en verbos, de manera que se anotan todas las ocurrencias de un verbo cada vez, comenzando por los verbos más comunes en el corpus. En este proceso se decide qué es argumento y qué rol juega el argumento, estando los argumentos restringidos a nodos en el árbol de análisis. Para cada rol se incluye un campo descriptor, como por ejemplo cosa aceptada, instrumento, etc., el cual es utilizad sólo durante el proceso de anotación. Este corpus fue desarrollado dentro del proyecto PropBank[1] (Palmer *et al.*, 2005) con el objetivo de proporcionar datos de entrenamiento para etiquetadores y analizadores semánticos estadísticos.

La oración (E41)muestra un ejemplo de anotación del corpus PropBank.

Universitat d'Alacant
Universidad de Alicante

(E41) [$_{Arg0}$ He] [$_{ArgM-MOD}$ would][$_{ArgM-NEG}$ n't] accept [$_{Arg1}$ anything of value] [$_{Arg2}$ from those he was writing about]. (wsj-0186)

FrameNet fue desarrollado dentro del proyecto FrameNet (Fillmore, 2002) con el objetivo de crear un recurso léxico para inglés con información detallada de las posibles realizaciones sintácticas de elementos semánticos. Dicha información es organizada en marcos semánticos (en inglés *semantic frames*), los cuales están

[1] http://verbs.colorado.edu/∼mpalmer/projects/ace.html Consultado en marzo 2007

formados entre otros, por las unidades léxicas (pares palabra-significado) más frecuentes que pueden servir como predicados del *frame*, una lista de roles semánticos y un conjunto de ejemplos anotados manualmente. El corpus del que se han extraído las oraciones ejemplo es el British National Corpus (BNC)[2], aunque también se utilizó en menor medida, el corpus American Newswire[3]. Durante el proceso de anotación se buscaban oraciones individuales que contenían a la unidad léxica objetivo, ya fuera verbo, nombre o adjetivo, y que rellenara todos los roles semánticos del *frame* correspondiente. Las oraciones ejemplo eran elegidas de manera que se asegurase cobertura de todas las realizaciones sintácticas de los roles semánticos, prevaleciendo oraciones sencillas frente a otras más complejas (Palmer *et al.*, 2005).

La oración (E42) muestra un ejemplo de anotación del corpus FrameNet.

(E42) [$_{(Speaker,NP,Ext)}$ Others] assert [$_{(Message,Sfin,Somp)}$ that anthropology is the tree and sociology the brach]

En este trabajo se ha elegido el corpus PropBank frente a FrameNet dado que los criterios de selección de ejemplos de FrameNet, basados en simplicidad y cobertura, hacen de FrameNet un recurso limitado y poco realista. Sin embargo, PropBank, creado con el objetivo de proporcionar datos de entrenamiento, añade información de roles semánticos al corpus WSJ. Esto hace de PropBank un recurso que genera menos dependencias y aporta mayor transportabilidad a los sistemas que lo utilicen. Además, FrameNet presenta una cobertura inferior, respecto a PropBank, para el caso de adjuntos.

Por otro lado, los problemas que PropBank pudiera presentar respecto a inferencia y generalización causados por el hecho de que cada etiqueta de rol semántico sea específica para cada verbo, no son lo suficientemente severos como para no justificar su uso en sistemas de aprendizaje automático (Zapirain *et al.*, 2008).

[2] http://www.natcorp.ox.ac.uk/ Consultado en Febrero de 2008
[3] http://americannationalcorpus.org/ Consultado en Febrero de 2008

Universitat d'Alacant
Universidad de Alicante

Concretamente se han utilizado las secciones 02-21 del corpus.

5.2.2 Conjunto de roles semánticos

()

El conjunto de roles utilizado es totalmente dependiente del corpus utilizado. Como se acaba de indicar en el apartado anterior, en el trabajo aquí presentado se hace uso del corpus PropBank.

Recordemos que el conjunto de roles correspondiente a un uso de un verbo es denominado en PropBank, *roleset*, el cual está asociado a un conjunto de *frames* o marcos sintácticos, dando lugar a un denominado *frameset*. El criterio para distinguir *framesets* se basa en semántica, de manera que dos significados de un verbo son distinguidos en *framesets* diferentes si toman diferente número de argumentos. De esta manera, un verbo polisémico puede tener más de un *frameset* cuando las diferencias en significado requieren un conjunto de roles diferentes, uno por cada *frameset*. El procedimiento general es examinar un número de oraciones del corpus y seleccionar los roles que parece que ocurren con más frecuencia y/o son necesarios semánticamente (Kingsbury *et al.*, 2002).

Dada la dificultad de definir un conjunto universal de roles semánticos o temáticos que cubran todos los tipos de predicados, en PropBank, los argumentos semánticos de un verbo son numerados, comenzando por 0 y hasta 5, expresando la proximidad semántica respecto al verbo. Para un verbo en particular, arg0 es generalmente el argumento que muestra las características de un proto-agente de los de Dowty (Dowty, 1991), mientras que arg1 es un proto-paciente o tema. Como muestra el cuadro 5.2, para argumentos de número mayor no se pueden hacer generalizaciones (Baker *et al.*, 2004).

No se ha intentado que las etiquetas de los argumentos tengan el mismo significado, de un sentido de un verbo, a otro. Por ejemplo, el rol jugado por arg2 en un sentido de un predicado dado, puede ser jugado por arg3 en otro sentido.

Rol	Tendencias
Arg0	Agente
Arg1	Objeto directo/tema/paciente
Arg2	Objeto indirecto/beneficiario/instrumento/atributo/estado final/extensión
Arg3	Punto de partida, origen/beneficiario/instrumento/atributo
Arg4	Punto de llegada, destino

Cuadro 5.2. Tendencias de los argumentos numerados de PropBank

Existe un rol específico especial etiquetado como argA. Dicha etiqueta es utilizada para capturar el agente de una acción inducida que ocurre con los verbos volitivos de movimiento. Tal es el caso de la oración (E43).

(E43) Mary volunteered John to clean the garage

Además de los roles numerados específicos de cada verbo, los verbos pueden tomar cualquiera del conjunto de roles generales o adjuntos definidos en PropBank. Se etiquetan como argM, más una etiqueta de función. Una lista detallada de los mismos puede verse en el cuadro 5.3.

5.2.3 Estrategia de anotación

Recordemos que las estrategias de anotación más habituales son, o bien únicamente realizar la anotación de roles semánticos considerando constituyentes ya identificados, o bien identificar constituyentes y anotar sus roles, ya sea en un único paso o en dos. Además, en algunas ocasiones son necesarios pre-procesos y/o post-procesos que ayuden a corregir inconsistencias.

En este trabajo se afronta únicamente la tarea de anotación de roles, suponiendo ya argumentos o constituyentes identificados, y sin necesidad de pre ni post-procesos. A su vez, este proceso de anotación, también denominado clasificación de roles, se ha afrontado desde dos perspectivas diferentes: la perspectiva de los sentidos y la perspectiva individual.

Rol	Descripción
LOC	Lugar
EXT	Extensión (argumento numérico)
DIS	Conectiva del discurso
ADV	Propósito general
NEG	Marca de negación
MOD	Verbo modal
CAU	Causa
TMP	Tiempo
PNC	Propósito
MNR	Modo
DIR	Dirección
PRD	Predicación secundaria (indica que existe relación entre los argumentos, o lo que es lo mismo, que el argumento en cuestión actúa como un predicado para algún otro argumento de la oración. Ej.: *Mary called John an idiot*, relación entre "Jonh" y "an idiot")

Cuadro 5.3. Lista de etiquetas de función de adjuntos en PropBank

Los posibles roles que pueden jugar los argumentos de un verbo dependen en gran medida del significado de dicho verbo, por ello la anotación de roles se afronta por sentidos y de forma única. Frente a la forma clásica de anotación que utiliza un único clasificador para todos los verbos, independientemente del sentido del verbo para el cual los argumentos están siendo clasificados, en este trabajo se propone otorgar un papel relevante al sentido del verbo utilizando, para ello, clasificadores diferentes para sentidos de verbos diferentes. En este caso, las clases consideradas son únicamente los roles de cada sentido de cada verbo. Sin embargo, cuando no se consideran los sentidos, las clases a tener en cuenta la constituyen el conjunto completo de posibles roles.

Consideremos por ejemplo el verbo ***give*** y tres de sus sentidos (#1,#4 y #6), cada uno con su propio conjunto de roles, tal y como se muestra en el cuadro 5.4.

En un proceso de anotación o clasificación por sentidos habría tres clasificadores, uno por sentido. Los clasificadores correspondientes a *give*#1 y *give*#6 utilizarían las clases A0, A1 y A2, mientras que el clasificador give#4 sólo contaría con las clases

Sentido	Rol	Descripción
Give#1	A0	Giver
	A1	Thing given
	A2	Entity given
Give#4	A0	Emitter
	A1	Thing emitted
Give#6	A0	Distributor
	A1	Thing distributed
	A2	Distributed

Cuadro 5.4. Algunos sentidos y sus roles semánticos para el verbo *give* en PropBank

A0 y A1. De esta manera, nunca se asignaría a un argumento de give#4 el rol A2. Además, sería imposible asignar a un argumento de *give*, sea cual sea su sentido, roles diferentes a A0, A1 y A2, como podrían ser por ejemplo, los roles A4 o A5, si consideramos más verbos y más posibles roles. Por contra, ante un verbo-sentido no entrenado, no se realizará ninguna anotación puesto que no existe clasificador para él.

Sin embargo, en una clasificación única, habría un único clasificador que utilizaría tres clases: A0, A1, A2. En este caso, sería posible que a un argumento de *give*#4 el clasificador le asignara el rol A2, o incluso, si se consideraran más verbos con sus roles correspondientes, que se anotara con el rol A4 o cualquier otro rol semántico. Como ventaja, considerar a todos los verbos por igual supone que verbos no entrenados serán igualmente anotados por el clasificador.

Por otro lado, y de manera independiente a considerar o no el sentido de los verbos, la anotación se puede considerar como un proceso único para todos los roles, o proponer clasificadores individuales para cada tipo de rol considerado en el corpus. En el primer caso, las clases consideradas en la tarea serían los posibles roles semánticos, mientras que en el segundo estaríamos ante un conjunto de clasificadores binarios, tantos como roles diferentes haya, que lo único que harían cada uno de ellos es determinar si un argumento juega o no el tipo de rol correspondiente.

Universitat d'Alacant
Universidad de Alicante

5.2.4 Algoritmo de aprendizaje

()

Los sistemas que tienen por objetivo llevar a cabo la anotación automática de roles semánticos han tenido tradicionalmente dos enfoques: i) hacer uso de conocimiento lingüístico previamente adquirido, ii) utilizar corpus anotados previamente construidos. Los primeros, denominados de forma genérica, sistemas basados en conocimiento, son sistemas que resuelven problemas utilizando una representación simbólica del conocimiento humano. La arquitectura de un sistema basado en conocimiento de alguna manera refleja la estructura cognitiva y los procesos humanos. Por ello, entre sus componentes fundamentales se encuentra la base de conocimiento, la cual encapsula en algún formalismo de representación el conocimiento del dominio que debe ser puesto en juego por el sistema para resolver el problema dado.

Los segundos, denominados sistemas basados en corpus o sistemas de aprendizaje automático, tratan de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Esta información no estructurada ha de ser por tanto, traducida o representada en algún formato legible computacionalmente. La representación se hace en forma de atributos o característica, los cuales se definen como la descripción de alguna medida de una muestra o entidad tratada en el problema de aprendizaje automático en estudio. Los atributos tienen un dominio, determinado por los valores que puede tomar el atributo. Y además, cada entidad pertenece a una clase o categoría.

El objetivo del aprendizaje automático es, por tanto, obtener una función que asigne una etiqueta de clase a una nueva muestra no etiquetada, es decir, anotar o clasificar una serie de muestras utilizando una de entre varias categorías. Por esta razón, estos métodos se llaman a veces clasificadores.

Teniendo en cuenta la forma del aprendizaje se puede hablar de aprendizaje supervisado o no supervisado. En el aprendizaje

supervisado el algoritmo determina la clase a la que pertenece una muestra nueva a partir de un conjunto de muestras etiquetadas con la clase a la que cada una de ellas pertenece. A este conjunto de muestras etiquetadas se le denomina conjunto de aprendizaje o entrenamiento. Al conjunto de muestras nuevas que se pretenden clasificar se le denomina conjunto de test. Cuando las clases se desconocen a priori y el algoritmo debe ser capaz de descubrirlas a base de agrupar ejemplos similares en categorías, se dice que el aprendizaje es no supervisado.

Debido a la cantidad de recursos disponibles y a la robustez que tales recursos presentan para ser utilizados en tareas de aprendizaje automático supervisado, en este trabajo se ha optado por seguir el enfoque de los sistemas basados en corpus, y más concretamente, de los enfoques supervisados.

En concreto, de entre los diferentes algoritmos de clasificación supervisados existentes, se han utilizado dos: máxima entropía (ME) y el sistema de aprendizaje basado en ejemplos, denominado TiMBL.

El hecho de haber utilizado dos algoritmos de aprendizaje supervisado diferentes cubre el objetivo de analizar la influencia de dicho algoritmo de aprendizaje en el proceso de selección de características. Es importante determinar si el conjunto de características a utilizar es o no dependiente del uso de uno u otro algoritmo.

ME se ha elegido porque ha demostrado ser uno de los algoritmos de aprendizaje supervisado que se ha comportado de forma correcta en otras tareas de PLN, como el análisis morfológico o la detección de límites de oraciones (Ratnaparkhi, 1998), el análisis sintáctico (Charniak, 2000), o la desambigüación del sentido de las palabras (Suárez, 2004). Las principales ventajas de este algoritmo se encuentran en que con características pobres se puede aplicar con precisión, y que en realidad, ME permite representar sin restricciones el conocimiento del problema específico en forma de características.

Universitat d'Alacant
Universidad de Alicante

Por otro lado, TiMBL se presenta como un sistema sencillo y eficiente a nivel computacional, que también ha demostrado su utilidad en muchas tareas de PLN, como analizadores de dependencias (Canisius & den Bosch, 2007), analizadores morfológicos (van den Bosch *et al.*, 2007) o reconocedores de entidades (Evans, 2003).

Las principales características de los algoritmos utilizados son:

- Un clasificador obtenido por medio de una técnica de ME consta de un conjunto de parámetros o coeficientes los cuales son estimados durante el procedimiento de optimización. Cada coeficiente está asociado con una característica observada en los datos de entrenamiento. El principal propósito es obtener la distribución de probabilidad que maximiza la entropía, esto es, se asume máxima ignorancia y no se considera nada aparte de los datos de entrenamiento de tal manera que no se induce ningún conocimiento que no esté propiamente en los datos (Suárez, 2004).
- El sistema TiMBL utiliza algoritmos basados en ejemplos, los cuales a su vez, se basan en memorizar todos los ejemplos de aprendizaje tal cual están, sin necesidad de intentar generalizar ninguna regla ni representación más concisa. Para clasificar un nuevo ejemplo, el procedimiento a seguir será obtener de la memoria de ejemplos el conjunto de ejemplos más parecidos al que estamos intentando clasificar y asignar la clase que sea la mayoritaria en ese grupo (Màrquez, 2002).

Validación cruzada. Con el fin de evitar que el trabajo realizado sea dependiente de la porción de datos del corpus PropBank utilizada para llevar a cabo el aprendizaje, se plantea el uso de procedimientos de validación cruzada. En concreto, se propone dividir el conjunto de aprendizaje en k particiones mutuamente excluyentes, de manera que el proceso de entrenamiento y test se repite k veces. A partir de las k particiones se obtienen k clasificadores, utilizando como conjunto de aprendizaje para el clasificador i-ésimo todas las particiones menos la partición i-ésima, y esti-

mando el error sobre las muestras de la partición no utilizada en el aprendizaje. Por último, el error se obtiene como la media de los errores de los k clasificadores.

Recordemos que la gran ventaja de este método es que ya no importa tanto cómo estén divididos los datos. Su gran desventaja es el elevado coste computacional que supone la ejecución del proceso k veces.

Para determinar un tamaño de k adecuado se han seguido criterios relativos a tiempos de ejecución y resultados obtenidos de la medida $F_{\beta=1}$, de manera que se ha buscado un punto intermedio entre coste computacional y bondad de los resultados. Para ello se han realizado los siguientes pasos:

1. Evaluar de forma individual cada una de las características sobre el corpus completo.
2. Seleccionar una característica que obtenga buenos resultados para los dos algoritmos de aprendizaje, TiMBL y ME, y tanto para la aproximación por sentidos como única. Este es el caso de la característica F23 para la aproximación única, tanto para TiMBL como para ME. En el caso de la aproximación por sentidos los mejores resultados se obtienen con la característica F1 y los segundos mejores con la característica F23, para los dos algoritmos de aprendizaje. Por esta razón, se elige para el proceso la característica F23.
3. Llevar a cabo el proceso de desambigüación de roles con la característica seleccionada, para diferentes tamaños de corpus, a fin de poder determinar un tamaño de k adecuado.

Como consecuencia de los pasos anteriores se concluye (ver resumen del proceso en cuadro 5.5):

- k=10. Supondría un corpus de entrenamiento de 900.006 palabras y un corpus de prueba de 89.854. Para estos tamaños de entrenamiento y utilizando TiMBL como algoritmo de aprendizaje, los tiempos de respuesta son demasiado altos, a pesar

	TiMBL		ME	
Palabras	VS	U	VS	U
500.039	52,36	55,76	53,35	55.76
700.001	53,73	56,00	54,49	56,10
800.017	54,36	**56,18**	55,18	time out
900.006	54,82	56,16	55,65	time out
989.860	**55,23**	56,15	**55,91**	time out

Cuadro 5.5. Resumen del proceso realizado para determinar un tamaño de k adecuado

de que para el tratamiento por sentidos los resultados son óptimos. Además, a partir de las 800.000 palabras el clasificador único no mejora los resultados. Para el caso de utilizar ME, los tiempos de respuesta para tratamiento único hacen imposible la obtención de resultados; y para el tratamiento por sentidos el resultado obtenido es óptimo. Los problemas con los tiempos de respuesta hacen que esta opción sea `RECHAZADA`.

- k=5. El tamaño del corpus de entrenamiento sería de 800.017 palabras y 189.843 para prueba. Para TiMBL, las diferencias con el mejor resultado son de menos de 1 punto en tratamiento por sentidos, mientras que para el tratamiento único los resultados son óptimos. A pesar de los buenos resultados, los tiempos de ejecución continúan siendo algo elevados. Para ME, los tiempos de respuesta para tratamiento único hacen imposible la obtención de resultados; y para el tratamiento por sentidos la diferencia con el mejor resultado está por debajo de 1 punto. Todo ello hace que esta opción sea `RECHAZADA`.
- k=3. Corpus de entrenamiento de 700.001 palabras y de prueba de 289.859. Los tiempos de respuesta mejoran considerablemente para TiMBL, aunque en el caso de ME continúan siendo aún algo costosos. Las diferencias aumentan a casi 2 puntos para el tratamiento por sentidos y menos de 0,2 para tratamiento único, en el caso de utilizar TiMBL. Con ME, la diferencia en el tratamiento por sentidos está por debajo de 1,5 puntos respecto al mejor resultado, y en el tratamiento único los tiempos de respuesta empiezan a ser aceptables. Resumiendo, con este

Universitat d'Alacant
Universidad de Alicante

tamaño de k, los tiempos de respuesta son aceptables a la vez que los resultados se mantienen muy cerca de los óptimos. Todo ello hace que esta opción sea `ACEPTADA`.

5.2.5 Información utilizada

Independientemente de la estrategia de anotación y del algoritmo de aprendizaje utilizados, los clasificadores necesitan algún tipo de información que les permita realizar una propuesta de anotación. En general, la información utilizada suele ser la proporcionada por el análisis sintáctico, ya sea manualmente anotado u obtenido con analizadores sintácticos estadísticos. Sin embargo, existen algunas excepciones que hacen uso de información sintáctica parcial. En este trabajo se ha optado por analizar el comportamiento de los clasificadores cuando sólo se facilita información sintáctica parcial, dado que supone utilizar información más robusta que si se utilizara análisis sintáctico completo. Trabajos posteriores de otros investigadores han demostrado que la elección ha sido adecuado debido, principalmente, a dos razones: i) la información proporcionada por el análisis sintáctico completo tiene una contribución significativa si el análisis es realizado a mano, sin embargo, si el análisis es automático esta contribución disminuye debido a que el análisis completo automático es menos robusto que el análisis parcial (Yi & Palmer, 2005; Surdeanu *et al.*, 2007; Punyakanok *et al.*, 2008); ii) la información proporcionada por el análisis sintáctico contribuye en mayor medida a la identificación de argumentos, pero no a la determinación del rol jugado por cada uno de ellos (Xue & Palmer, 2004; Punyakanok *et al.*, 2005b), que precisamente es la fase analizada en este trabajo.

Normalmente esta información sintáctica suele complementarse con información a otro niveles, como por ejemplo información sobre entidades nombradas o sobre frecuencias de aparición.

En este trabajo se han definido una serie de características que hacen uso de información a diferentes niveles:

- Nivel léxico-morfológico
  - Etiquetas de categoría gramatical.
  - Sentido de verbos.
- Nivel sintáctico
  - Cláusulas de las oraciones.
  - Sintagmas o *chunks* identificados.
- Nivel semántico
  - Entidades nombradas.
  - Argumentos de los verbos.

Antes de pasar a detallar este conjunto de características utilizadas, es necesario aclarar el uso o interpretación que se ha hecho en este trabajo de los conceptos: raíz, palabra con carga semántica y núcleo de sintagma.

Cuando una característica haga uso de la raíz de una palabra, en realidad, se está refiriendo a la palabra como tal, si la longitud de la misma es menor o igual a cuatro caracteres; y a la mitad de la palabra en caso contrario, es decir, si la longitud de dicha palabra es superior a cuatro caracteres.

Varias características hacen uso de palabras con carga semántica, entendiendo como tal, aquellas palabras que son nombres, verbos, adjetivos o adverbios; o lo que es lo mismo, palabras cuya etiqueta de categoría gramatical pertenece al siguiente conjunto (nombre (NN, NNS, NNP, MPS), adjetivo (JJ, JJR, JJS), verbo (VB, VBD, VBG, VBN, VBP, VBZ, MD), adverbio (RB, RBR, RBS).

Finalmente, aquellas características que hacen referencia a los núcleos de los sintagmas, hacen uso de los nombres de los sintagmas nominales y de los verbos de los sintagmas verbales. Dicho de otra manera, palabras del sintagma nominal con etiqueta de nombre (NN, NNS, NNP, MPS) y palabras del sintagma verbal con etiquetas de verbo (VB, VBD, VBG, VBN, VBP, VBZ, MD).

Una vez realizadas estas aclaraciones, podemos detallar cuál es la información utilizada por el clasificador.

Independientemente de que la información se obtenga a través de proceso léxico-morfológico, sintáctico o semántico, dicha información se organiza en dos grandes grupos, dependiendo de si es referente al argumento o a la oración en general. Con el objetivo de identificar y poder referenciarlas con claridad, cada característica o atributo tiene asignado un número del tipo Fx, siendo x un número secuencial comenzando por 0, por ejemplo F0, F1, etc. Un resumen de estas características se puede ver en los cuadros 5.6 y 5.7.

- Información extraída de los argumentos:
  - Posición del argumento respecto al verbo (F1). Posibles valores (+1, -1). Si el verbo se encuentra después del argumento -1, en caso contrario +1.
  - Distancia en palabras desde el final del argumento hasta el verbo (F3). Posibles valores (0,1,2). Si están juntos el valor será 0, si hay de una a tres palabras el valor será 1, y si hay más de tres palabras el valor será 2.
  - Distancia en sintagmas desde el final del argumento hasta el verbo (F4). Posibles valores (0,1,2). Si están juntos el valor será 0, si hay uno o dos sintagmas el valor será 1, y si hay más de dos sintagmas el valor será 2.
  - Distancia en argumentos desde el final del argumento hasta el verbo (F5). Posibles valores (0,1,2). Si están juntos el valor será 0, si hay un argumento de diferencia el valor será 1, y en cualquier otro caso el valor será 2.
  - Número de palabras entre el final del argumento y el verbo (F6).
  - Número de sintagmas entre el final del argumento y el verbo (F7).
  - Número de argumentos entre el final del argumento y el verbo (F8).

Universitat d'Alacant
Universidad de Alicante

- Tipos de entidades nombradas incluidas en el argumento (F9). Si las hay, se indicarán los tipos diferentes de entidades nombradas que estén incluidas dentro del argumento.
- Cadena de entidades nombradas incluidas en el argumento, con posición (F10). Si las hay, se detallará la cadena de entidades sin eliminar duplicados e indicando la posición dentro de la entidad de cada una de las palabras que la componen.
- Cadena de sintagmas que forman el argumento (F11). Se extraerán los tipos de sintagma que componen un argumento.
- Cadena de sintagmas que forman el argumento indicando posición (F12). Se extraerán las etiquetas de sintagma que componen un argumento indicando la posición de cada palabra dentro del sintagma.
- Preposición inicial (F13). Si el argumento comienza por una preposición, se extrae esa preposición.
- Núcleo de los sintagmas que forman el argumento (F14).
- Categoría gramatical de los núcleos de los sintagmas que forman el argumento (F15).
- Nombres que forman el argumento (F16). Se extraen las palabras del argumento con etiqueta de categoría gramatical correspondiente a un nombre (NN, NNS, NNP, MPS).
- Adjetivos que forman el argumento (F17). Se extraen las palabras del argumento con etiqueta de categoría gramatical correspondiente a un adjetivo (JJ, JJR, JJS).
- Adverbios que forman el argumento (F18). Se extraen las palabras del argumento con etiqueta de categoría gramatical correspondiente a un adverbio (RB, RBR, RBS).
- Palabras con carga semántica que forman el argumento (F19).
- Categoría gramatical de la preposición (F20). Si el argumento comienza por una preposición, la categoría gramatical de esa preposición.

Universitat d'Alacant
Universidad de Alicante

- Raíz de las palabras con carga semántica que forman el argumento (F21).
- Raíz del núcleo de los sintagmas que forman el argumento (F22).
- Categoría gramatical de las palabras que forman el argumento (F23).
- Categoría gramatical de las palabras con carga semántica que forman el argumento (F24).
- Nombres que forman el argumento y su categoría gramatical (F28). Se extraen las palabras del argumento con etiqueta de categoría gramatical correspondiente a un nombre (NN, NNS, NNP, MPS), junto a dicha etiqueta de categoría gramatical.
- Adjetivos que forman el argumento y su categoría gramatical (F29). Se extraen las palabras del argumento con etiqueta de categoría gramatical correspondiente a un adjetivo (JJ, JJR, JJS), junto a dicha etiqueta de categoría gramatical.
- Adverbios que forman el argumento y su categoría gramatical (F30). Se extraen las palabras del argumento con etiqueta de categoría gramatical correspondiente a un adverbio (RB, RBR, RBS), junto a dicha etiqueta de categoría gramatical.
- Palabras con carga semántica que forman el argumento, junto a su categoría gramatical (F31).
- Raíz de las palabras que forman el argumento y su categoría gramatical (F32).
- Raíz del núcleo de los sintagmas que forman el argumento y su categoría gramatical (F33).
- Número de palabras del argumento (F34).
- Primera y última palabra del argumento (F35).
- Primera y última palabra del argumento, junto a su categoría gramatical (F36).
- Raíz de la primera y última palabras del argumento (F37).

Universitat d'Alacant
Universidad de Alicante

- Raíz de la primera y última palabras del argumento, junto a su categoría gramatical (F38).
- Categoría gramatical de la primera y última palabra del argumento (F39).
- Palabras anterior y posterior al argumento (F40).
- Categoría gramatical de las palabras anterior y posterior al argumento (F41).
- Palabras anterior y posterior al argumento, junto con su categoría gramatical (F42).
- Raíz de las palabras anterior y posterior al argumento, junto con su categoría gramatical (F43).
- Raíz de las palabras anterior y posterior al argumento (F44).
- Tipo de sintagma de las palabras anterior y posterior al argumento (F45).
- Etiqueta de sintáctico de las primera y última palabras del argumento (F46).
- Etiqueta de sintáctico de los núcleos de los sintagmas del argumento (F47).
- Etiqueta de sintáctico de las palabras anterior y posterior al argumento (F48).
- Etiqueta de sintáctico de las palabras anterior y posterior al argumento, junto con su categoría gramatical (F49).
- Etiqueta de sintáctico de la primera y última palabra del argumento, junto con su categoría gramatical (F50).
- Etiqueta de sintáctico de los núcleos de los sintagmas del argumento, junto con su categoría gramatical (F51).

▪ Información extraída de la oración:

- Voz (F0). Posibles valores (A)ctiva / (P)asiva. Se considerará que la oración está en pasiva si el verbo tiene etiqueta de categoría gramatical igual a VBN, y en el sintagma verbal

Universitat d'Alacant
Universidad de Alicante

al que pertenezca el verbo aparece el verbo *be*. Será activa en cualquier otro caso.

- Pertenencia del verbo a la cláusula (F2). Posibles valores (0, +1, -1). Si el verbo pertenece a la misma cláusula que el argumento el valor de esta característica será 0, si está una una cláusula superior será +1, y si está en una inferior será -1.
- Infinitivo del verbo (F25).
- Sentido del verbo (F26).
- Infinitivo y sentido del verbo (F27).

Para una mayor claridad sobre las características que se acaban de presentar, los cuadros 5.9 y 5.10 ofrece un detalle de cada una de ellas para la oración (E44). Cada una de las columnas Arg$n$ corresponde a uno de los argumentos de la oración, de manera que n tiene que ver con el orden de aparición del argumento (ver cuadro 5.8).

(E44) The luxury auto marker last year sold 1.214 cars in the U.S.

5.2.6 Arquitectura de SemRol

SemRol es una herramienta de anotación de roles semánticos. Como tal, dada una oración, SemRol será capaz de determinar el papel jugado respecto al verbo, de cada uno de los argumentos o constituyentes de la oración. Pero SemRol también es una herramienta desarrollada para el análisis y estudio da la influencia de los diferentes tipos información utilizada (léxico-morfológica, sintáctica y semántica) en el proceso de desambigüación de roles. Esto hace que SemRol se organice en torno a dos módulos bien diferenciados: i) Módulo de procesamiento *off-line* de aprendizaje, el cual corresponde al proceso de ajuste de características , ii) Módulo de procesamiento *on-line* de anotación de roles semánticos, el cual corresponde al conjunto de clasificadores utilizados en la tarea de anotación.

Núm.	Característica
F0	Voz
F1	Posición del argumento respecto al verbo
F2	Pertenencia del verbo a la cláusula
F3	Distancia en palabras desde el final del argumento hasta el verbo
F4	Distancia en sintagmas desde el final del argumento hasta el verbo
F5	Distancia en argumentos desde el final del argumento hasta el verbo
F6	Número de palabras entre el final del argumento y el verbo
F7	Número de sintagmas entre el final del argumento y el verbo
F8	Número de argumentos entre el final del argumento y el verbo
F9	Tipos de entidades nombradas incluidas en el argumento
F10	Cadena de entidades nombradas en el argumento indicando su posición
F11	Cadena de sintagmas que forman el argumento
F12	Cadena de sintagmas que forman el argumento indicando su posición
F13	Preposición inicial
F14	Núcleo de los sintagmas que forman el argumento
F15	Categoría gramatical o PoS, de los núcleos de los sintagmas del argumento
F16	Nombres que forman el argumento
F17	Adjetivos que forman el argumento
F18	Adverbios que forman el argumento
F19	Palabras con carga semántica que forman el argumento
F20	Categoría gramatical de la preposición inicial
F21	Lema de las palabras con carga semántica que forman el argumento
F22	Lema del núcleo de los sintagmas que forman el argumento
F23	Categoría gramatical de las palabras que forman el argumento
F24	Categoría gramatical de las palabras con carga semántica del argumento
F25	Infinitivo del verbo

Cuadro 5.6. Detalle de las características utilizadas (1/2)

Universitat d'Alacant
Universidad de Alicante

Detalles sobre esta arquitectura se muestran en la figura 5.1. Los siguientes apartados presentan en profundidad ambos módulos.

Número	Característica
F26	Sentido del verbo
F27	Infinitivo y sentido del verbo
F28	Nombres que forman el argumento y su categoría gramatical
F29	Adjetivos que forman el argumento y su categoría gramatical
F30	Adverbios que forman el argumento y su categoría gramatical
F31	Palabras con carga semántica del argumento y su categoría gramatical
F32	Lema de las palabras del argumento con carga semántica y su PoS
F33	Lema del núcleo de los sintagmas del argumento y su categoría gramatical
F34	Número de palabras del argumento
F35	Primera y última palabra del argumento
F36	Primera y última palabra del argumento y su categoría gramatical
F37	Lemas de la primera y última palabras del argumento
F38	Lemas de la primera y última palabras del argumento y su PoS
F39	Categoría gramatical de la primera y última palabra del argumento
F40	Palabras anterior y posterior al argumento
F41	Categoría gramatical de las palabras anterior y posterior al argumento
F42	Palabras anterior y posterior al argumento con su categoría gramatical
F43	Lemas de las palabras anterior y posterior al argumento, con su PoS
F44	Lemas de las palabras anterior y posterior al argumento
F45	Tipo de sintagma anterior y posterior al argumento
F46	Etiqueta sintáctica de las palabras primera y última del argumento
F47	Etiqueta sintáctica de los núcleos de los sintagmas del argumento
F48	Etiqueta sintáctica de las palabras anterior y posterior al argumento
F49	Etiqueta sintáctica de las palabras anterior y posterior al argumento, con su PoS
F50	Etiqueta sintáctica de las palabras del argumento primera y última con su PoS
F51	Etiqueta sintáctica de los núcleos de los sintagmas del argumento y su PoS

Cuadro 5.7. Detalle de las características utilizadas (2/2)

Universitat d'Alacant
Universidad de Alicante

Argn	Argumento
Arg1	The luxury auto maker
Arg2	last year
Arg3	1 cars
Arg4	in the U.S.

Cuadro 5.8. Lista de argumentos de la oración (E44)

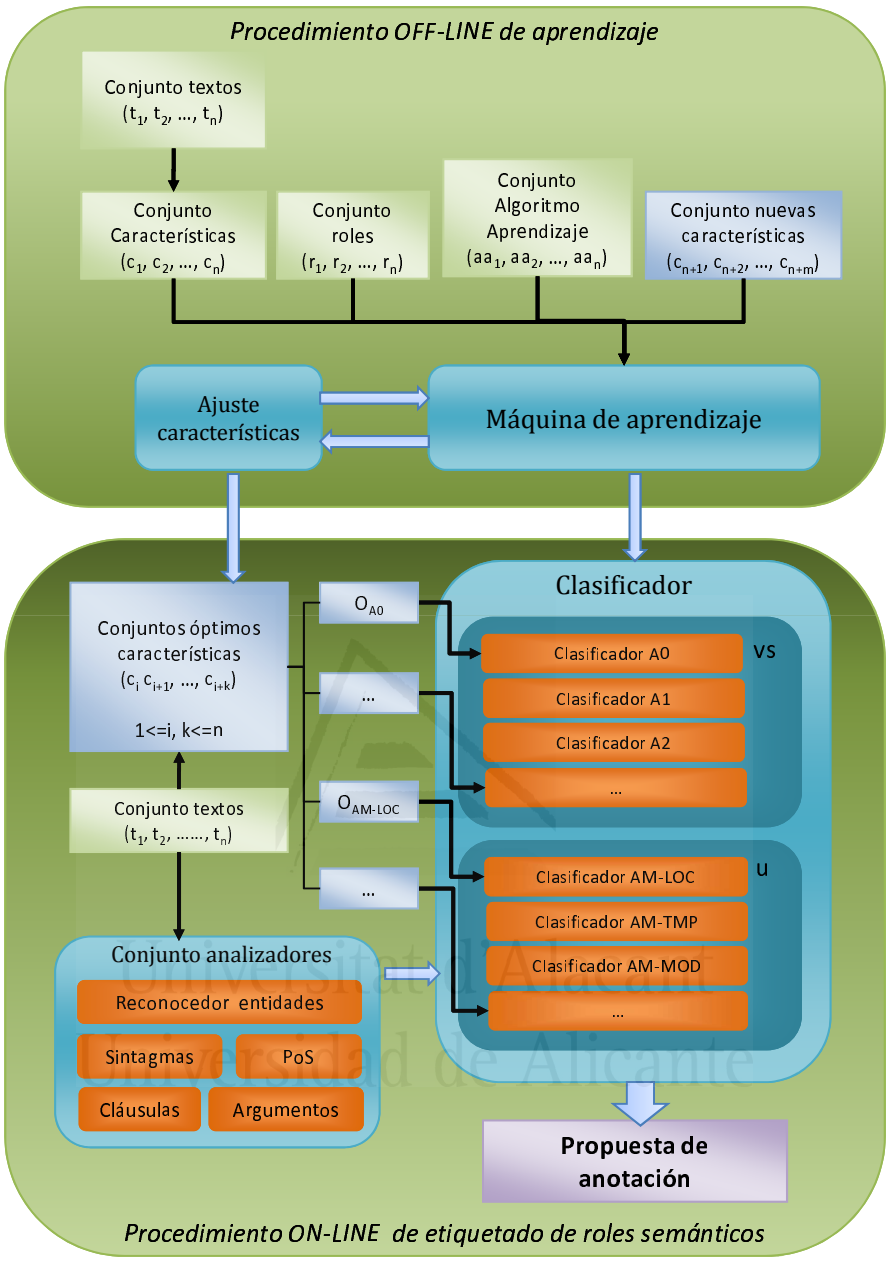


Figura 5.1. Arquitectura del sistema para anotación de roles semánticos: SemRol.

Fn	Arg1	Arg2	Arg3	Arg4
F0	A	A	A	A
F1	-1	-1	+1	+1
F2	0	0	0	0
F3	1	0	0	1
F4	1	0	0	1
F5	1	0	1	0
F6	3	1	1	3
F7	2	1	1	2
F8	1	0	0	1
F9	NULL	NULL	NULL	LOC
F10	NULL	NULL	NULL	B-LOC
F11	NP NP NP NP	NP NP	NP NP	PP NP PP
F12	B-NP I-NP I-NP I-NP	B-NP I-NP	B-NP I-N	B-PP B-NP I-PP
F13	NULL	NULL	NULL	in
F14	luxury auto maker	year	cars	U.S.
F15	NN NN NN	NN	NNS	NNP
F16	luxury auto maker	year	cars	U.S.
F17	NULL	last	NULL	NULL
F18	NULL	NULL	NULL	NULL
F19	luxury auto maker	last year	cars	U.S.
F20	NULL	NULL	NULL	IN
F21	lux auto ma	last year	cars	U.S.
F22	lux auto ma	year	cars	U.S.
F23	DT NN NN NN	JJ NN	CD NNS	IN DT
F24	NN	JJ NN	NNS	NNP
F25	sell	sell	sell	sell

Cuadro 5.9. Ejemplo de valores de las características utilizadas para la oración (E44) (1/2)

Universitat d'Alacant
Universidad de Alicante

5.3 Módulo de procesamiento *off-line* de SemRol

En los sistemas de aprendizaje automático la calidad del conocimiento inducido depende fuertemente de la calidad de las medidas utilizadas. En consecuencia, un problema importante dentro del aprendizaje automático es la selección de los atributos o ca-

Fn	Arg1	Arg2	Arg3	Arg4
F26	01	01	01	01
F27	sell01	sell01	sell01	sell01
F28	luxuryNN autoNN makerNN	yearNN	carsNNS	U.S.NNP
F29	NULL	lastJJ	NULL	NULL
F30	NULL	NULL	NULL	NULL
F31	luxuryNN autoNN makerNN	lastJJ yearNN	carsNNS	U.S.NNP
F32	luxNN autoNN maNN	lastJJ yearNN	carsNNS	U.S.NNP
F33	luxNN autoNN maNN	yearNN	carsNNS	U.S.NNP
F34	4	2	2	3
F35	The maker	last year	1 cars	in U.S.
F36	TheDT makerNN	lastJJ yearNN	1CD carsNNS	inIN U.S.NNP
F37	The ma	last year	1 cars	in U.S.
F38	TheDt maNN	lastJJ yearNN	1CD carsNNS	inIN U.S.NNP
F39	DT NN	JJ NN	CD NNS	IN NNP
F40	last	maker sold	sold in	cars
F41	JJ	NN VBD	VBD IN	NNS
F42	lastJJ	makerNN soldVBD	soldVBD inIN	carsNNS
F43	lastJJ	maNN soldVBD	soldVBD inIN	carsNNS
F44	last	ma sold	sold in	cars
F45	NP	NP VP	VP PP	NP
F46	(S1(S(NP* *)	(NP* *)	(NP* *)	(PP* *))))))
F47	* * *)	*)	*)	*))))))
F48	(NP*	*) (VP*	(VP* (PP*	*)
F49	(NP*JJ	*)NN (VP*VBD	(VP*VBD (PP*IN	*)NNS
F50	(S1(S(NP*DT *)NN	(NP*JJ *)NN	(NP*CD *)NNS	(PP*IN *)))))NNP
F51	*NN *NN *)NN	*)NN	*)NNS	*)))))NNP

Cuadro 5.10. Ejemplo de valores de las características utilizadas para la oración (E45) (1/2)

Universitat d'Alacant
Universidad de Alicante

racterísticas, de manera, que antes de que un algoritmo de aprendizaje se ejecute para hacer predicciones sobre casos nuevos, se debe decidir qué atributos utilizar en esas predicciones y cuáles no. De ahí la importancia de llevar a cabo un estudio y análisis exhaustivos de la influencia de las características utilizadas en el proceso de desambigüación o anotación de roles.

En SemRol este análisis es realizado por el denominado módulo de procesamiento *off-line* de aprendizaje. Los dos componentes fundamentales de dicho módulo son: i) el conjunto de características utilizado, ii) el proceso de selección de las mismas realizado, la denominada máquina de aprendizaje.

5.3.1 Características utilizadas

Como ya se ha comentado en el apartado 5.2.5, SemRol hace uso de información a tres niveles: léxico-morfológico, sintáctico y semántico. Para el caso de las oraciones presentes en el corpus PropBank, esta información es proporcionada por los corpus preparados para las *shared-tasks* de las conferencias CoNLL celebradas en 2004 (Carreras & Màrquez, 2004) y 2005 (Carreras & Màrquez, 2005). Estos corpus son el resultado de analizar las oraciones de PropBank, ya sea mediante el uso de diferentes herramientas, ya sea de forma manual. En concreto:

- Análisis morfológico. Esta información se ha obtenido con la herramienta desarrollada por (Giménez & Màrquez, 2003). Dicha herramienta utiliza técnicas de aprendizaje automático, en concreto máquinas de soporte vectorial, y hace uso de las etiquetas de PoS del Penn Treebank[4]. La herramienta se ha entrenado sobre las secciones 0-18 del corpus Penn Treebank, obteniendo unos resultados del 97 % de precisión. Ver segunda columna de los cuadros 5.11, y 5.12 y 5.13.
- Sentido de los verbos. Los verbos han sido manualmente desambigüados y asignados sentidos de VerbNet, en las secciones 2-21

[4] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html Consultado en julio 2007

y 24. Para cualquier otra palabra que no sea un verbo, esta información no se facilita. Ver columna sexta de los cuadros 5.11, y 5.12 y 5.13.

- Análisis sintáctico parcial. La herramienta desarrollada por (Carreras & Màrquez, 2003), basado en aprendizaje automático, en concreto perceptrón, proporciona información tanto sobre los sintagmas como sobre las cláusulas incluidas en una oración. Los resultados obtenidos por esta herramienta son 93,74 % y 84,36 % de $F_{\beta=1}$ para identificación de sintagmas e identificación de cláusulas, respectivamente. Esta herramienta utiliza las secciones 15-18 del corpus Penn Treebank para entrenamiento y la sección 20 para test.

  La información se presenta en formato *start*end*. En este formato cada etiqueta indica qué sintagmas o cláusulas empiezan y terminan en una determinada palabra. La parte de inicio, *start*, es una concatenación de k paréntesis, cada uno de los cuales representa que una cláusula o un sintagma empieza en esa palabra. La parte de fin, *end*, es una concatenación de paréntesis, cada uno de los cuales representa que una cláusula o un sintagma termina en esa palabra. Ver tercera y cuarta columnas para sintagmas y cláusulas, respectivamente, de los cuadros 5.11, y 5.12 y 5.13.

- Análisis sintáctico total obtenido con el analizador (Charniak, 2000) que utiliza técnicas de aprendizaje automático, en concreto máxima entropía, obteniendo unos resultados de 91,10 % de media entre precisión y recall para oraciones de 40 palabras o menos, y 89,50 % de media entre precisión y recall para oraciones de 100 palabras o menos. El analizador se ha desarrollado utilizando el corpus Penn Treebank, secciones 2-21 para entrenamiento y 23 para test. Por tanto, la información obtenida muestra las etiquetas de sintáctico utilizadas en el Penn Treebank (Marcus *et al.*, 1993). Ver columna séptima de los cuadros 5.11, y 5.12 y 5.13.

- Entidades nombradas. Información facilitada por el reconocedor de entidades desarrollado por (Chieu & Ng, 2003). Esta

herramienta está basada en aprendizaje automático, utilizando máxima entropía y el corpus Reuters, con 88,31 % de $F_{\beta=1}$ sobre los datos de test. Los tipos de entidades reconocidos son localización, organización, persona y *miscellaneus*.

La información relativa a las entidades se muestra en formato IOB2, según el cual palabras fuera de una entidad son etiquetadas con (O)*utside*, palabras comienzo de entidad son etiquetadas con (B)*egin* y palabras incluidas en una entidad son etiquetadas con (I)*nside*. Además, en el caso de B e I, se indica el tipo correspendiente de la forma B-k e I-k, pudiendo tomar k los valores LOC, ORG, PER o MISC. Ver quinta columna de los cuadros 5.11, y 5.12 y 5.13.

- Argumentos de la oración. Esta información es la facilitada por el propio corpus PropBank. Recordar que en este corpus los argumentos se han identificados manualmente para el corpus Penn Treebank. Esta información se presenta en el mismo format *start*end* comentado para sintagmas y cláusulas. La información ha de ser facilitada para cada verbo de la oración, de manera que si hay un único verbo, sólo habrá una información de argumentos; y si hay n verbos habrá n informaciones sobre argumentos. Ver columna novena del cuadro 5.11 y columnas novena y décima de los cuadros 5.12 y 5.13.

Los cuadros 5.11, y 5.12 y 5.13 muestran el resultado de todos estos procesos al ejecutarlos sobre las oraciones (E45) y (E46). La oración (E45) solamente tiene un verbo, por lo que el cuadro 5.11 únicamente tiene una columna de argumentos. Sin embargo, la oración (E46) tiene dos verbos, por lo que los cuadros 5.12 y 5.13 poseen dos columnas de argumentos, una para cada verbo. La primera columna (W) corresponde a las palabras de la oración; la columna (PoS) muestra la información de categoría gramatical de cada una de las palabras; la columna (Ph) la información sobre sintagmas; la columna (C) la información sobre cláusulas; (NE) sobre entidades nombradas; (VS) el sentido de los verbos; (FS) la información facilitada por el analizador sintáctico; (V) el infinitivo de los verbos; y las columnas (Arg$n$) los argumentos identificados.

(E45) The luxury auto marker last year sold 1.214 cars in the U.S.

(E46) The new suitor, Stevric Equity Ventures Inc., of Mineola, N.Y., characterized its proposal as the "first truly independent offer which does not pit one interest group against another within the Arby's franchisee community."

W	PoS	Ph	C	NE	VS	FS	V	Arg
The	DT	B-NP	(S*	O	-	(S1(S(NP*	-	(A*
luxury	NN	I-NP	*	O	-	*	-	*
auto	NN	I-NP	*	O	-	*	-	*
marker	NN	I-NP	*	O	-	*)	-	*)
last	JJ	B-NP	*	O	-	(NP*	-	(A*
year	NN	I-NP	*	O	-	*)	-	*)
sold	VBD	B-VP	*	O	01	(VP*	sell	(V*)
1.214	CD	B-NP	*	O	-	(NP*	-	(A*
cars	NNS	I-NP	*	O	-	*)	-	*)
in	IN	B-PP	*	O	-	(PP*	-	(A*
the	DT	B-NP	*	O	-	(NP*	-	*
U.S.	NNP	I-NP	*)	B-LOC	-	*))))	-	*)

Cuadro 5.11. Detalle de la información proporcionada por el corpus PropBank para la oración (E45). Oración de un sólo verbo. (W)ords-(PoS)-(Ph)rases-(Cl)auses-(N)amed (E)ntities-(V)erb (S)enses-(F)ull (S)yntactic Parser-(V)erb-(Arg)uments

Universitat d'Alacant
Universidad de Alicante

5.3.2 Máquina de aprendizaje

Una de las principales aportaciones en este trabajo es la selección de uno de los mejores conjuntos de características que resuelva la anotación automática de roles semánticos, considerando como mejor, aquel subconjunto de características para el que se obtenga la mayor medida $F_{\beta=1}$[5].

La selección de características se puede considerar como un problema de búsqueda en un cierto espacio de estados, donde ca-

[5] Media armónica entre precisión y cobertura.

W	PoS	Ph	C	NE	VS	FS	V	Arg1	Arg2
The	DT	B-NP	(S*	O	-	(S1(S(NP(NP*	-	(A*	*
new	JJ	I-NP	*	O	-	*	-	*	*
suitor	NN	I-NP	*	O	-	*)	-	*	*
,	,	O	*	O	-	*	-	*	*
Stevric	NNP	B-NP	*	B-ORG	-	(NP(NP*	-	*	*
Equity	NNP	I-NP	*	I-ORG	-	*	-	*	*
Ventures	NNP	I-NP	*	I-ORG	-	*	-	*	*
Inc.	NNP	I-NP	*	I-ORG	-	*)	-	*	*
,	,	O	*	O	-	*	-	*	*
of	IN	B-PP	*	O	-	(PP*	-	*	*
Mineola	NNP	B-NP	*	B-LOC	-	(NP(NP*)	-	*	*
,	,	O	*	O	-	*	-	*	*
N.Y.	NNP	B-NP	*	B-LOC	-	(NP*))))	-	*	*
,	,	O	*	O	-	*)	-	*)	*
characterized	VBD	B-VP	*	O	01	(VP*	characterize	(V*)	*
its	PRP$	B-NP	*	O	-	(NP*	-	(A*	*
proposal	NN	I-NP	*	O	-	*)	-	*)	*
as	IN	B-PP	*	O	-	(PP*	-	(A*	*
the	DT	B-NP	*	O	-	(NP(NP*	-	*	(A*
first	RB	I-NP	*	O	-	*	-	*	*

Cuadro 5.12. Detalle de la información proporcionada por el corpus PropBank para la oración (E46) (2/1). Oración de dos verbos: (W)ords-(PoS)-(Ph)rases-(Cl)auses-(N)amed (E)ntities-(V)erb (S)enses-(F)ull (S)yntactic Parser-(V)erb-(Arg)uments verbo (1)-(Arg)uments verbo (2)

Universitat d'Alacant
Universidad de Alicante

W	PoS	Ph	C	NE	VS	FS	V	Arg1	Arg2
truly	RB	I-NP	*	O	-	(ADJP*	-	*	*
independent	JJ	I-NP	*	O	-	*)	-	*	*
offer	NN	I-NP	*	O	-	*)	-	*	*)
which	WDT	B-NP	(S*	O	-	(SBAR(WHNP*)	-	*	(A*)
does	VBZ	B-VP	(S*	O	-	(S(VP*	-	*	*
not	RB	I-VP	*	O	-	*	-	*	(A*)
pit	VB	I-VP	*	O	01	(VP*	pit	*	(V*)
one	CD	B-NP	*	O	-	(NP*	-	*	(A*
interest	NN	I-NP	*	O	-	*	-	*	*
group	NN	I-NP	*	O	-	*)	-	*	*)
against	IN	B-PP	*	O	-	(PP*	-	*	(A*
another	DT	B-NP	*	O	-	(NP(NP*)	-	*	*
within	IN	B-PP	*	O	-	(PP*	-	*	*
the	DT	B-NP	*	O	-	(NP(NP*	-	*	*
Arby	NNP	I-NP	*	B-ORG	-	*	-	*	*
's	POS	B-NP	*	O	-	*)	-	*	*
franchisee	NN	I-NP	*	O	-	*	-	*	*
community	NN	I-NP	*))	O	-	*))))))))))	-	*)	*)
.	.	O	*)	O	-	*))	-	*	*

Cuadro 5.13. Detalle de la información proporcionada por el corpus PropBank para la oración (E46) (2/2). Oración de dos verbos: (W)ords-(PoS)-(Ph)rases-(Cl)auses-(N)amed (E)ntities-(V)erb (S)enses-(F)ull (S)yntactic Parser-(V)erb-(Arg)uments verbo (1)-(Arg)uments verbo (2)

Universitat d'Alacant
Universidad de Alicante

da estado se corresponde con una cierta característica o subconjunto de características, y el espacio engloba todas los posibles subconjuntos de características que se pueden generar. El proceso de selección de características puede entenderse como el recorrido de dicho espacio hasta encontrar un estado (característica o combinación de características) que optimice alguna función definida sobre un conjunto de atributos. Todo ello con un triple objetivo: simplificar el clasificador, mejorar la precisión del clasificador y reducir la dimensionalidad de los datos para el clasificador.

Una vez definido el espacio de búsqueda, es decir, el conjunto de características a utilizar por el clasificador, será necesario establecer cuatro aspectos:

- Punto de inicio para empezar la búsqueda, aspecto que viene determinado por la estrategia para recorrer el espacio de búsqueda.
- Estrategia para recorrer el espacio de búsqueda.
- Función de evaluación de cada subconjunto de características.
- Criterio de parada del proceso de búsqueda.

En el trabajo que aquí se presenta, se ha decidido, partiendo del conjunto vacío, ir gradualmente añadiendo características siempre y cuando la medida $F_{\beta=1}$ obtenida con la característica o conjunto de características vaya aumentando. Cuando $F_{\beta=1}$ deje de aumentar, habremos encontrado uno de los mejores conjuntos de características. Si el conjunto inicial de características, llegara a ser igual al conjunto completo, sin que $F_{\beta=1}$ disminuya, se encontraría el conjunto óptimo de características.

De manera formal:

Sea F el conjunto de características de cardinalidad m.

Sea f_i una característica perteneciente al conjunto F, con $0 <= i <= m$

Sea $E = ({C_m}^1(f_0, .., f_m), ..., {C_m}^m(f_0, .., f_m))$, el conjunto de estados formado por todas las posibles combinaciones de los f_i pertenecientes a F

Sea e_j un estado perteneciente a E, con $0 <= j <= 2^m$

Sea f_{e_j} el subconjunto de características de F que forman el estado e_j

Sea k la cardinalidad de un estado e_j perteneciente a E

1. Calcular la medida $\mathrm{F}_{\beta=1}$ para cada $e_j \in E$ con $k = 1$
2. Seleccionar el e_j con mejor medida $\mathrm{F}_{\beta=1}$
3. Guardar en P el subconjunto $f_{e_j} \in F$ del e_j seleccionado y $\mathrm{F}_{\beta=1}$(P)
4. Calcular la medida $\mathrm{F}_{\beta=1}$ para cada $e_j \in E$ con cardinalidad $k++$, obtenidos de la combinación de los f_i en P con cada una de las características $f_i \in F - P$
5. Seleccionar el e_j con mejor $\mathrm{F}_{\beta=1}$
6. Guardar en P' el subconjunto $f_{e_j} \in F$ del e_j seleccionado y $\mathrm{F}_{\beta=1}$(P')
7. Si $(F_{\beta=1}(P') > F_{\beta=1}(P))$ y $(P' \subset F)$, $P = P'$ y volver al paso 3
8. Si $(F_{\beta=1}(P') > F_{\beta=1}(P))$, $P = P'$
9. Devolver el subconjunto $f_{e_j} \in F$ en P y $\mathrm{F}_{\beta=1}$(P)

Universitat d'Alacant
Universidad de Alicante

Lo que se traduce en:

- Punto de inicio. Conjunto vacío.
- Estrategia de recorrido. *Forward Selection* (FS). Dicha estrategia consiste en llevar a cabo una búsqueda que añade gradualmente un atributo cada vez, hasta alcanzar el criterio de parada.
- Función de evaluación. Medida $\mathrm{F}_{\beta=1}$.

- Criterio de parada. Cuando los resultados no se mejoren o se hayan probado todas las características.

La estrategia de recorrido del espacio de búsqueda elegida utiliza un algoritmo secuencial puesto que como (Aha & R.L.Bankert, 1994) plantea, entre algoritmos de búsqueda secuenciales, exponenciales y aleatorios, es preferible utilizar secuenciales si atendemos a su coste computacional, el cual resulta prohibitivo en el caso de algoritmos exponenciales; y si atendemos al tamaño de los subconjuntos de características obtenidos, los aleatorios tienden a producir subconjuntos más grandes que los obtenidos con estrategias secuenciales.

Por otro lado, y teniendo en cuenta que entre los algoritmos secuenciales no es posible determinar si uno es mejor que otro (Caruana & Freitag, 1994), en este trabajo se ha elegido el algoritmo *Forward Selection* que, previsiblemente, tendrá un coste computacional menor que *Backward Selection*, puesto que se parte del conjunto vacío en lugar del conjunto completo de características.

La sencillez y bajo coste computacional de la estrategia de recorrido ayuda a compensar el gasto que supone utilizar un método envolvente de evaluación, en lugar de un método filtro. En este caso, el criterio del coste computacional no es suficiente criterio puesto que se ha demostrado que los métodos envolventes ofrecen mejores resultados que los filtro.

Por otra parte, este proceso de selección se aborda desde dos perspectivas diferentes:

- Clasificación por sentidos vs única. Frente a la forma clásica de clasificación que utiliza un único clasificador independientemente del sentido del verbo para el cual los argumentos están siendo clasificados, se propone otorgar un papel relevante al sentido del verbo utilizando, para ello, clasificadores diferentes para sentidos de verbos diferentes.
- Clasificación global vs individual. Generalmente el proceso de selección de características se lleva a cabo sobre un conjunto de roles vistos como un todo. Este trabajo propone, además,

Universitat d'Alacant
Universidad de Alicante

analizar de forma individual qué características serán más adecuados para predecir cada uno de los diferentes roles semánticos considerados.

Añadir características nuevas. Si una vez terminado el proceso *off-line* se quisiera probar la influencia de una nueva característica o un conjunto de ellas, bastaría con repetir el proceso *off-line* para estas nuevas características, con un par de modificaciones. Por un lado, el punto de inicio ya no sería el conjunto vacío, si no que sería el conjunto de características obtenido en el proceso *off-line* inicial. Por otro lado, con el fin de evitar posibles conflictos de las nuevas características con las ya seleccionadas, sería necesario cambiar la estrategia de búsqueda de manera que se permitiera la eliminación secuencial hacia atrás de características. Este procedimiento, denominado *Forward Stepwise Elimination*, consiste en añadir en cada paso la característica que provoque mayor incremento de la función de evaluación, y luego comenzar un proceso de eliminación condicional. Esta eliminación supone ir extrayendo características de forma que la cardinalidad del conjunto de características seleccionado se vaya reduciendo, siempre que la función de evaluación para cada dimensión obtenida por eliminación de una característica sea mayor que la que se obtuvo por adición una característica.

5.3.3 Mejor Conjunto de características

Universitat d'Alacant
Universidad de Alicante

Como resultado del proceso anterior, la máquina de aprendizaje, partiendo del conjunto de características mostrado en los cuadros 5.6 y 5.7, devolverá los conjuntos de características que mejores resultados hayan obtenido. Estos resultados se muestran en el cuadro 5.14. Las diferentes columnas de este cuadro indican, AA: Algoritmo de aprendizaje, EA: Estrategia de anotación, Rol: conjunto de roles para los que se ha hecho el ajuste y por último, Características: el detalle del subconjunto de características obtenido. Así, por ejemplo, La segunda fila muestra los resultados para el conjunto de roles completo, utilizando TiMBL como algoritmo de aprendizaje y siguiendo una estrategia

de anotación basada en sentidos; mientras que en la tercera fila los resultados corresponden a una estrategia de anotación que no tiene en cuenta los sentidos de los verbo; o la fila sexta, que muestra los resultados correspondientes al ajuste para el clasificador del rol de lugar sin considerar los sentidos de los verbos.

Para alcanzar estos resultados se han llevado a cabo una gran cantidad de pruebas con diferentes combinaciones de características. Como ejemplo, los resultados intermedios obtenidos en la ejecución del proceso de ajuste para el caso del clasificador por sentidos para el conjunto completo de roles, utilizando TiMBL como algoritmo de aprendizaje, se muestran en los diferentes cuadros del anexo (apartado 8). Recordemos que el proceso de ajuste se afrontaba haciendo uso de estrategias de validación cruzada con k=3, por lo que los resultados mostrados corresponden a los valores medios obtenidos con dicho proceso de validación cruzada. Cada uno de los cuadros corresponden a un paso en el proceso de ajuste, en concreto, el cuadro 8.1 muestra los resultados obtenidos respecto a precisión, cobertura y medida $F_{\beta=1}$, para cada una de las características consideradas de forma individual. En este caso los mejores resultados de medida $F_{\beta=1}$ (61,33 %), se obtienen para la característica F39. Por ello, el cuadro 8.2 muestra los resultados para las combinaciones de F39 con el resto de características. En esta ocasión, el máximo valor de la medida $F_{\beta=1}$ (69,41 %) se obtiene para la combinación F1,F39. Este par de características se combina con el resto en el cuadro 8.3, obteniendo la mejor medida $F_{\beta=1}$ (71,87 %) para la combinación F1,F13,F39. Y así sucesivamente hasta que ninguna de las combinaciones supera los resultados obtenidos en el paso anterior. Esta condición ocurre con las combinaciones de trece características en adelante, cuadros 8.23 a 8.30, en los que ninguna de las combinaciones de trece y catorce características supera el mejor valor obtenido con combinaciones de doce características (76,34 %) cuadro 8.20.

En estos datos puede observarse que cuando más de una combinación de características obtiene el mismo resultado todas esas combinaciones pasan a la fase siguiente. Este arrastre de combinaciones tiene lugar hasta que o bien se produce un desempate

entre las diferentes combinaciones, o bien la cantidad de combinaciones es tan grande que se elige aquella que suponga menos necesidades de información. Por ejemplo, cuando se realizan las combinaciones de ocho características 8.8, las combinaciones F0,F1,F2,F13,F18,F35,F39,F43 y F0,F1,F2,F13,F30,F35,F39,F43 obtienen la mejor medida $F_{\beta=1}$ (75,91 %), por lo que ambos subconjuntos de características son base de combinación para el resto de características en subconjuntos de cardinalidad nueve. Esta situación se repite en combinaciones de nueve, diez, once, doce y trece características. En este último caso, se obtienen doce combinaciones con la mejor $F_{\beta=1}$. Como arrastrar tantas posibilidades supone un alto coste computacional, recordemos que evitamos estrategias exponenciales, se toma la decisión de seleccionar aquellas características que utilizan menos información ante los mismos resultados, reduciendo las doce combinaciones a cinco.

No hay que olvidar, que el proceso de selección de características propuesto no recorre el espacio E entero, lo que supone obtener uno de los mejores conjuntos de características para la tarea, que no tiene por qué coincidir necesariamente con el mejor de todos.

AA	EA	Rol	Características
TiMBL	vs	Completo	F0,F1,F2,F13,F18,F20,F22,F35,F37,F39,F43,F45
TiMBL	u	Completo	F0,F1,F2,F3,F13,F18,F25,F27,F38,F42
ME	vs	Completo	F12,F37,F39,F45
ME	u	Completo	Time out
TiMBL	u	LOC	F2,F9,F10,F13,F23,F25,F27,F35,F37
TiMBL	u	TMP	F2, F13, F14, F19, F32, F35, F36, F38
TiMBL	u	EXT	F0,F1,F13,F27,F30,F39,F42
TiMBL	u	MOD	F1,F35
TiMBL	u	NEG	F0,F25,F36
TIMBL	u	CAU	F12,F13,F27,F32,F37,F38
TiMBL	u	ADV	F1,F2,F13,F19,F34,F35,F38,F39,F43
TiMBL	vs	A2	F0,F1,F2,F13,F20,F35,F37,F39,F42,F45

Cuadro 5.14. Resultados del proceso de selección de características

5.4 Módulo de procesamiento *on-line* de SemRol

Una vez determinada la información que cada tipo de clasificador necesita, tales clasificadores están disponibles para anotar con roles semánticos cualquier texto nuevo. Para ello, simplemente será preciso determinar el clasificador o clasificadores que se desea utilizar, e incorporar la información que dichos clasificadores necesiten.

Por ejemplo, supongamos que es necesario determinar los argumentos con rol de lugar en un conjunto de oraciones no conocidas por el clasificador. Como se muestra en el cuadro 5.15, será necesario extraer de esas oraciones, información sobre las cláusulas, las entidades nombradas, el sentido del verbo, y la categoría gramatical de las palabras.

C	Descripción	Inf.
F2	Pertenencia a la cláusula del verbo (0, +1, -1)	Cláusulas
F9	Tipos de NE que componen el argumento. Sin posición	NE
F10	Cadena de NE's que componen el argumento. Sin posición	NE
F13	Si el argumento comienza por una preposición, la preposición	PoS
F23	Categoría gramatical de las palabras del argumento	PoS
F25	Infinitivo del verbo objetivo	PoS
F27	Infinitivo y sentido del verbo	WSD
F35	Primera y última palabra del argumento	
F37	Lemas de la primera y última palabra del argumento	

Cuadro 5.15. Detalle de las características del clasificador de lugar

Universitat d'Alacant
Universidad de Alicante

Por tanto, en términos generales, SemRol necesitará de un conjunto de analizadores que proporcionen la información a nivel sintáctico, semántico y léxico-morfológico, utilizada por los clasificadores para llevar a cabo una propuesta de anotación. En concreto las herramientas que proporcionarán tal información son:

- Nivel léxico-morfológico
  - Etiquetas de categoría gramatical. El analizador de *The Cognitive Computation Group*[6]
  - Sentido de verbos. Desambiguador de sentidos del Grupo de Procesamiento del Lenguaje Natural de la Universidad de Alicante (Montoyo *et al.*, 2005).
- Nivel sintáctico
  - Cláusulas de las oraciones. El analizador sintáctico parcial desarrollado por The cognitive Computation Group[7].
  - Sintagmas o *chunks* identificados. El analizador sintáctico parcial desarrollado por *The Cognitive Computation Group*[8].
- Nivel semántico
  - Entidades nombradas. LingPipe[9], uno de los reconocedores de entidades que mejores resultados ofrece para inglés.
  - Argumentos de los verbos. El identificador de argumentos desarrollado por *The Cognitive Computation Group*[10]

De esta manera, cuando se realiza una petición de anotación de roles a SemRol, en primer lugar se utilizará el conjunto de analizadores para extraer toda la información necesitada, y luego, utilizando los conjuntos de características devueltos por la máquina de aprendizaje del proceso *off-line*, se ejecutarán los clasificadores deseados, obteniendo la propuesta de anotación.

[6] http://l2r.cs.uiuc.edu/ cogcomp/demo.php?dkey=POS Consultado en marzo 2008
[7] http://l2r.cs.uiuc.edu/ cogcomp/demo.php?dkey=SP Consultado en marzo 2008
[8] http://l2r.cs.uiuc.edu/ cogcomp/demo.php?dkey=SP Consultado en marzo 2008
[9] http://www.roseindia.net/software-tutorials/detail/5325 Consultado en marzo 2008
[10] http://l2r.cs.uiuc.edu/ cogcomp/demo.php?dkey=SRL Consultado en marzo 2008

Universitat d'Alacant
Universidad de Alicante

5.5 Evaluación de SemRol

La evaluación de SemRol se ha hecho desde varias perspectivas. En primer lugar se presentarán los resultados del proceso de ajuste de características, mostrando cómo otros conjuntos de características, tanto mayores como menores al elegido, obtienen peores resultados, y cómo este comportamiento se mantiene aún cambiando el algoritmo de aprendizaje utilizado (apartado 5.5.1).

Por otro lado, se mostrarán los resultados de las pruebas realizadas con clasificadores únicos y clasificadores por sentido, y cómo el algoritmo de aprendizaje automático supervisado que se utilice, así como la cantidad de información sintáctica, influyen en tales resultados (apartado 5.5.2). Los resultados mostrados determinarán bajo qué circunstancias será más adecuada una clasificación por sentidos o una clasificación única para todos los predicados.

También se verá cómo la definición de clasificadores específicos para cada tipo de rol contribuye a la mejora de los resultados totales del anotador (apartado 5.5.3).

Finalmente, se compararán los resultados obtenidos en el proceso de anotación utilizando SemRol, con los obtenidos por otros sistemas de características similares (apartado 5.5.4).

La evaluación se ha realizado respecto a precisión, cobertura y medida $F_{\beta=1}$. Precisión (P) es la porción de roles predichos por el sistema que son correctos ($P = correctos/predichos$). Cobertura (C) es la porción de roles correctos que son predichos por el sistema ($R = correctos/(predichos + no_predichos)$). La medida $F_{\beta=1}$ calcula la media armónica entre precisión y *recall* ($F_{\beta=1}$=(2pc)/(p+c)).

Los datos que se mostrarán a continuación se han obtenido utilizando el programa de evaluación facilitado por la organización de las *shared task* de las conferencias CoNLL 2004 y 2005.

5.5.1 Proceso de ajuste

Los cuadros 5.16 y 5.17 muestran los resultados obtenidos utilizando el algoritmo de aprendizaje TiMBL, tanto para clasificadores individuales para cada sentido, como para un clasificador único. Las filas dos a trece del cuadro 5.16 muestran como los resultados van mejorando conforme se van añadiendo características durante el proceso de ajuste, hasta llegar a obtener los mejores resultados con un conjunto de doce características ($F_{\beta=1}$ 76,34 %). Las filas catorce y quince muestran que conjuntos de características mayores, con cardinalidades de trece y catorce características, obtenidos en el proceso de ajuste no mejoran los resultados ($F_{\beta=1}$ 76,34 %). La fila dieciséis, corresponde a un conjunto de veinticinco características, cuyo resultado es peor que el mejor conjunto de doce características en más de 13 puntos ($F_{\beta=1}$ 63,05 %). La última fila muestra los resultados obtenidos para el conjunto de las diez características con mejores resultados individuales y cómo estos resultados están algo más de 4 puntos por debajo de los mejores resultados ($F_{\beta=1}$ 71,96 %).

Además, el cuadro 5.17 ofrece unos resultados similares, pero para el caso de un clasificador único para todos los predicados. En este caso, el proceso de ajuste ha obtenido los mejores resultados con un conjunto de diez características ($F_{\beta=1}$ 81,09 %), tal y como muestran las filas dos a once del cuadro. Combinaciones de características de cardinalidades superiores, once y doce, ofrecen resultados algo menores, como muestran las filas doce y trece del cuadro ($F_{\beta=1}$ 81,01 % y 80,89 %, respectivamente).

Por otro lado, el cuadro 5.18 muestra la información para el proceso de ajuste de características utilizando clasificadores individuales para cada sentido de cada verbo, pero en este caso con un algoritmo de aprendizaje diferente, ME. En este caso, los mejores resultados se obtienen para un conjunto de cuatro características ($F_{\beta=1}$ 71,62 %) tal y como muestran las filas dos a cuatro del cuadro. Conjuntos de características mayores como el mostrado en la fila cinco supone un ligero empeoramiento de los resultados ($F_{\beta=1}$ 71,60 %).

Estos resultados ponen de manifiesto la obligatoriedad de un proceso de ajuste en procesos que hacen uso de aprendizaje automático, independientemente de la estrategia de anotación utilizada o del algoritmo de aprendizaje elegido, con el fin de eliminar aquellas características que interfieran con otras más útiles.

Características	P(%)	C(%)	$F_{\beta=1}$(%)
F39	61,69	60,98	61,33
F1,F39	69,83	68,99	69,41
F1,F13,F39	72,31	71,43	71,87
F1,F13,F39,F43	74,19	73,15	73.67
F1,F13,F35,F39,F43	75,99	74,52	75,05
F1,F2,F13,F35,F39,F43	76,15	75,05	75,60
F0,F1,F2,F13,F35,F39,F43	76,33	75,22	75,77
F0,F1,F2,F13,F18,F35,F39,F43	76,47	75,36	75,91
F0,F1,F2,F13,F18,F22,F35,F39,F43	76,50	75,47	76,02
F0,F1,F2,F13,F18,F22,F35,F39,F43,F45	76,71	75,58	76,14
F0,F1,F2,F13,F18,F20,F22,F35,F39,F43,F45	76,82	75,78	76,24
F0,F1,F2,F13,F18,F20,F22,F35,F37,F39,F43,F45	**76,91**	**75,78**	**76,34**
F0,F1,F2,F13,F16,F18,F20,F22,F35,F37,F39,F43,F45	76,91	75,78	76,34
F0,F1,F2,F13,F16,F18,F20,F22,F27,F35,F37,F39,F43,F45	76,91	75,78	76,34
Set of twenty five	64,90	61,30	63,05
F1,F11,F12,F23,F24,F39,F41,F43,F44,F45	72,48	71,45	71,96

Cuadro 5.16. Comportamiento de las características en la clasificación por sentidos. Algoritmo TiMBL.

Características	P(%)	C(%)	$F_{\beta=1}$(%)
F38	63,22	63,92	63,57
F38,F42	70,25	70,95	70,59
F27,F38,F42	73,66	74,15	73,91
F13,F27,F38,F42	76,55	77,06	76,81
F1,F13,F27,F38,F42	78,87	79,38	79,12
F1,F2,F13,F27,F38,F42	80,00	80,52	80,26
F0,F1,F2,F13,F27,F38,F42	80,38	80,89	80,63
F0,F1,F2,F13,F18,F27,F38,F42	80,48	81,03	80,76
F0,F1,F2,F13,F18,F25,F27,F38,F42	80,56	81,09	80,82
F0,F1,F2,F3,F13,F18,F25,F27,F38,F42	**80.84**	**81.34**	**81.09**
F0,F1,F2,F3,F13,F18,F25,F27,F38,F39,F42	80,75	81,26	81,01
F0,F1,F2,F3,F13,F18,F25,F27,F38,F39,F42,F44	80,63	81,14	80,89

Cuadro 5.17. Comportamiento de las características en la clasificación única. Algoritmo TiMBL.

Universitat d'Alacant
Universidad de Alicante

Características	P(%)	C(%)	$F_{\beta=1}$(%)
F39	61,91	62,44	62,17
F39,F45	68,38	38,85	68,61
F37,F39,F45	71,24	71,81	71,53
F12,F37,F39,F45	**71,33**	**71,92**	**71,62**
F12,F37,F39,F43,F45	72,06	71,15	71,60

Cuadro 5.18. Comportamiento de las características en la clasificación por sentidos. Algoritmo ME.

5.5.2 Clasificador por sentidos frente Clasificador único

Se han realizado pruebas de los resultados que se obtienen cuando los clasificadores consideran conjuntos de roles diferentes para diferentes sentidos de verbos (vs) y cuando no (u), es decir, cuando sólo se considera un clasificador para cualquier verbo y sentido de verbo. El cuadro 5.19 muestra estos resultados. En concreto, la columna `AA` muestra el algoritmo de aprendizaje utilizado; la columna `EA`, la estrategia de anotación seguida, específica para cada sentido de cada verbo o única; y las columnas `P`, `C` y $\mathtt{F}_{\beta=1}$, los valores, expresados en tanto por ciento, relativos a precisión, cobertura y la medida $F_{\beta=1}$, respectivamente.

Los resultados ponen de manifiesto que el algoritmo de aprendizaje utilizado influye en la obtención de mejores o peores resultados. En general, se puede concluir que se obtienen mejores resultados con TiMBL, si bien la cantidad de información, es decir, de características que necesita ME es mucho menor (ver cuadro 5.14 para información sobre las características utilizadas por cada clasificador). El clasificador que hace uso de TiMBL obtiene una medida $F_{\beta=1}$ de 76,34 % con doce características, frente a los mejores resultados alcanzados con ME_2[11] de 74,97 % para la medida $F_{\beta=1}$ para cuatro características.

Puesto que la diferencia no llega a 1,5 puntos de mejora y la cantidad de características utilizadas es 3 veces mayor, se podría pensar que ME puede resultar más beneficioso que TiMBL. Sin

[11] Algoritmo ME mejorado. Realiza un proceso especial de la información proporcionada por las características

embargo, ME es un algoritmo muchísimo más costoso a nivel computacional que TiMBL. El cuadro 5.20 es una clara muestra de ello. Este cuadro muestra para diferentes algoritmos de aprendizaje (AA), y diferentes estrategias de anotación (EA), los tiempos medios de ejecución, en horas:minutos:segundos, atendiendo al número de características utilizadas (CARACTERÍSTICAS). Por ejemplo, las filas tercera y sexta, muestran que utilizar TiMBL con estrategia de anotación única, tarda, de media, un minuto y medio aproximadamente (0:01:34), frente a las casi tres horas de media de ME (2:42:53), cuando se utiliza una característica; o que, las filas segunda, quinta y octava, muestran que utilizar TiMBL con estrategia de anotación por sentidos, tarda, de media, algo más de un minuto (0:01:19), frente a los casi seis minutos de ME (0:05:53), o los diecinueve y pico de ME_2 (0:19:42), cuando se utiliza una característica.

De hecho, utilizando ME como algoritmo de aprendizaje, no se ha podido realizar el proceso de ajuste cuando la estrategia de anotación seguida era un único clasificador para todos los verbos Esta es la razón por la que el cuadro 5.19, no muestra estos resultados. Además, el hecho de que la versión mejorada de ME, tenga aún un coste computacional aún mayor, ha impedido también realizar un proceso de ajuste para cualquiera de las dos estrategias de anotación, sentidos o única. Por esta razón los resultados de ME_2 se han obtenido por sustitución de las características equivalentes que permiten el tratamiento especial requerido.

Universitat d'Alacant
Universidad de Alicante

AA	EA	P (%)	C (%)	$F_{\beta=1}$ (%)
TiMBL	vs	76,91	75,78	76,34
ME	vs	71,33	71,92	71,62
ME_2	vs	75,45	74,51	74,97
TiMBL	u	80,84	81,34	81,09

Cuadro 5.19. Resultados de los clasificadores por sentidos (vs) y únicos (u)

AA	EA	Tiempo	Características
TiMBL	vs	0:01:19	1
	u	0:01:34	1
	vs	0:03:07	2
ME	vs	0:05:53	1
	u	2:42:53	1
	vs	1:44:10	2
ME_2	vs	0:19:45	1
	u	2:50:43	1

Cuadro 5.20. Comparativa de tiempos de ejecución entre algoritmos de aprendizaje y estrategias de anotación

Por otro lado, se podría concluir que se obtienen mejores resultados cuando se adopta una estrategia de anotación única que en el caso contrario. Utilizando TiMBL, la aproximación por sentidos obtiene una medida $F_{\beta=1}$ de 76,34 %, mientras que en la clasificación única se obtiene una medida $F_{\beta=1}$ de 81,09 %. Sin embargo, un estudio más detallado de estos resultados muestra que cuando lo que se quiere clasificar son los argumentos numerados (A1, A2, A3, A4, A5, AA), generalmente, los mejores resultados se obtienen con la estrategia de anotación por sentidos. Por contra, si lo que se desea es anotar adjuntos (LOC, TMP, MOD, etc.), los mejores resultados se suelen obtener con la estrategia de anotación única. Ver cuadro 5.21 para detalle de las medidas $F_{\beta=1}$ para cada tipo de rol.

Por ejemplo, los argumentos con rol A2 obtienen una medida $F_{\beta=1}$ de 81,97 % y 66,92 % en los clasificadores por sentido con TiMBL y ME, respectivamente. Sin embargo, la medida $F_{\beta=1}$ desciende a 73,46 % y 33,48 % para el caso del clasificador único. De la misma manera, un argumento con rol AM-LOC obtiene una medida $F_{\beta=1}$ de 64,62 % y 46,28 % para el clasificador único utilizando TiMBL y ME, respectivamente, y de 41,36 % y 23,26 % si se utiliza el clasificador por sentidos.

Roles	TiMBL(vs)	TiMBL(u)	ME(vs)	ME(u)
A0	83,95	84,93	72,76	76,57
A1	84,49	83,99	73,39	70,27
A2	81,97	73,46	66,92	34,48
A3	74,01	58,82	56,87	24,37
A4	76,80	64,00	60,47	57,36
A5	50,00	50,00	0,00	0,00
AA	100,00	0,00	0,00	0,00
AM-ADV	36,82	55,24	26,49	46,44
AM-CAU	16,00	25,40	6,15	3,64
AM-DIR	64,79	50,00	51,92	40,00
AM-DIS	57,22	85,99	43,28	81,53
AM-EXT	50,00	52,00	59,26	34,48
AM-LOC	41,36	64,62	32,00	46,28
AM-MNR	39,79	54,15	23,26	31,13
AM-MOD	89,79	96,57	80,30	98,59
AM-NEG	76,06	96,62	65,55	88,22
AM-PNC	43,17	39,46	25,60	29,75
AM-PRD	100,00	57,14	50,00	0,00
AM-TMP	48,96	77,52	35,16	61,63

Cuadro 5.21. Comportamiento de los clasificadores para cada tipo de rol cuando se sigue una estrategia de anotación por sentidos del verbo (vs) y cuando no (u). Resultados de $F_{\beta=1}$ medida.

El cuadro 5.22 muestra los valores medios para todos los argumentos numerados y adjuntos atendiendo a las diferentes estrategias de anotación y a los diferentes algoritmos de aprendizaje. Estos datos muestran que en cualquier caso, siempre la estrategia por sentidos es preferible para argumentos numerados, y que la estrategia de anotación única es preferible para los adjuntos. De hecho, los promedios para vs con argumentos numerados es de 78,75 % para TiMBL y 47,20 % para ME, frente a 59,31 % para TiMBL y 37,58 % para ME en el caso de estrategia de anotación u. Además, los promedios para u con adjuntos son de 55,33 % para TiMBL y 41,58 % para ME, frente a 62,90 % para TiMBL y 46,81 % para ME.

Roles	TiMBL(vs)	TiMBL(u)	ME(vs)	ME(u)
Numerados	78,75	59,31	47,20	37,58
Adjuntos	55,33	62,90	41,58	46,81

Cuadro 5.22. Promedios e incrementos de mejora cuando se sigue una estrategia de anotación por sentidos del verbo (vs) y cuando no (u). Resultados de $F_{\beta=1}$ medida.

Finalmente, se han realizado algunas pruebas para medir la influencia del uso de información sintáctica total y parcial, en este proceso de anotación de roles. Para ello, se han añadido a los clasificadores características relativas a la información proporcionada por un analizador sintáctico completo. Como muestra el cuadro 5.23, la información sintáctica completa no solo no mejora este proceso de anotación, si no que incluso en el caso de la clasificación única con TiMBL, los resultados empeoran (81,09 % frente a 80,49 %). La columna `Sint.` del cuadro indica cuando se han utilizado características relativas al análisis sintáctico total y parcial (C+P) y cuando sólo parcial (P).

AA	EA	Sint.	P (%)	R (%)	$F_{\beta=1}$ (%)
TiMBL	vs	P+C	77,19	76,05	76,61
TiMBL	vs	P	76,91	75,78	76,34
TiMBL	u	P+C	80,23	80,74	80,49
TiMBL	u	P	80,84	81,34	81,09

Cuadro 5.23. Influencia del análisis sintáctico en la anotación de roles

Universitat d'Alacant
Universidad de Alicante

De todo lo comentado anteriormente, se puede concluir: i) utilizar un algoritmo de aprendizaje u otro puede afectar a los resultados obtenidos y a los tiempos de ejecución necesitados para la obtención de esos datos; ii) la estrategia de anotación por sentidos es preferible cuando se quieren anotar argumentos numerados, mientras que la estrategia única presenta mejores resultados para la anotación de adjuntos; iii) la información sintáctica total no

aporta beneficios notables a la tarea de identificación de los roles semánticos jugados por los argumentos de un verbo.

5.5.3 Clasificador individual frente Clasificador global

Como se acaba de comentar en el apartado anterior (apartado 5.5.2), los resultados obtenidos respecto a la estrategia de anotación por sentidos o única (cuadro 5.21), establecen que cuando lo que se quiere clasificar son los argumentos numerados (A0, A1, A2, A3, A4, A5, AA), generalmente, los mejores resultados se obtienen con la clasificación por sentidos. Por contra, si lo que se desea es anotar adjuntos (LOC, TMP, MOD, etc.), los mejores resultados se suelen obtener con la clasificación única.

Esta conclusión ha dado lugar al desarrollo de clasificadores específicos para cada tipo de rol atendiendo a las estrategias de anotación más apropiadas en cada caso. En concreto, se ha llevado a cabo un proceso de ajuste de la información significativa para cada tipo de clasificador siguiendo una estrategia de anotación por sentidos, para argumentos numerados, y una estrategia de anotación única, para adjuntos. Los resultados a cerca de este proceso se muestran en el cuadro 5.24.

Rol	AA	EA	P (%)	C (%)	$F_{\beta=1}$ (%)
TMP	TiMBL	u	87,70	79,53	83,41
LOC	TiMBL	u	71,96	68,67	70,26
MOD	TiMBL	u	99,96	99,36	99,66
EXT	TiMBL	u	77,93	67,48	72,08
NEG	TiMBL	u	99,63	98,92	99,27
CAU	TiMBL	u	70,56	32,76	44,71
ADV	TiMBL	u	60,78	63,54	62,13
A2	TiMBL	vs	80,69	81,18	80,92

Cuadro 5.24. Resultados de los clasificadores específicos para cada tipo de rol

Universitat d'Alacant
Universidad de Alicante

Los resultados obtenidos para cada clasificador individual obtienen mejoras significativas con respecto a los obtenidos en la clasificación global, tal y como se muestra en el cuadro 5.25. Por ejemplo, para adjuntos de tipo lugar la mejora es de 5,64 puntos, o para adjuntos de causa, la mejora es de 20,08 puntos.

Rol	Global$_{vs}$ ($F_{\beta=1}$)	Global$_{u}$ ($F_{\beta=1}$)	Individual ($F_{\beta=1}$)	Mejora$_{vs}$	Mejora$_{u}$
LOC	41,36	64,62	70,26	28,90	5,64
TMP	48,96	77,52	83,41	34,45	5,89
EXT	50,00	52,00	72,08	22,08	20,08
NEG	76,06	96,62	99,27	23,46	2,65
MOD	89,79	96,57	99,66	9,87	3,09
ADV	36,82	55,24	62,13	25,31	6,89
CAU	16,00	25,40	44,71	28,71	19,31
A2	81,97	73,46	80,92	-	7,46

Cuadro 5.25. Comparativa de los resultados obtenidos con clasificadores individuales frente a los globales.

Los resultados mostrados permiten concluir que el hecho de diseñar clasificadores individuales que hacen uso de información útil para cada tipo de rol diferente, repercute en una mejora significativa de los resultados para cada uno de esos clasificadores, respecto al clasificador obtenido como resultado de un proceso de ajuste global para todos los tipos de roles como un todo.

Universitat d'Alacant
Universidad de Alicante

5.5.4 Comparación con otros sistemas de anotación

Para terminar, se mostrarán resultados comparativos entre SemRol y otros sistemas de sus mismas características. En concreto, la comparación se hará con los sistemas presentados en la *shared task* CoNLL 2004[12].

[12] No se comparan los resultados con la shared task CoNLL 2005 porque la información utilizada por SemRol sólo hace uso de análisis sintáctico parcial.

Para que los resultados sean realmente comparables, se han utilizado los mismos corpus de entrenamiento y test que en la citada conferencia: secciones 15-18 para entrenamiento y 21 para test, del corpus PropBank.

El cuadro 5.26 muestra un resumen de los resultados obtenidos por sistemas participantes en la *shared task* del CoNLL 2004, atendiendo únicamente a la determinación del rol que juega un determinado argumento, y suponiendo, por tanto, argumentos conocidos. Como muestra dicho cuadro, el clasificador único de SemRol obtiene resultados realmente competitivos, con respecto al resto de sistemas. En concreto, aporta una mejora de 2,57 puntos en la medida $F_{\beta=1}$ con respecto al mejor de los sistemas (SemRol 77,99 frente a (Hacioglu *et al.*, 2004) 75,42).

Por otro lado, la versión de SemRol que sigue una estrategia de anotación por sentidos también presenta unos resultados bastante aceptables, situando a SemRol justo en la mitad del cuadro.

Sistema	P (%)	C (%)	$F_{\beta=1}$ (%)
SemRol$_u$	**77.75**	**78.23**	**77.99**
(Hacioglu *et al.*, 2004)	78.61	72.47	75.42
(Punyakanok *et al.*, 2004)	77.82	70.04	73.72
(Carreras & Màrquez, 2004)	79.22	67.41	72.84
(Park *et al.*, 2004)	73.64	70.05	71.80
(Lim *et al.*, 2004)	75.43	67.76	71.39
SemRol$_{vs}$	**72.97**	**69.31**	**71.10**
(Higgins, 2004)	70.72	63.40	66.86
(van den Bosch *et al.*, 2004)	75.48	61.23	67.61
(Kouchnir, 2004)	66.52	58.43	62.21
(Baldewein *et al.*, 2004a)	75.13	48.70	59.09
(Williams *et al.*, 2004)	70.62	42.25	52.87

Cuadro 5.26. Comparación de SemRol con otros sistemas de SRL

Estos resultados ponen de manifiesto, una vez más, la obligatoriedad de llevar a cabo en proceso de ajuste de la información a utilizar, en procesos de clasificación automática.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

6. Los Roles Semánticos en aplicaciones de Búsqueda de Respuestas

Si bien son muchas las áreas de PLN, como por ejemplo extracción de información, resúmenes, o implicación textual, en las que se considera que la aportación de los roles semánticos puede ser interesante (S. Wen-tau Yih and K. Toutanova, 2006), los primeros estudios importantes sobre asignación automática de roles (Gildea & Jurafsky, 2002) ya indicaban que una de las áreas en las que los roles semánticos tendrían su contribución más destacable sería en los sistemas de búsqueda de respuestas (BR) (en inglés, *Question Answering* -QA-). La razón de tal suposición se debe al hecho, tal y como se muestra a continuación, de que ambos, roles semánticos y sistemas de BR, plantean objetivos complementarios.

Un sistema de BR tiene interés en encontrar la respuesta a preguntas como las mostradas en los ejemplos (E47) a (E51), en oraciones como las mostradas en los ejemplos (E52) a (E59):

Universitat d'Alacant
Universidad de Alicante

(E47) Who hit John with a baseball yesterday in the park?

(E48) Whom did Mary hit with a baseball?

(E49) What did Mary hit John with?

(E50) When did Mary hit John with a baseball in the park?

(E51) Where did Mary hit John with a baseball yesterday?

(E52) Mary hit John with a baseball yesterday in the park

(E53) Yesterday, Mary hit John with a baseball in the park

(E54) John was hit by Mary yesterday with a baseball in the park

(E55) Yesterday, John was hit with a baseball by Mary in the park

(E56) With a baseball, Mary hit John yesterday in the park

(E57) Yesterday John was hit by Mary with a baseball in the park

(E58) In the park, Mary hit John with a baseball yesterday

(E59) Mary hit John with a baseball in the park yesterday

Universitat d'Alacant
Universidad de Alicante

Es decir, su objetivo es averiguar "quién hizo qué a quién, dónde, cuándo, cómo y por qué" (Hacioglu & Ward, 2003).

Por otro lado, los roles semánticos permiten interpretar los textos determinando las relaciones semánticas entre las entidades y los eventos en los que participan. Ejemplos de roles semánticos son agente, paciente, beneficiario, etc., o también adjuntos, como causa, manera o temporal. De esta manera, las oraciones anteriores se representarían como se muestra a continuación en las oraciones (E60) a (E67).

(E60) [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{INSTRUMENT}$ with a baseball] [$_{TEMP}$ yesterday] [$_{LOC}$ in the park]

(E61) [$_{TEMP}$ Yesterday], [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{INSTRUMENT}$ with a baseball] [$_{LOC}$ in the park]

(E62) [$_{PACIENT}$ John] was hit [$_{AGENT}$ by Mary] [$_{TEMP}$ yesterday] [$_{INSTRUMENT}$ with a baseball] [$_{LOC}$ in the park]

(E63) [$_{TEMP}$ Yesterday], [$_{PACIENT}$ John] was hit [$_{INSTRUMENT}$ with a baseball] [$_{AGENT}$ by Mary] [$_{LOC}$ in the park]

(E64) [$_{INSTRUMENT}$ With a baseball], [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{TEMP}$ yesterday] [$_{LOC}$ in the park]

(E65) [$_{TEMP}$ Yesterday] [$_{PACIENT}$ John] was hit [$_{AGENT}$ by Mary] [$_{INSTRUMENT}$ with a baseball] [$_{LOC}$ in the park]

(E66) [$_{LOC}$ In the park], [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{INSTRUMENT}$ with a baseball] [$_{TEMP}$ yesterday]

(E67) [$_{AGENT}$ Mary] hit [$_{PACIENT}$ John] [$_{INSTRUMENT}$ with a baseball] [$_{LOC}$ in the park] [$_{TEMP}$ yesterday]

Es decir, el conocimiento que representan los roles semánticos puede ayudar a determinar, capturar y representar "quién hizo qué a quién, dónde, cuándo, cómo y por qué" en una oración, y por tanto ayudar a dar respuesta a las preguntas anteriores con cualquiera de las oraciones, tal y como se muestra en los ejemplos de la figura 6.1.

Aunque ya se han realizado algunas propuestas en esta línea, como se verá en el apartado 6.2, es mucho el trabajo que aún queda por realizar hasta alcanzar un análisis profundo de la influencia de los roles semánticos en esta clase de sistemas. Por ello, este trabajo se centra precisamente en aportar información y con-

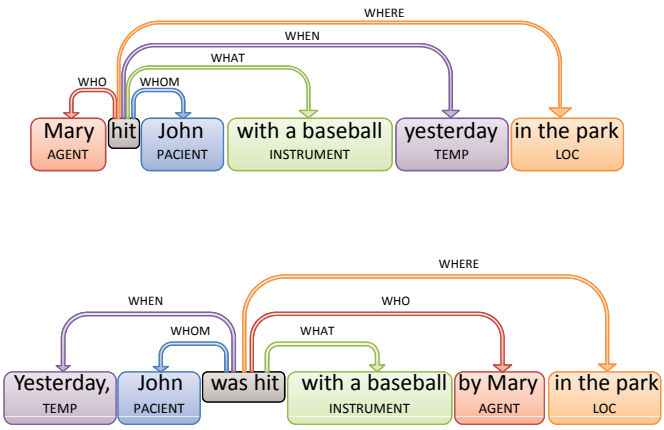


Figura 6.1. Uso de roles semánticos en búsqueda de respuestas.

clusiones sobre la influencia de los roles semánticos en sistemas de BR.

Para ello, en primer lugar se hará un breve resumen de los aspectos generales más importantes relacionados con sistemas de BR (apartado 6.1). Posteriormente se analizarán las contribuciones llevadas a cabo hasta la fecha respecto al uso de roles semánticos en sistemas de BR (apartado 6.2). A continuación, se presentará el sistema de BR desarrollado en este trabajo, haciendo especial hincapié en cómo dicho sistema hace uso de los roles semánticos (apartado 6.3), y se analizarán los resultados obtenidos con dicho sistema (apartado 6.4). Para finalizar se mostrará un ejemplo del proceso de creación automática de patrones semánticos (apartado 6.5).

Universitat d'Alacant
Universidad de Alicante

6.1 Sistemas de Búsqueda de Respuesta

Ante el enorme crecimiento de la información disponible en la red, los sistemas de BR han pasado a jugar un papel relevante, convirtiéndose en los potenciales sucesores de los buscadores tradicionales de información, como pueden ser Google o Yahoo. Frente

al objetivo de estos sistemas tradicionales de devolver una relación de documentos ordenada en función de su relevancia respecto a la pregunta realizada por el usuario, los sistemas de BR persiguen un objetivo mucho más ambicioso: devolver el trozo de texto que contiene la información requerida por el usuario, no un documento completo. Además, es importante también destacar que mientras que los buscadores tradicionales requieren como entrada una secuencia de palabras clave, los sistemas de BR trabajan con la pregunta expresada en su forma completa en lenguaje natural (Ferrández, 2003).

Realizando un repaso de los sistemas de BR desarrollados hasta el momento se puede determinar una arquitectura general a la que se suelen ajustar la mayoría de estos sistemas. Esta arquitectura se podría resumir en cinco módulos (Ferrández, 2003), siendo dos de ellos, la recuperación de documentos y la extracción de respuestas, los módulos núcleo (Saggion & Gaizauskas, 2006) de cualquier sistema. Veamos a continuación cada uno de estos módulos:

- *Análisis de la pregunta.* Con el fin de extraer cualquier información que pueda ser útil para el resto de los módulos, el primer paso consiste en llevar a cabo el análisis de la pregunta (Mollá, 2006). La información a extraer suele ser :
  - El tipo de la pregunta, y asociado a él, el tipo de la respuesta esperada. Por ejemplo, la pregunta *"When did Mary hit Scott with a baseball?"* requiere como respuesta una expresión temporal, o la pregunta *"Where did Mary hit Scott with a baseball yesterday?"* requiere como respuesta una expresión de lugar. Para llevar a cabo esta tarea se han desarrollado diferentes taxonomías de preguntas, como por ejemplo la taxonomía jerárquica de (Li & Roth, 2002).
  - El foco de la pregunta. A la hora de determinar el tipo de respuesta esperada, el foco de la pregunta juega un papel fundamental. Generalmente, el foco es una expresión en la pregunta que ayuda a determinar el tipo de repuesta esperada, y que se caracteriza porque no suele aparecer en la oración que contiene la respuesta. Por ejemplo, en la pregunta *"In*

what year died Cristobal Colon?", el foco es la palabra *year*, la cual determina que la respuesta esperada sea una expresión temporal. Además, ocurre que dicha palabra no es usual que aparezca en oraciones que contengan la respuesta.

- Información acerca del contenido de la pregunta, como por ejemplo el conjunto de términos de la pregunta que pueden ayudar a localizar documentos candidatos a contener la respuesta, ampliación de la pregunta realizada por el usuario con palabras relacionadas semánticamente, reformulación de los términos de la pregunta, o el tema de la pregunta.

- *Recuperación de documentos.* Teniendo en cuenta que las técnicas de comprensión e interpretación de textos que nos permiten extraer la respuesta suelen tener un coste computacional elevado, es deseable reducir el volumen de texto sobre el que aplicarlas. La herramienta natural para llevar a cabo este proceso de selección de documentos suele ser un sistema de recuperación de información. Estos sistemas devuelven un conjunto de documentos supuestamente relevantes para la pregunta en cuestión, elegidos de entre el conjunto de documentos completo disponible para la tarea.

  Es importante destacar que dependiendo de si los documentos se buscan bien en la Web o en un corpus finito de documentos, bien en ontologías o bases de conocimiento, los sistemas de BR son catalogados como sistemas de dominio abierto o restringido, respectivamente.

- *Selección de párrafos.* Con el fin de eliminar aquellos documentos considerados como imposibles para contener la respuesta, los documentos devueltos por el módulo anterior son procesados, generalmente, mediante técnicas de procesamiento de lenguaje natural. Hasta la fecha, el método más común para ello es eliminar aquellos que no contengan instancias de texto del tipo de la respuesta esperada. Es decir, se etiquetan las entidades nombradas (en inglés, *Named Entity* -NE-) en las oraciones, y se eliminan aquellas que no tengan entidades correspondientes al tipo de respuesta esperado.

- *Extracción de la respuesta.* Determinar qué parte de las oraciones seleccionadas en el módulo anterior constituyen una respuesta es el módulo más subjetivo de todos dentro de un sistema de BR. Hasta el momento, el método más sencillo consiste en devolver el texto etiquetado como una NE del tipo esperado.
- *Clasificación de respuestas candidatas.* Normalmente el sistema extrae varias respuestas candidatas, por lo que se hace obligatorio establecer un método de ordenación de las mismas atendiendo a su probabilidad de contener la respuesta. Se han utlizado diferentes criterios para conseguir este objetivo, como por ejemplo (Mollá, 2006):
  - Similitud. Recompensar respuestas en contextos similares.
  - Popularidad. Recompensar respuestas populares.
  - Patrones. Utilizar patrones específicos de pares pregunta-respuesta de manera que si la pregunta satisface un patrón determinado, se le da prioridad a respuestas candidatas que satisfagan el patrón de respuesta asociado.
  - Validación. Comprobar que la respuesta tiene valores aceptables mediante, por ejemplo, el uso de reglas, consulta de ontologías o consultas en la Web (Magnini *et al.*, 2002).

6.2 Uso de roles semánticos en sistemas de BR

Como se ha comentado, desde el primer sistema de SRL automático ya se proponía el uso de roles semánticos como una ayuda a la hora de encontrar la respuesta en sistemas de BR. Desde entonces, muchos autores han recalcado y justificado el desarrollo de herramientas SRL respaldándose en su aplicación a esta clase de sistemas de BR. Sin embargo, no es hasta el año 2004 cuando tal propuesta toma forma y se muestran los primeros resultados. A partir de ese momento, y como se mostrará a continuación, varias propuestas se han llevado a cabo utilizando roles semánticos, bien como complemento a otros métodos, bien como método base para la búsqueda de las respuestas.

Todos los sistemas que se analizarán a continuación tendrán una arquitectura más o menos próxima a la arquitectura presentada en el apartado 6.1. Puesto que los sistemas de BR en sí mismos no son un objetivo de este trabajo, el análisis que aquí se muestra se realizará atendiendo únicamente al uso que estos sistemas hagan de los roles semánticos. Se analizarán, por tanto, dos aspectos principales: el conjunto de roles utilizado y el papel concreto que éstos han tenido dentro del sistema de BR.

Un breve resumen de la información a nivel general para cada uno de los sistemas puede consultarse en el cuadro 6.1. Este cuadro muestra, para cada sistema, si es de dominio abierto o restringido (columna dominio), las preguntas a las que es capaz de dar respuesta (columna pregunta), el conjunto de documentos en los que busca la respuesta (columna documentos), si alguna parte o todo el proceso se realiza de forma manual o por el contrario es completamente automático (columna proceso), y si los roles constituyen la base principal del sistema de BR o son un complemento a otro método (columna núcleo).

Los siguiente apartados analizarán en detalle el conjunto de roles utilizado por cada uno de estos sistemas, así como el uso que de ellos hayan hecho a la hora de afrontar la tarea de la búsqueda de respuestas.

Universitat d'Alacant
Universidad de Alicante

6.2.1 Conjunto de roles semánticos utilizados

La mayoría de los sistemas han utilizado los corpus PropBank y FrameNet, ya sea:

- de forma individual
  - PropBank (Stenchikova *et al.*, 2006; Sun *et al.*, 2005; Melli *et al.*, 2006; Moschitti *et al.*, 2007)
  - FrameNet (Ofoghi *et al.*, 2006; Shen *et al.*, 2007; Frank *et al.*, 2007; Fliedner, 2007)

Sistema	Dominio	Pregunta	Documentos	Proceso	Núcleo
(Narayanan & Harabagiu, 2004)	Restrin.	Complejas	AQUAINT[1]	Manual	Roles
(Sun *et al.*, 2005)	Abierto	Factual	Web	Auto.	NE
(Stenchikova *et al.*, 2006)	Abierto	Factual	Web	Auto.	Roles
(Ofoghi *et al.*, 2006)	Abierto	Factual	AQUAINT	Manual	Roles
(Lo & Lam, 2006)	Abierto	Factual	TREC, Wikipedia	Auto.	NE
(Melli *et al.*, 2006)	Abierto	Complejas	TREC	Auto.	Roles
(Kaisser, 2007)	Abierto	Factual	Web	Auto.	Roles
(Shen *et al.*, 2007)	Abierto	Factual	TREC, Wikipedia	Auto.	NE
(Moschitti *et al.*, 2007)	Abierto	Descripción	Web	Auto.	Roles
(Fliedner, 2007)	Abierto	Factual	Textos noticias	Auto.	Roles
(Frank *et al.*, 2007)	Restrin.	Definición, Si-No	Base de conocimiento	Manual	Roles

Cuadro 6.1. Resumen de las principales características de los sistemas de BR que hacen uso de roles semánticos

- o una combinación de ambos (Narayanan & Harabagiu, 2004; Kaisser, 2007).

Otros sistemas han seleccionado un subconjunto de los conjuntos de roles proporcionados por estos corpus, como (Lo & Lam, 2006) que sólo utiliza los roles de PropBank A0, A1 y A2.

Universitat d'Alacant
Universidad de Alicante

6.2.2 Papel de los roles semánticos

Respecto al papel que los roles han jugado en los sistemas de BR, se podrían distinguir dos grandes grupos, los que utilizan los roles semánticos sólo como un apoyo o complemento a otros métodos, como pueden ser las entidades nombradas; o bien los que basan su sistema de BR, ya sea de forma exclusiva o principal, en los roles semánticos.

Roles como complemento. Algunos sistemas de BR basados en reconocimiento de entidades han realizado ampliaciones mediante el uso de roles semánticos. En concreto,

- En (Sun *et al.*, 2005), se utilizan los roles única y exclusivamente cuando el sistema no es capaz de obtener una respuesta mediante entidades, situación en la que se mide la similitud entre los conjuntos de argumentos de la pregunta y de las respuestas candidatas;
- los sistemas (Lo & Lam, 2006) y (Shen *et al.*, 2007) establece una clasificación sobre la lista de respuestas candidatas obtenidas también mediante el uso de entidades, atendiendo a criterios de similitud entre roles;
- finalmente, la propuesta de (Melli *et al.*, 2006) plantea un sistema de BR que responde a las preguntas realizando resúmenes de los documentos. Para ello propone utilizar los roles sobre el conjunto de respuestas candidatas obtenidas mediante entidades, tanto a la hora de construir *clusters* de oraciones candidatas atendiendo a medidas de similitud, como a la hora de establecer una clasificación de la lista de respuestas candidatas.

En cualquier caso, estos usos no permiten llevar a cabo un estudio real de cuál es la influencia de los roles semánticos en sistemas de BR. El hecho de que cualquiera de las propuestas anteriores aporte o no una mejora en el sistema de BR, sólo sirve para establecer la mejor manera en la que los roles semánticos pueden complementar a las entidades nombradas.

Roles como núcleo. Otros sistemas de BR han centrado el peso principal de alguna subtarea de los mismos en los roles semánticos. Un resumen de las principales características de estos sistemas puede verse en el cuadro 6.2. En dicho cuadro se indica para cada sistema, el conjunto de roles utilizado, ya sea PropBank (PB) o FrameNet (FN) (columna roles), el módulo o tarea del sistema de BR en el que se han utilizado los roles semánticos (columna uso) y de qué manera se han utilizado (columna método).

Veamos a continuación con algo más de detalle estos sistemas.

Sistema	Roles	Uso	Método
(Narayanan & Harabagiu, 2004)	FN	Tipo resp.	Mapeo Patrón preg. − Patrón resp.
(Stenchikova *et al.*, 2006)	PB	Extrac. resp.	Reglas tipo preg. − Rol resp.
(Ofoghi *et al.*, 2006)	FN	Extrac. resp.	Mapeo Patrón preg. − Patrón resp.
(Kaisser, 2007)	FN,PB	Extrac. resp.	Mapeo Patrón preg. − Patrón resp.
(Moschitti *et al.*, 2007)	PB	Tipo preg., Tipo resp., Reclas. resp.	Clasificador basado en aprendizaje automático supervisado
(Fliedner, 2007)	FN	Extrac. resp.	Mapeo *frame* preg. − *frame* resp.

Cuadro 6.2. Resumen del uso de roles semánticos en sistemas de BR

- La propuesta de (Narayanan & Harabagiu, 2004), primera en hacer uso de los roles semánticos en sistemas de BR, plantea un sistema para resolución de preguntas complejas (que conciernan a aspectos temporales y causales de eventos complejos) por división en preguntas más sencillas, haciendo para ello uso de información semántica y mecanismos de inferencia. Con este objetivo se plantea un uso doble de los roles semánticos. Por un lado, utilizar las estructuras predicado-argumento de la pregunta para determinar el modelo del tema de la misma; y por otro, utilizar los marcos semánticos, o *frames*, relacionados con la pregunta y los pasajes relevantes, con el fin de determinar el tipo de la respuesta y extraer las respuestas esperadas, lo cual implica el uso de inferencias complejas sobre estructuras de evento y de causa derivadas manualmente de los *frames*. Los resultados mostrados[2] indican que en el 73,5 % de las veces las estructuras para inferencia desarrolladas a partir de la información semántica ayudaron a detectar de forma correcta el tipo de respuesta en cuestiones complejas. La evaluación se realizó so-

[2] Los únicos resultados mostrados se refieren a la detección del tipo de respuesta, dejando como trabajo futuro la evaluación del proceso de extracción de la respuesta.

bre un subconjunto de 400 preguntas creadas específicamente para este trabajo y restringidas a cuatro temas concretos.

- El sistema propuesto en (Stenchikova *et al.*, 2006), denominado QASR, también realiza un doble uso de los roles semánticos: búsqueda de documentos y extracción de la respuesta. En primer lugar, la búsqueda en la Web de documentos conteniendo la respuesta se realiza atendiendo a los argumentos detectados en la pregunta, siempre y cuando una búsqueda utilizando la transformación de la pregunta a su forma enunciativa, no devolviera nada. En segundo lugar, la extracción de respuestas candidatas se realiza o bien mediante el uso de un conjunto de reglas que relacionan preguntas del tipo *who*, *when*, o *where* con tipos de roles esperados como respuesta, o bien mediante el uso de un clasificador para el caso de preguntas de tipo *what*. Debido a la herramienta SRL utilizada, QASR presenta limitaciones a la hora de trabajar con oraciones que contengan el predicado *to be*. Los mejores resultados obtenidos con este sistema muestran una precisión de 30 %, y un *Mean Reciprocal Ranking* (MRR)[3] de 35 %. La evaluación se realizó utilizando un subconjunto de 190 preguntas del TREC-9.

- En (Ofoghi *et al.*, 2006) se realiza una prueba manual sobre un conjunto de 15 preguntas, con el objetivo de extraer respuestas candidatas a una pregunta haciendo uso de los roles semánticos. Para ello se lleva a cabo un proceso de mapeo entre la información semántica de la pregunta y la respuesta. Este proceso en tres pasos, supone primero, detectar los marcos semánticos, o *frames*, evocados por la pregunta y la respuesta; segundo, determinar los elementos de *frame*, tanto para la pregunta como la respuesta; y, por último, extraer el contenido del elemento de *frame* de la respuesta que concuerda con el elemento de *frame* ausente en la pregunta. Resultados utilizando un conjunto de 15 preguntas del TREC2004 y su correspondiente colección

[3] MRR asigna a cada pregunta una puntuación igual a la inversa de la posición de la primera pregunta correcta, dentro de la lista ordenada de posibles respuestas, o 0 en caso de no encontrarse la respuesta correcta entre las cinco primeras respuestas.

de documentos del corpus AQUAINT, muestran una MRR de 38,89 %[4]

- En el sistema Alyssa (Kaisser, 2007) se presenta una propuesta similar a la anterior puesto que para determinar respuestas candidatas también se lleva a cabo un mapeo que busca el rol ausente en la pregunta, en los roles de respuestas candidatas. Concretamente, en este sistema se propone determinar el rol buscado como respuesta, o bien mapeando la pregunta contra las estructuras del verbo en cuestión proporcionadas por FrameNet, PropBank o VerbNet, cuando el rol buscado no es un adjunto; o bien mediante el uso de reglas en caso contrario. Sin embargo, a diferencia de otras propuestas similares, el mapeo se realiza por posiciones y no por tipo de rol. Esto obliga tanto a generar variaciones de las estructuras del verbo atendiendo a voz y tiempos verbales, como al desarrollo de las reglas mencionadas para el caso de los adjuntos. Como aspecto positivo, la generación de estas variaciones tiene en cuenta las relaciones entre *frames* de FrameNet, lo que resulta en un mayor número de variaciones al tener en cuenta sinónimos y palabras semánticamente relacionadas con el predicado en cuestión. Por otro lado, este proceso se complementa, en el caso de no obtener respuesta, con procesos de mapeo entre las oraciones candidatas y las oraciones ejemplo proporcionadas por cualquiera de los corpus utilizados. Esta comparación utiliza reglas basadas en roles y asigna pesos a las oraciones, de manera que se devolverá como respuesta la oración de mayor peso. La propuesta obtiene una precisión de 36,70 % utilizando un subconjunto de 264 preguntas del TREC2002 que no contenían al predicado *to be*.

- La propuesta de (Moschitti *et al.*, 2007) realiza pruebas de la influencia de la información sintáctica y semántica en cuanto a clasificación de preguntas, clasificación de respuestas y reclasificación de respuestas. En concreto, analiza la utilidad de la infor-

[4] Estos resultados se refieren al uso exclusivo de roles semánticos. La propuesta incluye un proceso adicional para los casos en los que los *frames* de la pregunta y la respuesta no sean comparables, recurriendo al uso de entidades: MRR 22,44 %. MRR total 61,33 %.

Universitat d'Alacant
Universidad de Alicante

mación sobre estructuras predicado-argumento, en cada una de las tareas indicadas. Para ello, utiliza clasificadores automáticos supervisados que hacen uso de información obtenida a partir del árbol de análisis semántico, es decir, del árbol formado por el predicado de la oración y sus argumentos etiquetados con roles semánticos. Los resultados obtenidos demuestran la utilidad de esta información semántica en la tarea de clasificación (MRR 56,21 %) y reclasificación (MRR 81,12 %) de respuestas, pero no en la de clasificación de preguntas, debido principalmente a la dificultad de anotar con roles semánticas preguntas y a los problemas de anotación de la herramienta SRL utilizada, con el verbo *to be*. Para las pruebas se utilizaron las preguntas del TREC201.

- En (Fliedner, 2007) se propone la representación, tanto de la pregunta como de los documentos en los que buscar la respuesta, en forma de estructuras al estilo de FrameNet. La respuesta se obtiene mediante un proceso de mapeo entre ambas estructuras. El conjunto de documentos sobre los que se ha trabajado corresponde a textos de noticias, aunque el sistema no es dependiente de dominio. Las pruebas realizadas demostraron una precisión de 66 % y una cobertura de 33 %.
- Por último, en el área de la traducción de lenguaje natural a sentencias SQL, y restringido a los dominios de los ganadores de premios nobel o de la tecnología del lenguaje, el sistema Quetal (Frank *et al.*, 2007) utiliza un conjunto de reglas manualmente definidas, que permiten mapear los *frames* y sus roles, manualmente asignados, a tablas y columnas de una base de conocimiento.

El cuadro 6.3[5] muestra un resumen de los resultados obtenidos, ya sea respecto a precisión o MRR, dependiendo de los datos facilitados por los autores.

[5] El objetivo de este cuadro no es más que resumir la información de los sistemas analizados, relativa a los resultados obtenidos, y no mostrar una comparación de tales resultados, ya que como se comentará a continuación una comparación directa entre estos sistemas no es posible.

Sistema	Precisión (%)	MRR (%)
(Narayanan & Harabagiu, 2004)	73,50	
(Stenchikova *et al.*, 2006)		30,00
(Ofoghi *et al.*, 2006)		38,89
(Kaisser, 2007)	36,70	
(Moschitti *et al.*, 2007)		56,21 % Clasif. resp. 81,12 % Reclasif. resp.
(Fliedner, 2007)	66,00	

Cuadro 6.3. Resultados del uso de roles semánticos en sistemas de BR

6.2.3 Principales conclusiones

Una vez analizados estos trabajos, parece obvio que la posible aportación de los roles semánticos en sistemas de BR se centra en el módulo de extracción de la respuesta[6]. Sin embargo, tales trabajos no permiten extraer ideas concluyentes, debido principalmente a dos razones. Por un lado, el uso de conjuntos de preguntas y documentos, así como de roles, diferentes, hace que los sistemas presentados se sitúen en contextos diferentes y que por tanto, no puedan ser directamente comparados. Por otro, el hecho de que estos sistemas se hayan desarrollado sobre conjuntos de preguntas cuya respuesta esperada es una entidad nombrada, dificulta la tarea de medir la aportación real de los roles semánticos a esta clase de sistemas.

Universitat d'Alacant
Universidad de Alicante

Respecto a los diferentes contextos de los sistemas presentados, existen:

- Tantos conjuntos de preguntas diferentes como sistemas se han analizado. En (Narayanan & Harabagiu, 2004) se crea un conjunto de preguntas específico para su objetivo, en (Stenchikova *et al.*, 2006) se trabaja con un subconjunto de las preguntas

[6] Recordemos que aunque Narayanan (2004) presenta resultados únicamente del tipo de respuesta, su objetivo final no es otro que la extracción de la respuesta, y estos resultados referentes al tipo de respuesta son sólo un paso intermedio en el proceso.

utilizadas en la edición TREC-9 del año 2000, en (Ofoghi *et al.*, 2006) con el conjunto de preguntas del TREC2002, en (Moschitti *et al.*, 2007) del TREC2001, y (Kaisser, 2007) del TREC2004.

- Diferentes conjuntos de documentos sobre los que buscar las respuestas candidatas. Si bien, en (Narayanan & Harabagiu, 2004) y (Ofoghi *et al.*, 2006) se utiliza el corpus AQUAINT, en (Kaisser, 2007; Stenchikova *et al.*, 2006; Moschitti *et al.*, 2007) se realizan las búsquedas en la Web, y en (Fliedner, 2007) se trabaja sobre textos de noticias.
- Dos conjuntos de roles con los que anotar preguntas y/o respuestas, diferentes y con criterios de anotación y coberturas diferentes, PropBank y FrameNet.

Todo ello hace imposible realizar comparaciones entre sistemas que permitan establecer de forma clara cuándo un determinado uso de los roles semánticos puede ser más o menos beneficioso que otro.

Respecto al tipo de respuesta esperada, como bien se señala en (Stenchikova *et al.*, 2006), todos los sistemas trabajan sobre conjuntos de preguntas cuya respuesta es una entidad nombrada. En estos casos, la aportación de los roles semánticos no puede ser directamente medida ni comparada con sistemas basados en entidades, que claramente obtendrán mejores resultados. Es más, casi parece que intentar utilizar los roles semánticos en un entorno poco apropiado, puede carecer de sentido. Ahora bien, qué pasaría si consideramos preguntas como la mostrada en el ejemplo (E68). En este caso, la respuesta no es una entidad nombrada, y por lo tanto, ninguno de los sistemas basados en entidades sería capaz de encontrar la respuesta. Sin embargo, para un sistema basado en roles semánticos, sería muy sencillo dado que bastaría con localizar el argumento del verbo que juegue el rol cosa_creada (ver ejemplo (E69).

(E68) What did Antonio Meucci invent? The telephone.

(E69) [$_{A0:creador}$ Antonio Meucci] invented [$_{A1:cosa_creada}$ the telephone] [$_{TMP}$ in 1876].

Por todo ello el trabajó aquí presentado intenta establecer un contexto adecuado y estable que permita determinar de manera concluyente y justificada i) cuál es la verdadera aportación de los roles semánticos a los sistemas de BR, ii) bajo qué circunstancias el aporte de los roles semánticos a los sistemas de búsqueda de respuestas es preferible frente al proporcionado por las entidades nombradas.

6.3 SemRol en sistemas de BR

Como se acaba de comentar, uno de los objetivos de esta Tesis es analizar y demostrar la validez de los roles semánticos en aplicaciones de PLN, y más concretamente en sistemas de BR. Para ello, se llevará a cabo el desarrollo de un prototipo de un sistema de BR cuyo módulo de extracción de respuestas haga uso de la información proporcionada por los roles semánticos. Este uso de los roles semánticos se afrontará desde dos perspectivas diferentes: i) el uso de reglas semánticas que relacionen tipos de preguntas con tipos de roles semánticos, ii) el uso de patrones semánticos que hagan uso de la información proporcionada por los roles semánticos.

A continuación se presentarán las principales características del prototipo desarrollado, haciendo especial hincapié en su módulo de extracción de respuestas basado en roles semánticos.

Universitat d'Alacant
Universidad de Alicante

6.3.1 Sistema de BR desarrollado

El sistema de BR desarrollado en este trabajo, sigue las indicaciones establecidas en (Pizzato & Mollá-Aliod, 2005) respecto a la creación de un sistema de BR que trabaja en dominios no restringidos, haciendo uso de los resultados de motores de búsqueda disponibles en la red. El sistema, cuya arquitectura se muestra

en la figura 6.2, se compone de cinco módulos, los mismos que tendría cualquier otro sistema de BR: análisis de la pregunta, recuperación de documentos, selección de *snippets*, extracción de respuestas, y clasificación de respuestas.

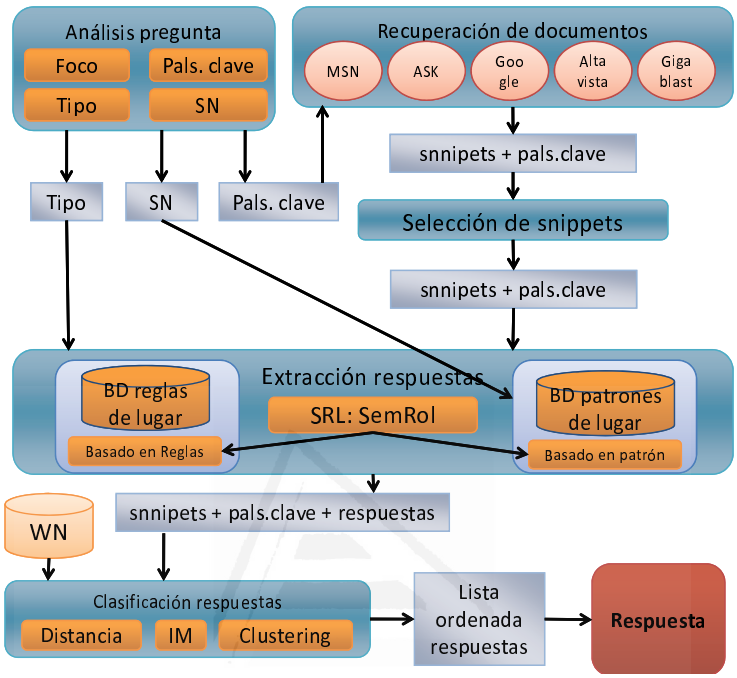


Figura 6.2. Arquitectura de un sistema de BR basado en roles semánticos.

Universitat d'Alacant
Universidad de Alicante

Veamos a continuación cada uno de estos módulos con más detalle:

- *Análisis de la pregunta.* A partir de un conjunto de reglas manualmente definidas y del análisis sintáctico de la pregunta, se determina:

  - Tipo de pregunta. Ampliando la propuesta de (Mollá, 2003) se han utilizado un conjunto de reglas que determinan que el tipo de la pregunta, y por tanto el tipo de la respuesta esperada, puede ser uno de entre: *person, date, location, number,*

organization. En concreto, se ha utilizado un conjunto de 35 reglas. Un detalle de las reglas utilizadas para la detección de preguntas del tipo *location* puede consultarse en la figura 6.3.

```
/^[¿]*[Ww]here('s)?
/^[¿]*(\w+ )?[Ww](hat|hich)('s)? (\w+ )*(town|province)(s)?
/^[¿]*(\w+ )?[Ww](hat|hich)('s)? (\w+ )*(cit)(ylies)
/^[¿]*(\w+ )?[Ww](hat|hich)('s)? (((\w(.)?)*)+ )*(state(s)?|communit(ylies))
/^[¿]*(\w+ )?[Ww](hat|hich)('s)? (\w+ )*capital(s)? (city )?of
/^[¿]*(\w+ )?[Ww](hat|hich)('s)? (\w+ )*(count(r)?(ylies)|nation(s)?)
/^[¿]*(\w+ )?[Ww](hat|hich)('s)? (\w+ )*continent(s)?
/^[¿]*(\w+ )?[Ww](hat|hich)('s)? (\w+ )*(place|arealsite)(s)?
/^[¿]*[Ww](hat|hich)('s)? (\w+ )*team(s)? (\w+ )*world cup
/[Rr]iver(s)?
/[Mm]ountain(s)?
/([sS]eal[Oo]cean)(s)?
/([Bb]each|[Cc]oast)(s)?
/[Ii]sland(s)?
/^[¿]*[Ww](hat|hich)('s)? world(s)?
/([Cc]athedral|[Mm]useum)(s)?
```

Figura 6.3. Reglas utilizadas para identificar las preguntas de tipo lugar.

- Lista de palabras clave. A partir de la información facilitada por Minipar[7], un analizador sintáctico basado en dependencias, se extraen los sintagmas nominales, una vez se hayan eliminado tanto las *stopwords*[8], como el verbo principal de la oración.
- Foco. Nodo raíz del árbol sintáctico devuelto por Minipar[9]. Una vez detectado, se elimina de la lista de palabras clave ya que es poco probable que aparezca en oraciones candidatas a contener la respuesta.

[7] http://www.cs.ualberta.ca/ lindek/minipar.htm Consultado en marzo 2008
[8] http://members.unine.ch/jacques.savoy/clef/ Consultado en marzo 2008
[9] http://www.cs.ualberta.ca/ lindek/minipar.htm Consultado en marzo 2008

Universitat d'Alacant
Universidad de Alicante

- *Recuperación de documentos.* La lista de palabras clave se lanza en cinco motores de búsqueda: MSN[10], AskJeeves[11], Google[12], Altavista[13] y Gigablast[14], cuyas salidas, a diferencia de la propuesta de (Pizzato & Mollá-Aliod, 2005), son tratadas todas por igual.
- *Selección de snippets.* El sistema de BR desarrollado trabaja con unidades más pequeñas que un párrafo, en concreto, con los fragmentos de texto que devuelven los buscadores, los denominados *snippets*. De los *snippets* devueltos por los cinco buscadores, se seleccionan los 50 primeros (Pizzato & Mollá-Aliod, 2005) de cada uno de ellos, que contengan todas las palabras clave en una misma oración.
- *Extracción de respuestas.* Se utilizan dos submódulos diferentes. Uno que hace uso de un conjunto de reglas semánticas manualmente definidas, que determinan dada una pregunta el tipo de rol semántico esperado como respuesta, y otro que hace uso de un conjunto de patrones semánticos construidos automáticamente, que sirven para identificar respuestas candidatas a partir de su rol semántico. En cualquiera de los dos casos, este módulo devolverá un conjunto de posibles respuestas. Información detallada sobre cada uno de estos submódulos se presentará en el apartado 6.3.2.
- *Clasificación de respuestas.* Con el fin de establecer una lista ordenada de las respuestas obtenidas en el paso anterior se asignará un peso a cada una de las respuestas candidatas. Para ello se hace uso de dos medidas: distancia e información mutua. Además, en la lista ordenada se eliminarán duplicados acumulando sus pesos. La respuesta con mayor peso, siempre que dicho peso sea superior a un umbral fijado, será la respuesta seleccionada.

[10] http://es.msn.com/ Consultado en marzo 2008
[11] http://es.ask.com/#subject:ask—pg:1 Consultado en marzo 2008
[12] http://www.google.es/ Consultado en marzo 2008
[13] http://es.altavista.com/ Consultado en marzo 2008
[14] http://beta.gigablast.com/ Consultado en marzo 2008

Universitat d'Alacant
Universidad de Alicante

- Distancia. Se calculará la distancia de cada posible respuesta a cada una de las palabras de la lista de palabras clave.

  $$S(R) = \sum_{i=1}^{n}(\delta(R, f_i)^{-1}/n)$$

  donde, $F = \{f_1, f_2, ..., f_n\}$ es la lista de palabras clave, $\delta(a, b)$ es el número de palabras entre a y b y R la posible respuesta.

- Información mutua (IM). De manera informal, la información mutua de dos palabras, a y b, compara la probabilidad de observar a y b juntas, con la probabilidad de observar a y b independientemente. De manera formal,

  $$I(a, b) = log(P(a, b)/P(a)P(b))$$

  donde $P(a, b)$ es la probabilidad del bigrama a,b, y $P(x)$ es la estimación del unigrama x:

  $P(x) = frec(x)/num_total_de_unigramas.$

  Así, si hay relación entre a y b, $P(a, b)$ será mucho mayor que $P(a)P(b)$, y por tanto $I(a, b) > 0$.

  Esta medida plantea algunos problemas cuando hay escasez de datos, como es el caso que aquí se presenta, puesto que no se trabaja con documentos completos, si no con *snippets*. En estos casos se recomienda no tratar pares de palabras que aparezcan menos de 5 veces (Church & Hanks, 1989).

  En concreto:

  1. Calcular IM para cada uno de los unigramas, bigramas y trigramas que se obtienen de los *snippets* seleccionados como candidatos a contener la respuesta, una vez se hayan eliminado las *stopwords*[15], y siempre y cuando aparezcan más de 5 veces en los *snippets*.

  2. Eliminar los unigramas, bigramas y trigramas en los que todas las palabras pertenezcan a la lista de palabras clave.

[15] http://members.unine.ch/jacques.savoy/clef/

Universitat d'Alacant
Universidad de Alicante

3. Si la información mutua es mayor que 1[16], se buscan en WordNet hiperónimos de las palabras que forman los n-gramas. La hipótesis es que tales n-gramas tienen más probabilidad de ser la respuesta a la pregunta (Pizzato & Mollá-Aliod, 2005).

   a) Si algún hiperónimo coincide con alguna de las palabras clave, su $IM = IM/10$

   b) Si alguna palabra del n-grama coincide con alguna de las palabras clave, $IM = IM/10$

   c) En cualquier otro caso $IM = IM/20$

4. Si el n-grama incluye alguna de las respuestas candidatas, acumular el valor obtenido para la IM, al valor de la distancia.

- Agrupar las respuestas atendiendo tanto a criterios de popularidad como de inclusión, manteniendo la respuesta más popular, o la más larga, y acumulando pesos.
- Seleccionar como respuesta la palabra con más peso, siempre y cuando se supere el umbral $numero_de_snippets/40$[17].

De todos los módulos comentados, el que realmente es centro de interés para este trabajo es el módulo de extracción de la respuesta, el cual como se ha indicado ya, se compone de dos submódulos diferentes. Uno que hace uso de un conjunto de reglas semánticas manualmente definidas, que determinan dada una pregunta el tipo de rol semántico esperado como respuesta, y otro que hace uso de un conjunto de patrones semánticos construidos automáticamente, que sirven para identificar respuestas candidatas a partir de su rol semántico. Veamos a continuación en detalle cada uno de ellos.

[16] Un IM mayor que 1 significa que el n-grama ocurre más a menudo que su probabilidad de ocurrencia aleatoria (Pizzato & Mollá-Aliod, 2005)

[17] Este umbral se ha obtenido de manera experimental

6.3.2 Extracción de respuestas basada en roles semánticos

Para analizar el papel de los roles semánticos en sistemas de BR se va a estudiar el comportamiento del módulo de extracción de respuestas candidatas atendiendo, por un lado a reglas que relacionen tipos de preguntas con posibles roles semánticos respuesta, y por otro, a patrones construidos a partir de la información facilitada por los roles semánticos.

Dado que existen diferentes tipos de preguntas atendiendo al tipo de rol respuesta esperado, y que puede ocurrir que el comportamiento de los sistemas de BR no sea generalizado para cualquier tipo de rol, el trabajo que aquí se presenta, afronta dar respuesta a preguntas cuya respuesta esté representada por un rol de lugar. Esta especialización, sin embargo, no significa que no sea posible hacer el mismo análisis sobre cualquier otro tipo de pregunta. Realizarlo supondría repetir el proceso para cada tipo de pregunta diferente.

Extracción de respuestas basada en reglas. Tal y como se mostró en (Moreda *et al.*, 2007) se puede establecer una relación entre el tipo de pregunta y el rol semántico que la respuesta esperada tenga en la oración. Por ejemplo, preguntas como "*where*", "*in where*", "*in what* + expresión de lugar" o "*at what* + expresión de lugar", pueden responderse con argumentos etiquetados con el rol de lugar y nunca con roles como agente, paciente, temporal, causa o modo. El detalle de estas relaciones se muestra en el cuadro 6.4.

Universitat d'Alacant
Universidad de Alicante

En función de estas relaciones es posible definir un conjunto de reglas semánticas, que determinen dada una pregunta y su tipo, qué rol tendrá el texto buscado como respuesta. Conforme a estas reglas, este módulo seleccionará como respuestas candidatas todos aquellos argumentos, de los *snippets* devueltos por el módulo

Pregunta	Rol	No Rol
Where	Location	ProtoAgent
In where		Mode
In what + exp		Temporal
At what + exp		Cause
		ProtoPatient
When	Temporal	ProtoAgent
In what + exp		Mode
What + exp		Location
		Cause
		ProtoPatient
How	Mode	ProtoAgent
	Theme (if it is a diction verb)	Location
		Temporal
		Cause
		Patient
		Beneficiary
Who	Agent − ProtoAgent	Mode
	Patient − ProtoPatient	Temporal
		Location
		Theme
		beneficiary
What	Cause	
	Theme	
Whose	Receiver	Agent
	Beneficiary	Location
	Patient	Mode
	ProtoPatient	Temporal
		Theme
		Cause

Cuadro 6.4. Conjunto de relaciones semánticas pregunta-rol_semántico

Universitat d'Alacant
Universidad de Alicante

de selección de *snippets*, que tengan rol de lugar (etiqueta AM-LOC), para aquellas preguntas cuya respuesta esperada sea un lugar. Para ello se utilizará el clasificador de SemRol que mejores resultados obtiene para roles de lugar.

Como ejemplo, considerar la pregunta mostrada en (E70). El módulo de extracción de respuestas basado en reglas, anota las oraciones contenidas en los *snippets* seleccionados y extrae como posible respuesta aquellos argumentos anotados con el rol de lu-

gar, que en el ejemplo (E71) correspondería a "*in the island of Samos*".

(E70) Where was Pythagoras born?

(E71) Samos Pitagoras was born [$_{[}AM-LOC$] in the island of Samos]

Es importante destacar que a diferencia de otros sistemas que también hacen uso de reglas, como el caso de (Stenchikova *et al.*, 2006), sólo es necesario anotar las oraciones candidatas a contener la respuesta, y no la pregunta. De esta manera se evitan los problemas que presentan las herramientas de anotación automática de roles semánticos, a la hora de anotar oraciones interrogativas. Estos problemas se deben, principalmente, a la escasez de oraciones interrogativas en los corpus utilizados por tales herramientas de SRL.

Extracción de respuestas basada en patrones. Recordemos que el conjunto de roles utilizado en PropBank no permite hacer generalizaciones respecto al significado de tales roles. Aunque este conjunto de roles es único para todos los verbos, también ocurre que es variable en significado dependiendo del verbo y sentido con el que los roles aparezcan. Esto significa que la regla definida en el apartado anterior no siempre funcionará adecuadamente. Consideremos las dos oraciones mostradas en (E72) y (E73). En ambos casos, "to the John's house" y "to the park" son argumentos que tienen el rol de lugar. Sin embargo, en ninguno de los dos casos el argumento está etiquetado como un adjunto del tipo AM-LOC. Es más, en un caso el lugar está representado por un rol de tipo A2 y en otro de tipo A4.

(E72) [$_{A0}$ Mary] is going [$_{A2}$ to the John's house]

(E73) [$_{A0}$ Mary] is going [$_{A4}$ to the park]

PropBank	Moreda et al.
Arg0	Proto-agent
Arg1	Proto-patient: T-P
Arg2	Proto-patient: B-R
	Proto-agent: Instrument
	Location
Arg3	Proto-patient
	Proto-agent: Instrument
	Location
Arg4	Location: Goal
AM-LOC	Location
AM-MNR	Mode
AM-TMP	Temporal

Cuadro 6.5. Correspondencia entre PropBank y la propuesta de Moreda et al.

Como se indicaba en (Moreda *et al.*, 2007), y tal y como el cuadro 6.5 muestra, en PropBank el lugar puede ser representado por roles A2, A3, o A4, o por el adjunto AM-LOC. Sin embargo, el módulo basado en reglas semánticas sólo extraerá respuestas candidatas cuando se anoten adjuntos del tipo AM-LOC. Una primera mejora podría ser ampliar la regla de manera que si no se ha anotado ningún argumento con el rol de tipo AM-LOC, se busque cualquiera de los otros posibles roles. Esto es posible puesto que cuando los roles A2, A3 o A4 representan lugar, ningún otro argumento puede tener el rol de lugar. Esta solución plantearía dos problemas: i) cómo saber cuando el rol A2 está representando lugar y cuando no, ii) cómo saber qué rol es el que representa el lugar, caso de aparecer juntos en la misma oración A2, A3 y/o A4. Por ejemplo, en la oración (E74) uno de los argumentos tiene el rol A2, pero no representa lugar; o la oración (E75) en la que aparecen argumentos con roles A3 y A4 y no se puede saber cuál es el que está representando el lugar.

(E74) [$_{A0}$ Mary] is talking [$_{A2}$ with John] [$_{A1}$ about the party]

(E75) [$_{A0}$ Mary] is going [$_{A3}$ with John] [$_{A4}$ to the park]

A fin de salvar este problema, y partiendo del trabajo presentado en (Yousefi & Kosseim, 2006) para entidades nombradas, se propone extraer de forma automática un conjunto de patrones semánticos construidos a partir de la información que los roles semánticos proporcionan. Estos patrones contemplarán todos los posibles casos en los que los roles están expresando lugar. De esta manera los problemas de cobertura detectados en el módulo de reglas semánticas quedarán solucionados.

La construcción automática de patrones basados en roles semánticos proporciona dos beneficios importantes. Por una parte, un argumento con un rol concreto en una oración siempre tiene el mismo rol independientemente del tiempo, voz o variación sintáctica de la oración. Por ejemplo, ante la pregunta *Who is the president of U.S.?*, un sistema que haga uso de patrones debería encontrar que las oraciones (E76), (E77) y (E78), son todas ellas oraciones candidatas a contener la respuesta. Y además, determinar que *Bill Clinton* es la respuesta a extraer. Por otra parte, con un proceso automático, es posible obtener muchos más patrones que con un proceso manual, y además, con menor coste. En consecuencia, dos de los principales problemas achacables al uso de patrones quedarían solucionados: su cobertura y el coste de su construcción.

(E76) The president of the U.S. is [$_{A}$2 Bill Clinton]

(E77) [$_{A0}$ Bill Clinton, the president of U.S.] gave a conference yesterday

(E78) [$_{A0}$ Bill Clinton, also known as the leader of the US], gave a conference yesterday

Veamos a continuación cuál es este proceso de construcción de patrones semánticos y cómo, una vez construidos, tales patrones pueden ayudar a la búsqueda de respuestas.

Universitat d'Alacant
Universidad de Alicante

Extracción de patrones: Este proceso tiene como objetivo generalizar *snippets* extraídos de la Web en patrones semánticos, utilizando para ello un conjunto de pares pregunta−respuesta. Dicho proceso se resume en cuatro pasos:

1. Recuperación de oraciones. Para cada par pregunta-respuesta se obtiene el conjunto de términos que cualquier documento relevante contendría. Estos términos son utilizados para realizar una consulta en la Web que devuelva *snippets* conteniendo los términos seleccionados.

   a) El conjunto de términos relevantes está formado por todas las posibles combinaciones de los sintagmas nominales de la pregunta, una vez eliminado el foco de la misma, con todos los subsintagamas de la respuesta.

      1) Para extraer los sintagmas nominales se hará uso de la herramienta de análisis sintáctico parcial desarrollada por The cognitive Computation Group[18].

      2) Para determinar el foco de la pregunta se utilizará el árbol de análisis devuelto por Minipar[19], seleccionando el nodo raíz de dicho árbol.

   b) Las cadenas formadas por todas las posibles combinaciones de los sintagmas nominales de la pregunta con todos los subsintagmas de la respuesta, se lanzan en varios motores de búsqueda disponibles en la red: MSN[20], AskJeeves[21], Google[22], Altavista[23] y Gigablast[24]

   c) Seleccionar los 100 primeros *snippets* devueltos por cada buscador, que contengan tanto los sintagmas de la pregunta, como al menos uno de la respuesta, en una misma oración.

[18] http://l2r.cs.uiuc.edu/ cogcomp/demo.php?dkey=SP Consultado en marzo 2008
[19] http://www.cs.ualberta.ca/ lindek/minipar.htm Consultado en marzo 2008
[20] http://es.msn.com/ Consultado en marzo 2008
[21] http://es.ask.com/#subject:ask—pg:1 Consultado en marzo 2008
[22] http://www.google.es/ Consultado en marzo 2008
[23] http://es.altavista.com/ Consultado en marzo 2008
[24] http://beta.gigablast.com/ Consultado en marzo 2008

Universitat d'Alacant
Universidad de Alicante

2. Filtrado semántico de *snippets*. Los *snippets* seleccionados son validados atendiendo a las relaciones semánticas que contegan. Puesto que la relación semántica generalmente aparece respecto al verbo principal de una pregunta, todos los verbos de los *snippets* seleccionados son examinados. Aquellas oraciones de los *snippets* que contengan sinónimos, hiperónimos o hipónimos del verbo de la pregunta, se seleccionan.

   a) Para ello se hace uso de WordNet y de la herramienta para desambiguador de sentidos del Grupo de Procesamiento del Lenguaje Natural de la Universidad de Alicante (Montoyo *et al.*, 2005).

3. Generación de patrones. Las oraciones seleccionadas son generalizadas en patrones semánticos utilizando información a cerca de roles semánticos.

   a) Cada oración se anota con roles semánticos, utilizando los clasificadores de SemRol específicos para cada tipo de rol. En primer lugar se anotan los adjuntos de tipo AM-LOC, y si no hubiera, entonces se anotan los argumentos con rol A2, A3 y A4.

   b) Reemplazar cada argumento en la respuesta candidata que coincida o contenga un sintagma o subsintagma nominal de la respuesta correcta, con el rol asignado en el paso anterior. Para los casos, en los que un mismo *snippet* contenga más de un rol numerado que cumpla la condición de inclusión, se generará un patrón para cada tipo de rol.

   c) Reemplazar cada argumento en la respuesta candidata que coincida o incluya un sintagma nominal de la pregunta por una etiqueta $< QARG_n >$, siendo n un contador de sintagmas. Si un argumento contiene más de un sintagma, se indicarían los diferentes sintagmas como sustitución de un mismo argumento.

   d) Reemplazar resto de argumentos por etiquetas $< ARG_n >$, siendo n un contador.

   e) Eliminar el resto de información a excepción de las preposiciones.

   f) Sustituir el verbo de la oración por la lista de verbos obtenida.

4. Eliminar patrones duplicados. Formar una única lista de verbos que incluya a los verbos del patrón eliminado.

Un ejemplo de este proceso de creación de patrones se muestra en el apartado 6.5.

Todos los patrones generados mediante este proceso automático se almacenan para su posterior uso a la hora de extraer respuestas candidatas.

Extracción respuesta: Ante una nueva pregunta para la que se desconoce su respuesta, será necesario comprobar si alguno de los *snippets* candidatos obtenidos con el sistema de BR se ajusta a alguno de los patrones semánticos generados en el proceso anterior. En caso de coincidencia, la respuesta candidata se obtendrá del texto correspondiente a la etiqueta de rol dentro del patrón semántico. Para ello será necesario anotar las oraciones candidatas a contener la respuesta con roles semánticos, de manera que se puedan generalizar en patrones semánticos.

1. Anotar los snippets seleccionados con información sobre roles semánticos (AM-LOC, o si no hubiera, A2, A3, A4) utilizando los clasificadores de SemRol correspondientes.
2. Generar el patrón para la respuesta candidata:

   a) Seleccionar las oraciones en las que el verbo de la oración pertenezca a la clase del verbo de alguno de los patrones.

   b) Obtener la lista numerada de sintagmas nominales de la pregunta, que no sean foco de la misma.

   c) Reemplazar cada argumento en el *snippet* que coincida o contenga un sintagma nominal de la pregunta por una etiqueta $< QARGn >$, siendo n un contador. Si un argumen-

Universitat d'Alacant
Universidad de Alicante

to contiene más de un sintagma, se indicarán los diferentes sintagmas como sustitución de un mismo argumento.

d) Reemplazar en el *snippet* el argumento con el rol AM-LOC, o bien, si éste no existiera, con el rol A2, A3 ó A4, por su etiqueta de rol correspondiente. En caso de que una misma oración contenga más de un rol numerado generar un patrón para cada uno de ellos.

e) Reemplazar resto de argumentos por etiquetas $< ARGn >$, siendo n un contador.

f) Eliminar el resto de información a excepción de las preposiciones.

3. Si los patrones obtenidos coinciden con alguno de los patrones extraídos en el paso 1, seleccionar como respuesta candidata, el texto del *snippet* correspondiente a la etiqueta de rol del patrón.

Al igual que en el módulo de extracción de repuestas basado en reglas, no es necesario realizar en ningún momento la anotación de la pregunta con roles semánticos, evitando así los problemas de cobertura que ello conllevaría (Narayanan *et al.*, 2002; Ofoghi *et al.*, 2006; Kaisser, 2007; Moschitti *et al.*, 2007).

6.4 Análisis de la utilidad de los roles semánticos en sistemas de BR

Uno de los objetivos de este trabajo es medir la influencia de los roles semánticos en sistemas de BR. Por ello, se han llevado a cabo varios experimentos cuyos resultados se analizarán desde tres perspectivas diferentes. Por un lado, se realizará un estudio comparativo de los dos submódulos de extracción de respuestas basados en roles semánticos (apartado 6.4.1). Por otro lado, se estudiarán los casos en los que los sistemas de BR basados en roles semánticos son preferibles respecto a los sistemas clásicos

basados en NE (apartado 6.4.2). Finalmente, se llevará a cabo una comparación entre el sistema desarrollado en este trabajo y el resto de propuestas que siguen una línea común (apartado 6.4.3). Todo ello centrado en preguntas cuya respuesta es un lugar.

6.4.1 Extracción de respuesta basada en Reglas frente a Patrones

Para evaluar los dos módulos de extracción de respuesta basados en roles semánticos se ha utilizado un conjunto de 100 preguntas de lugar, extraídas de los conjuntos de preguntas del TREC8 y TREC9. Además, el módulo basado en patrones semánticos ha requerido de un conjunto adicional de 200 preguntas de lugar que permitiera la construcción automática de un conjunto adecuado de patrones. Estas preguntas se han extraído de los conjuntos de preguntas del TREC2003 y TREC2006 y de OpenTrivia[25].

El cuadro 6.6 resume los resultados obtenidos por el sistema de BR cuando se utiliza para la extracción de respuestas candidatas, tanto el submódulo basado en reglas semánticas, como el basado en patrones semánticos. Este cuadro muestra el incremento, expresado en tanto por ciento, que se obtiene cuando se utilizan patrones, respecto al uso de reglas. La evaluación se ha realizado en cuanto a precisión, cobertura, medida $F_{\beta=1}$ y MRR.

	P (%)	C (%)	$F_{\beta=1}$ (%)	MRR (%)
Reglas	65,60	21,00	31,80	52,25
Patrones	**88,20**	**30,00**	**44,88**	**58,33**
Incremento (%)	+33,40	+42,80	+40,80	+13,00

Cuadro 6.6. Resultados para un sistema de BR basado en roles semánticos

Como se puede ver en el cuadro 6.6 el submódulo de patrones semánticos mejora a reglas, tanto en precisión como en cobertu-

[25] http://www.opentrivia.com/

ra. La mejora respecto a cobertura se debe a dos razones: i) la inclusión de los argumentos de lugar con roles A2, A3 y A4, ii) la inclusión de los sinónimos, hiperónimos e hipónimos del verbo de la pregunta. Por otro lado, la mejora respecto a precisión se debe a que el módulo basado en patrones sólo selecciona aquellos argumentos con rol de lugar incluidos en patrones que coincidan con alguno de los patrones previamente generados con los pares pregunta-respuesta conocidos. Sin embargo, el módulo basado en reglas extrae como posible respuesta cualquier argumento con rol de lugar. Es decir, los patrones establecen un mayor filtro de las sentencias seleccionadas, lo que se traduce en una mayor precisión en la extracción de la respuesta.

Es importante destacar que la cobertura obtenida, tanto con reglas como con patrones, acusa el hecho de que los procesos se han realizado de forma completamente automática, sin ningún ajuste ni revisión manual. Incluso a pesar de la sabida degradación que sufren los sistemas de SRL basado en aprendizaje automático supervisado al cambiar de corpus de trabajo, respecto al corpus utilizado para el entrenamiento de la herramienta (Carreras & Màrquez, 2005; Surdeanu *et al.*, 2007; Pradhan *et al.*, 2008).

En conclusión, el módulo de extracción de respuestas basado en patrones semánticos obtiene mejores resultados que el basado en reglas. La mejora es lo suficientemente significativa, la medida $F_{\beta=1}$ mejora en un 40 %, como para justificar el coste del proceso de construcción de los patrones.

Universitat d'Alacant
Universidad de Alicante

6.4.2 Comparación con sistemas de BR basados en NE

Un aspecto importante a tener en cuenta y que afecta a la cobertura obtenida por los módulos basados en roles, es el hecho de que la mayoría de las preguntas utilizadas en la evaluación esperan como respuesta una entidad nombrada. Como ya se ha comentado, y como parece obvio, los sistema de BR basados en roles semánticos difícilmente podrán mejorar a los basados en NE en este tipo de preguntas.

Sin embargo, qué pasaría si se consideraran preguntas como la mostrada en el ejemplo (E79). La respuesta a este tipo de preguntas ya no es una entidad nombrada, por lo que un sistema basado en entidades, nunca sería capaz de encontrar la respuesta. Sin embargo, a un sistema basado en roles le bastaría con encontrar el argumento de la oración que represente el rol de lugar (ejemplo E80).

(E79) Where is pancreas located? Abdomen.

(E80) [$_{A1}$ The pancreas] is located [$_{AM-LOC}$ deep in the abdomen]

Por ello, en este apartado se analizará el comportamiento de ambos tipos de sistemas de BR sobre dos conjuntos de respuestas bien diferenciados. Un conjunto de 50 preguntas de lugar, extraídas de los conjuntos de preguntas del TREC8 y TREC9, y para las que la respuesta esperada es una entidad nombrada; y un conjunto de 50 preguntas desarrolladas específicamente para este trabajo, y para las que la respuesta no era una entidad nombrada.

De nuevo la evaluación se ha realizado respecto a precisión, cobertura, medida $F_{\beta=1}$ y MRR.

El cuadro 6.7 muestra los resultados de este proceso de evaluación. Estos resultados confirman, claramente, que mientras que el módulo basado en entidades funciona mejor para preguntas cuya respuesta es una entidad (MRR $+66,98\,\%$ sobre reglas y $+49,57\,\%$ sobre patrones), las aproximaciones basadas en roles superan a las NE en el caso contrario (MRR $+142,25\,\%$ para reglas y $+223,48\,\%$ para patrones).

Destaca en estos resultados el hecho de que el módulo de extracción de respuestas basado en entidades haya sido capaz de contestar preguntas cuya respuesta esperada no es una entidad (MRR 12,50 %). Un análisis de dichas preguntas descubre que el acierto, en realidad, se debe a un error de anotación de la herramienta utilizada para reconocer y clasificar las entidades nombra-

das. Por ejemplo, para la pregunta (E81), el módulo de extracción de respuestas basado en entidades encuentra la respuesta correcta dado que LingPipe siempre anota como entidad de lugar la palabra "*nectar*" si aparece junto a la palabra "*flower*".

(E81) Where de bees produce honey from? Nectar

Aproximación	Medida(%)	NE	No NE
NE	Precisión	87,50	15,62
	Cobertura	84,00	10,00
	$F_{\beta=1}$	85,70	12,19
	MRR	**87,25**	12,52
Reglas	Precisión	91,54	75,00
	Cobertura	52,00	30,00
	$F_{\beta=1}$	66,32	42,85
	MRR	52,25	30,33
Patrones	Precisión	93,54	95,23
	Cobertura	58,00	40,00
	$F_{\beta=1}$	71,60	56,33
	MRR	58,33	**40,50**

Cuadro 6.7. Resultados para sistemas de BR basados en roles semánticos y en entidades para respuestas NE y no NE

Universitat d'Alacant
Universidad de Alicante

Respecto al comportamiento de los diferentes módulos en precisión y cobertura, cabe resaltar la alta precisión de los módulos basados en roles semánticos, y la baja cobertura para cualquiera de las aproximaciones cuando las respuestas no son entidades nombradas. La alta precisión (91,54 % para reglas y 93,54 % para patrones, frente a 87,50 % para NE) se debe al hecho de que las aproximaciones basadas en roles únicamente extraen como respuestas candidatas argumentos cuyos roles representan lugar. Sin embargo, para el caso del módulo basado en entidades, cualquier entidad del tipo lugar es seleccionada, independientemente del papel que esa entidad juegue en la oración. Por ejemplo, la oración (E82) nunca sería seleccionada como respuesta por el módulo de

extracción basado en roles, pero si por el que hace uso de las entidades (E83), a pesar de no ser una respuesta a una pregunta de lugar.

(E82) [$_{A0}$ Spain] has a population of 45 million.

(E83) [$_{LOC}$ Spain] has a population of 45 million.

La baja cobertura para preguntas cuya respuesta no es una entidad, se debe a la menor cantidad de información para preguntas de este tipo disponible en Internet.

Con todo ello, se puede concluir que, en general, respecto al módulo de extracción de respuestas basado en entidades nombradas, el comportamiento de los módulos basados en roles semánticos: i) es bastante más preciso, ii) muestra mejor comportamiento ante preguntas cuya respuesta no es una entidad, iii) presenta una menor cobertura sólo si la respuesta esperada es una entidad.

6.4.3 Comparación con otros sistemas de BR basados en roles

En realidad, los resultados obtenidos por el sistema desarrollado en este trabajo no son directamente comparables con ninguno de los sistemas presentados en el apartado 6.2. El sistema presentado aquí se diferencia del resto, principalmente por: i) haber restringido el análisis a preguntas de lugar; iii) haber utilizado en la evaluación un subconjunto de preguntas del TREC8 y TREC9.

En cualquier caso, y con la intención de dar una idea aproximada de los resultados obtenidos hasta la fecha, el cuadro 6.8 muestra los resultados de los módulos de extracción de respuesta basados en roles semánticos tanto de nuestro sistema, como del resto de aproximaciones que han hecho uso de roles semánticos en dicho módulo. La comparación se realiza respecto a precisión o medida MRR, dependiendo de la información facilitada de cada sistema.

Sistema	P(%)	MRR(%)	Preguntas
Patrones	88,20	58,33	Lugar: TREC8 TREC9
Reglas	65,60	52,25	Lugar: TREC8 TREC9
(Stenchikova *et al.*, 2006)		30,00	General: TREC9
(Ofoghi *et al.*, 2006)		38,89	General: TREC2004
(Kaisser, 2007)	36,70		General: TREC2002
(Fliedner, 2007)	66,00		Textos de noticias

Cuadro 6.8. Comparación de diferentes sistemas de BR basados en roles semánticos

Si bien es probable que los resultados del sistema de BR aquí presentado, sufran algún decremento al generalizar los submódulos basados en roles semánticos a otros tipos de preguntas, también es destacable que los resultados obtenidos para preguntas de lugar resultan bastante prometedores.

6.5 Ejemplo de construcción de patrones semánticos

A continuación se muestra en detalle un ejemplo del proceso de construcción de patrones semánticos. En concreto, el proceso muestra para la pregunta y respuesta del cuadro 6.9, qué patrones son obtenidos como resultado de dicho proceso. Los pasos seguidos son:

Universitat d'Alacant
Universidad de Alicante

Pregunta	*Where is the actress, Marion Davies, buried?*
Respuesta	*Hollywood Memorial Park*
Patrones	[[QARG$_1$] [QARG$_2$]] [bury] [AM-LOC] [ARG$_1$].
	[[QARG$_1$] [QARG$_2$]] [bury, inter] [AM-LOC].

Cuadro 6.9. Ejemplos de patrones generados para la pregunta *Where is the actress, Marion Davies, buried?*

1. Recuperación de oraciones

   a) Lista numerada de sintagmas nominales de la pregunta una vez eliminado el foco:

   {$_{QARG_1}$ "the actress", $_{QARG_2}$ "Marion Davies"}

   b) Conjunto de subsintagmas de la respuesta:

   {"Hollywood Memorial Park", "Hollywood Memorial", "Hollywood Park", "Memorial Park", "Hollywood", "Memorial", "Park"}

   c) Cadenas de búsqueda:

   {"the actress", "Marion Davies", "Hollywood Memorial Park" }

   {"the actress", "Marion Davies", "Hollywood Memorial" }

   {"the actress", "Marion Davies", "Hollywood Park" }

   {"the actress", "Marion Davies", "Memorial Park" }

   {"the actress", "Marion Davies", "Hollywood" }

   {"the actress", "Marion Davies", "Memorial" }

   {"the actress", "Marion Davies", "Park" }

   d) Búsqueda en la Web. Ejemplo de tres *snippets* devueltos por los buscadores:

   The actress Marion Davies is buried in Hollywood in 1961.

   The actress Marion Davies is buried in the Hollywood Forever Memorial Park Cemetery in Hollywood.

   The actress Marion Davies was much loved by her friends and by Hollywood in general.

2. Filtrado semántico de oraciones.

   a) Examinar el verbo principal de la oración. De las tres oraciones anteriores, sólo las dos primeras cumplen que el verbo de la oración coincida o sea sinónimo, hiperónimo o hipónimo del verbo de la pregunta. En el caso de la tercera oración, el verbo *to love* no está relacionado con el verbo *to bury*. Por tanto, esa oración es eliminada.

   The actress Marion Davies is buried in Hollywood in 1961.

   The actress Marion Davies is interred in the Hollywood Forever Memorial Park Cemetery in Hollywood.

   b) Formar la lista de verbos

   The actress Marion Davies [bury] in Hollywood in 1961.

Universitat d'Alacant
Universidad de Alicante

The actress Marion Davies [bury, inter] in the Hollywood Forever Memorial Park Cemetery in Hollywood.

3. Generación de patrones:

   a) Anotar las oraciones con los roles de lugar (AM-LOC, y si no hubiera, A2, A3 y A4):

   [The actress Marion Davies] [bury] [$_{AM-LOC}$ in Hollywood] in 1961.

   [The actress Marion Davies] [bury, inter] [$_{AM-LOC}$ in the Hollywood Forever Memorial Park Cemetery in Hollywood].

   b) Reemplazar los argumentos conteniendo cualquiera de los subsintagmas de la respuesta por su etiqueta de rol:

   [The actress Marion Davies] [bury] [AM-LOC] in 1961.

   [The actress Marion Davies] [bury, inter] [AM-LOC].

   c) Reemplazar los argumentos conteniendo sintagmas nominales de la pregunta por su correspondiente etiqueta numerada:

   [[QARG$_1$] [QARG$_2$]] [bury] [AM-LOC] *in 1961*

   [QARG$_1$] [QARG$_2$]] [bury, inter] [AM-LOC]

   d) Reemplazar el resto de argumentos por etiquetas numeradas:

   [[QARG$_1$] [QARG$_2$]] [bury] [AM-LOC] [ARG$_1$]

   [[QARG$_1$] [QARG$_2$]] [bury, inter] [AM-LOC]

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

7. Conclusiones y trabajos futuros

Para terminar, se presentará un resumen de las principales conclusiones de este trabajo (apartado 7.1), así como un detalle de las aportaciones más importantes al conocimiento de la investigación en roles semánticos (apartado 7.2), y una lista analizada de las publicaciones más relevantes relacionadas con el trabajo (apartado 7.3). Finalmente, se comentarán los principales trabajos, tanto en curso como futuros (apartado 7.4).

7.1 Conclusiones

La principal conclusión que se puede extraer de la investigación desarrollada en esta Tesis doctoral, es la importancia de los roles semánticos en aplicaciones de la tecnología del lenguaje humano, y más concretamente en la búsqueda de respuestas. Dicha importancia se traduce en la necesidad de una herramienta de anotación de roles semánticos eficiente y eficaz. Es decir, una herramienta con buena precisión en sus resultados y con un coste computacional razonable para su uso en aplicaciones reales. Para conseguir tal herramienta hace falta, por un lado, un conjunto de roles semánticos adecuado, y si fuera posible consensuado por todos e independiente de la lengua; y por otro, un conjunto de recursos lingüísticos anotados que permitan desarrollar automáticamente dicha herramienta.

Por ello, el trabajo que aquí se ha presentado parte de tres objetivos principales. Por un lado, investigar en los conjuntos de roles semánticos y recursos lingüísticos definidos sobre ellos que

hay disponibles hasta la fecha; así como en proyectos cuyo objetivo haya sido desarrollar tales recursos. Por otro, abordar el desarrollo de una herramienta para la anotación automática de roles semánticos, SemRol, que permita realizar un análisis de las necesidades de dichas herramientas. Y finalmente, aportar información y conclusiones sobre la influencia de los roles semánticos en sistemas de búsqueda de respuestas.

En cuanto a la investigación en conjuntos de roles semánticos y recursos lingüísticos existentes, se puede concluir, que:

- Existe una gran variedad de conjuntos de roles semánticos definidos por los investigadores, generalmente independientes de la lengua, y tanto de uso general como específicos para aplicaciones determinadas.
- La variedad de conjuntos de roles se traduce en una gran variedad de recursos lingüísticos definidos sobre ellos, dependientes de la lengua en su mayoría, y, casi siempre, con una clara orientación hacia el aprendizaje automático.
- Se han llevado a cabo varios proyectos relacionados con la anotación automática de roles semánticos, con el fin de generar los recursos lingüísticos necesarios.
- La falta de consenso entre los investigadores a la hora de definir un conjunto de roles semánticos estándar, ha generado la necesidad de: i) establecer relaciones entre los recursos lingüísticos creados, con el objetivo de conseguir independencia respecto al recurso utilizado; ii) empezar por especificar qué roles se van a utilizar, al desarrollar cualquier recurso o herramienta que quiera hacer uso de la información proporcionada por los roles semánticos.
- Actualmente, existe un fuerte apoyo gubernamental en cuanto a proyectos relacionados con la semántica, más concretamente con los roles semánticos, y en general con la tecnología del lenguaje humano, como demuestra el plan nacional español, o el sexto y séptimo programa marco.

Universitat d'Alacant
Universidad de Alicante

Respecto a la investigación y desarrollo de la herramienta de anotación, la cual hace uso de estrategias de aprendizaje automático supervisado, cabe resaltar que dicha herramienta posee un fuerte componente de análisis y ajuste de la información utilizada en la anotación. Dicho componente de análisis da lugar a que el proceso de anotación de roles semánticos se realice desde dos perspectivas diferentes: i) clasificación por sentidos frente única, ii) clasificación global frente individual.

Los resultados obtenidos de este proceso de investigación y desarrollo de la herramienta de anotación establecen:

- La obligatoriedad de un proceso de ajuste de la información en herramientas que hacen uso de aprendizaje automático, con el fin de eliminar aquellas características que interfieran con otras más útiles. Esta obligatoriedad es independiente de la estrategia de anotación utilizada o del algoritmo de aprendizaje elegido
- El algoritmo de aprendizaje elegido para el proceso de ajuste puede afectar a los resultados obtenidos y a los tiempos de ejecución necesitados para la obtención de tales resultados.
- La estrategia de anotación por sentidos es preferible cuando se quieren anotar argumentos numerados, mientras que la estrategia única presenta mejores resultados para la anotación de adjuntos.
- La información sintáctica total no aporta beneficios notables a la tarea de identificación de los roles semánticos jugados por los argumentos de un verbo, cuando dichos argumentos son conocidos.
- El diseño de clasificadores individuales que hagan uso de información útil para cada tipo de rol semántico diferente, repercute en una mejora significativa de los resultados para cada uno de esos clasificadores, en comparación con el clasificador obtenido como resultado de un proceso de ajuste global para todos los tipos de roles semánticos como un todo.

Universitat d'Alacant
Universidad de Alicante

- Se ha demostrado, con el uso de dos algoritmos de aprendizaje automático supervisado diferentes, el elevado coste computacional de máxima entropía frente a TiMBL.

Respecto a la aportación de información y conclusiones sobre la influencia de los roles semánticos en sistemas de búsqueda de respuestas, se ha desarrollado un sistema con tres posibles módulos de extracción de respuesta: uno basado en entidades, otro basado en reglas semánticas que relacionan tipos de preguntas con posibles roles semánticos respuesta, y otro basado en patrones semánticos que hacen uso de la información que las reglas semánticas proporcionan.

La evaluación realizada sobre cada uno de estos módulos establece:

- El módulo de extracción de respuestas basado en patrones semánticos obtiene mejores resultados que el basado en reglas. La mejora es lo suficientemente significativa, la medida $F_{\beta=1}$ mejora en un 40 %, como para justificar el coste del proceso de construcción de los patrones.
- Comparado con el módulo basado en entidades, los módulos basados en roles semánticos son más precisos, muestran mejor comportamiento ante preguntas cuya respuesta no es una entidad, y presentan una menor cobertura, sólo si la respuesta esperada a la pregunta es una entidad nombrada.

Universitat d'Alacant
Universidad de Alicante

7.2 Aportaciones al conocimiento de la investigación en roles semánticos

Las principales aportaciones de este trabajo al conocimiento de la investigación en roles semánticos se pueden resumir en:

- Amplia recopilación de la gran diversidad de propuestas de conjuntos de roles semánticos realizadas hasta la fecha, confirmando

con ello la falta de consenso entre los investigadores a la hora de definir un conjunto de roles semánticos estándar, aceptado por todos y adecuado para cualquier aplicación.

- Propuesta de un conjunto de roles semánticos propio, desarrollado atendiendo a principios de aplicabilidad, generalidad, jerarquía y conexión con otras propuestas de anotación. Aplicabilidad, puesto que no se pretende definir unos roles semánticos universales, sino establecer un conjunto de roles semánticos que tenga una aplicación clara a búsqueda de respuestas. Generalidad, ya que son roles generales, aplicables a diferentes verbos que compartan rasgos semánticos similares, es decir, a toda una clase verbal. Jerarquía, puesto que es posible establecer una jerarquía entre roles semánticos, haciendo al conjunto más consistente. Y conexión, dado que la lista de roles propuesta está basada en los roles generales de PropBank y VerbNet y tiene en cuenta los utilizados en FrameNet.

- Evaluación y estudio exhaustivos de los diferentes recursos lingüísticos, tales como corpus anotados o léxicos, que la gran diversidad de conjuntos de roles semánticos ha generado. En concreto para cada recurso, se detalla el tipo de recurso que es; las lenguas para las que está disponible; el tipo de construcción, si ha sido manual o semiautomática; el origen de los datos utilizados para su construcción; las extensiones a otras lenguas y dominios, si las tiene; y el nivel al que los roles se definen, como pueden ser *frames*, verbos, o general. Este análisis hace especial hincapié en los trabajos realizados dentro de los proyectos PropBank y FrameNet, en los que los recursos desarrollados destacan por su completitud y usabilidad.

- Estudio de las relaciones que se han definido entre los recursos lingüísticos existentes, con el objetivo de conseguir independencia respecto al recurso concreto utilizado.

- Análisis de los principales enfoques seguidos por las herramientas de anotación de roles semánticos. En concreto, se analizan en profundidad las principales características de los sistemas que hacen uso de corpus anotados, los denominados sistemas

basados en corpus, así como los principales algoritmos desarrollados, ya sean supervisados, semi-supervisados o no supervisados. En este análisis se presta especial atención a los procesos de selección de características, tan importantes en esta clase de sistemas, y a los métodos que implementan dichos procesos de selección. Además, se analizan, aunque de manera algo más breve, los sistemas que hacen uso de conocimiento lingüístico previamente adquirido, los sistemas basados en conocimiento

- Investigación en cuanto a sistemas de anotación automática de roles semánticos se refiere. Dicha investigación se realiza atendiendo al corpus utilizado por la herramienta, y por tanto, la lengua para la que han sido definidas, así como el conjunto de roles utilizado en la anotación; a la información requerida para llevar a cabo el proceso de anotación; y a la estrategia o enfoque seguido en dicha anotación.
- Desarrollo de una herramienta propia para la anotación automática de roles semánticos, denominada SemRol, caracterizada por poseer un fuerte componente de análisis y ajuste de la información utilizada. Dicho componente de análisis da lugar a que el proceso de anotación de roles se realice desde dos perspectivas diferentes: i) clasificación por sentidos frente única, ii) clasificación global frente individual.
- Experimentación y prueba de un proceso de ajuste de la información requerida por SemRol, con el fin de determinar uno de los mejores conjuntos de características a utilizar en el proceso de anotación de roles semánticos.
- Evaluación de diferentes estrategias de anotación de manera que la tarea de anotación se pueda afrontar en función de las necesidades: por sentidos para roles numerados, única para adjuntos, y con clasificadores específicos para cada tipo rol.
- Desarrollo de un sistema de búsqueda de respuestas modular, basado en Web, que permite extraer posibles respuestas atendiendo a diferentes criterios: i) entidades nombradas, ii) reglas semánticas, iii) patrones semánticos.

Universitat d'Alacant
Universidad de Alicante

- Definición de un conjunto de reglas semánticas que permiten: i) establecer relaciones entre tipos de preguntas y tipos de roles semánticos que han de jugar los argumentos a contener respuestas candidatas; ii) crear de manera automática un conjunto de patrones semánticos que permiten obtener listas adecuadas de respuestas candidatas.
- Evaluación de la aportación de los roles semánticos a los sistemas de búsqueda de respuesta, desde dos perspectivas diferentes. Por un lado, determinando de qué manera los roles semánticos prestan mejor servicio a estos sistemas; y por otro, determinando bajo qué circunstancias el aporte de los roles semánticos a los sistemas de búsqueda de respuestas es preferible frente al proporcionado por las entidades nombradas.

7.3 Lista de publicaciones relevantes

A continuación se presenta una lista detallada de las publicaciones más relevantes que mantienen relación con el trabajo expuesto:

- (Moreda *et al.*, 2007).

  Este trabajo propone analizar la influencia de diferentes algoritmos de aprendizaje automático supervisado en la tarea de anotación de roles semánticos. Los algoritmos de aprendizaje utilizados son: Máxima entropía y una herramienta que hace uso de un algoritmo basado en memoria, TiMBL. El uso de uno y otro algoritmo de aprendizaje influye en el proceso de ajuste de características, de manera que si bien al utilizar aprendizaje basado en memoria, el clasificador necesita doce características para obtener una de las combinaciones de características que mejores resultados aporta a la tarea; máxima entropía, no necesita más de cuatro para obtener unos resultados bastante próximos.

  Por otro lado, se propone utilizar la información sobre roles semánticos obtenida por la herramienta de anotación como una

extensión a un sistema de recuperación de información. La idea es que sólo oraciones conteniendo roles semánticos adecuados sean seleccionadas, de manera que se reduzca la cantidad de pasajes devueltos por el sistema como candidatos a contener la respuesta. Para ello, el sistema hace uso de un conjunto de heurísticas que establecen relaciones entre preguntas y roles semánticos. Además, se analiza cómo la tarea se vería reforzada con la utilización del conjunto de roles semánticos específicos para ella.

- (Moreda & Palomar, 2006).

  En este artículo se lleva a cabo una evaluación exhaustiva del comportamiento de las características más relevantes utilizadas en sistemas de anotación de roles semánticos. En concreto, la anotación de roles se realiza desde dos perspectivas diferentes: i) clasificación para cada sentido de cada verbo, ii) clasificación única para todos los verbos por igual. Los principales resultados ponen de manifiesto que la anotación global obtiene los mejores resultados para la herramienta de anotación. Sin embargo, un análisis más detallado muestra que la aproximación por sentidos se comporta mejor en la anotación de roles específicos del verbo, como A2 o A3,y la anotación global se comporta mejor en la anotación de adjuntos.

- (Moreda & Palomar, 2005).

  Este artículo propone una metodología para seleccionar uno de los mejores conjuntos de características a utilizar en el proceso de anotación de roles semánticos. A partir del conjunto vacío de características, dicho proceso consiste en ir gradualmente añadiendo características, una cada vez, y calculando la precisión, cobertura y medida $F_{\beta=1}$ para todos las posibles combinaciones de características que se puedan hacer en cada paso. El proceso termina, cuando o bien la medida $F_{\beta=1}$ no mejora, o bien se haya alcanzado el conjunto completo de características. Aún a pesar de que este tipo de métodos no garantizan la obtención del conjunto óptimo de características, sino, sólo uno de los mejores, los propios resultados demuestran cómo atributos

Universitat d'Alacant
Universidad de Alicante

adicionales pueden interferir con otros más útiles, destacando la importancia de dicho proceso de ajuste de características. Los mejores resultados se obtienen para un conjunto de doce características, y se muestra como conjuntos de más y menos características no mejoran esos resultados.

- (Moreda *et al.*, 2004a).

  Las limitaciones de los sistemas de recuperación de información basados enteramente en sintaxis, plantea la posibilidad de utilizar técnicas de procesamiento de lenguaje natural que permitan a estos sistemas superar sus limitaciones. Entre las posibles propuestas, este trabajo presenta cómo incorporar la información proporcionada por SemRol, una herramienta de anotación de roles semánticos, en sistemas de recuperación de información. El objetivo es doble. Por un lado, limitar la cantidad de documentos o pasajes devueltos por el sistema, y por otro, asegurar que dichos documentos o pasajes son buenos candidatos a contener la respuesta.

- (Moreda *et al.*, 2004b).

  El artículo presenta los primeros intentos de desarrollo de una herramienta de anotación de roles semánticos que combina un conjunto de heurísticas, con estrategias de aprendizaje automático supervisado. La herramienta, denominada SemRol, determina dos argumentos para cada verbo, uno a la izquierda y otro a la derecha, atendiendo a las reglas definidas. El algoritmo de aprendizaje automático asigna el rol que los argumentos identificados juegan. Para ello, SemRol hace uso, únicamente, de las palabras que forman cada argumento y de su categoría gramatical.

- (Moreda *et al.*, 2004c).

  Una nueva versión de las heurísticas consideradas en el trabajo anterior incorporan cierta mejoría ($F_{\beta=1}$ +3 puntos) a la hora de identificar argumentos de un verbo. En esta ocasión información acerca de sintagmas y cláusulas juega un papel fundamental.

- (Moreda *et al.*, 2005).

La incorporación de una nueva fase en la tarea de anotación de roles semánticos que determine el sentido del verbo en la oración a anotar, supone una novedad para esta clase de sistemas. La necesidad de dicha incorporación surge del hecho, de que dependiendo del sentido del verbo, el conjunto de roles semánticos a considerar puede ser diferente. En consecuencia, la clasificación se afronta como una multitarea donde cada verbo y su sentido es tratado por un clasificador diferente.

- (Navarro *et al.*, 2004).

  Este trabajo expone los principios generales y los principales roles semánticos con los que anotar el corpus 3LB. Puesto que no existe un acuerdo en la comunidad científica en lo que a roles semánticos se refiere, lo primero que se debe hacer al anotar un corpus, es determinar qué roles se van a utilizar. La propuesta de roles realizada tiene una clara aplicación en tareas de búsqueda de respuestas, sigue una organización jerárquica y define un conjunto de roles generales, aplicables a diferentes verbos que compartan rasgos semánticos similares. Además, la lista de roles propuesta se basa en los roles generales de PropBank y VerbNet, y tiene en cuenta los utilizados en FrameNet.

- (Moreda *et al.*, 2008b).

  La principal aportación de este artículo es analizar la influencia de los roles semánticos en sistemas de búsqueda de respuestas. Con este fin, se construye un sistema de búsqueda de respuestas, que hace uso de dos módulos de extracción de respuestas candidatas basados en roles semánticos. Uno de los módulos utiliza reglas semánticas que determinan, dada una pregunta, el tipo de rol que debe tener un argumento candidato a contener la respuesta. El otro, construye un conjunto de patrones semánticos que permiten extraer los argumentos candidatos a contener la respuesta siempre que su patrón sea uno de los contemplados. Ambas aproximaciones son evaluadas y comparadas utilizando un subconjunto de preguntas de tipo lugar de las preguntas del TREC8 y TREC9. Los resultados de dicha evaluación demuestran que la aproximación basada en patrones supera a la apro-

ximación basada en reglas en un 40,80 % respecto a la medida $F_{\beta=1}$.

- (Moreda *et al.*, 2008a).

  El objetivo de este artículo es llevar a cabo una comparativa entre los sistemas de búsqueda de respuestas clásicos basados en entidades nombradas, y los basados en roles semánticos. Dado que los conjunto de preguntas disponibles actualmente contienen, en su gran mayoría, preguntas cuya respuesta es una entidad nombrada, es de esperar que los roles semánticos no sean capaces de mejorar a tales sistemas. Sin embargo, utilizando un conjunto de preguntas cuya respuesta no sea una entidad nombrada, el comportamiento de ambos tipos de sistemas está aún por demostrar. Por ello, se realizan pruebas sobre un conjunto de 50 preguntas de lugar cuya respuesta es una entidad, extraídas del TREC8 y TREC9; y otro conjunto de 50 preguntas de lugar creadas especialmente para la prueba, y cuya respuesta no es una entidad nombrada. Los resultados muestran que, si bien como era de esperar las entidades superan a los roles en el caso de preguntas con respuestas en entidades (MRR +66,98 % sobre reglas y MRR +49,57 % sobre patrones); el comportamiento cambia radicalmente para el caso de preguntas cuya respuesta no es una entidad (MRR +142,25 % para reglas y MRR +223,48 % para patrones).

Universitat d'Alacant
Universidad de Alicante

7.4 Trabajo en progreso y futuro

Sin duda alguna un trabajo de estas características, integrado en un grupo de investigación consolidado e inmerso en diferentes proyectos de investigación nacional e internacional requiere una continuación que permita reforzar la investigación en roles semánticos.

Respecto a los conjuntos de roles semánticos y los recursos lingüísticos disponibles, y atendiendo a las investigaciones de este trabajo, se pretende liderar la consolidación de los conjuntos de

roles semánticos y sus relaciones. Para ello, se trabajará en la línea de consensuar o estandarizar los conjuntos de roles para recursos lingüísticos, principalmente para el caso del español y del catalán. Más concretamete, y dado que en la actualidad se dispone de herramientas que establecen de forma automática relaciones entre PropBank y VerbNet, se quiere utilizar estas herramientas y VerbNet como un paso intermedio que establezca relaciones entre PropBank y nuestro conjunto de roles.

Respecto al desarrollo de herramientas de anotación de roles semánticos eficaces y eficientes, es preciso desarrollar un motor de anotación de roles semánticos adaptable a casos de uso. Actualmente, no existe una herramienta estándar para todos los casos de uso de la tecnología del lenguaje humano. Por ello, el objetivo es desarrollar una herramienta con un módulo central adaptable a cada una de las aplicaciones como puede ser la búsqueda de respuestas, o la implicación textual, entre otras.

Respecto a la influencia de los roles semánticos en los sistemas de búsqueda de respuestas,y una vez establecidas las relaciones de forma adecuada y completa entre PropBank y nuestro conjunto de roles, se dispondrá de un corpus adecuado y suficientemente grande como para afrontar la evaluación de la utilidad de dicho corpus en la tarea para la que ha sido definido. Se pretende llevar a cabo una comparativa entre el comportamiento de los sistema de búsqueda de respuesta basados en conjuntos de roles de carácter general, como PropBank, frente a un conjunto de roles diseñado *ad-hoc* para la búsqueda de respuestas.

Además, se están reforzando las reglas utilizadas por los módulos de extracción de respuestas basados en roles, con el fin de definir un proceso automático de generalización de reglas y patrones. De esta manera el sistema de búsqueda de respuestas estaría capacitado para responder cualquier tipo de pregunta de tipo factual, y no sólo de lugar. Esto, a su vez, permitiría analizar el comportamiento de otros tipos de roles semánticos respecto a los sistemas de búsqueda de respuestas.

Por otro lado, las investigaciones realizadas en este trabajo, indican que hay una hipótesis de trabajo en el dominio de la biomedicina, según la cual los roles semánticos representando tiempo, lugar y negación, son necesarios para definir relaciones semánticas en el proceso de extracción de conocimiento y deducción del mismo. La adaptación de SemRol a este tipo de dominios permitiría ahondar en el estudio de tal hipótesis.

Finalmente, se pretende que la semántica, y en concreto los roles semánticos, formen parte de cada una de las aplicaciones de la tecnología del lenguaje humano vinculadas a proyectos de investigación que se desarrollan en el seno del grupo de investigación.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

8. Anexo

En este apartado se muestra el detalle del proceso de ajuste realizado con el algoritmo de aprendizaje TiMBL y para la estrategia de anotación por sentidos cuando se considera el conjunto completo de roles semánticos.

Los cuadros mostrados siguen el orden de cardinalidad de los conjuntos de características, de manera que el primer cuadro (cuadro 8.1) muestra las combinaciones de conjuntos de características de cardinalidad uno, el segundo (cuadro 8.2), las de cardinalidad dos, y así sucesivamente. En casa paso, la combinación de características con mejor medida $F_{\beta=1}$ es seleccionada para combinar con el resto de características en el paso siguiente. Por ejemplo, el mejor resultado para combinaciones de una característica ($F_{\beta=1}$ = 61,33 %) se obtiene con la característica F39 (cuadro 8.1, por lo que en el siguiente paso se calcularán los valores de la medida $F_{\beta=1}$ para todas las posibles combinaciones de F39 con el resto de características (cuadro 8.2. A su vez, el mejor resultado para combinaciones de dos características ($F_{\beta=1}$ = 69,41 %), se obtiene con la combinación F1,F39, la cual se utilizará en el siguiente paso para calcular los valores de la medida $F_{\beta=1}$ para las combinaciones de F1,F39 con el resto de características (cuadro 8.3. Y así sucesivamente, hasta que ninguna de las combinaciones supere los resultados obtenidos en el paso anterior respecto a la medida $F_{\beta=1}$. Este es el caso de los cuadros 8.23 a 8.30, en los que ninguna de las combinaciones de trece y catorce características supera el mejor valor obtenido con combinaciones de doce características ($F_{\beta=1}$ = 76,34 %, cuadro 8.20).

En estos cuadros puede observarse que cuando más de una combinación de características obtiene el mejor resultado, todas esas combinaciones pasan a la fase siguiente. Por ejemplo, cuando se realizan las combinaciones de ocho características (cuadro 8.8), las combinaciones F0,F1,F2,F13,F18,F35,F39,F43 y F0,F1,F2,F13,F30,F35,F39,F43 obtienen la mejor medida $F_{\beta=1}$ (75,91 %), por lo que ambos subconjuntos de características son base de combinación para el resto de características en subconjuntos de cardinalidad nueve (cuadros 8.9 y 8.10). Esta situación se repite en combinaciones de nueve, diez, once, doce y trece características.

El arrastre de combinaciones tiene lugar hasta que o bien se produce un desempate entre las diferentes combinaciones, o bien la cantidad de combinaciones es tan grande que resulta demasiado costoso seguir arrastrándolas, recordemos que evitamos estrategias exponenciales, y se hace necesario llevar a cabo un proceso de selección. El criterio de selección establecido es simplicidad de la característica.

Un ejemplo de desempate se muestra en el cuadro 8.14, en el que ninguna de las combinaciones supera el valor máximo obtenido hasta el momento. El mejor valor alcanzado por estas combinaciones es $F_{\beta=1} = 76{,}13\,\%$, mientras que combinaciones con otros conjuntos de características alcanzan una medida $F_{\beta=1} = 76{,}14\,\%$. Por ello, la combinación de características F0,F1,F2,F13,F14,F30,F35,F39,F43, no se arrastra al siguiente nivel.

Un ejemplo en el que se ha aplicado el criterio de simplicidad se encuentra en el caso de las combinaciones de trece características (cuadros 8.23 a 8.25), donde trece combinaciones obtienen el máximo valor para la medida $F_{\beta=1}$ (76,34 %). Dado que arrastrar tantas combinaciones resulta demasiado costoso, se analizan las características con el fin de seleccionar las más fáciles de generar. Concretamente en todos los casos F18 y F30 son características muy similares que consideran los adverbios del argumento, y los adverbios del argument y su PoS, respectivamente, por lo que a igualdad de resultados, se elige simplicidad de características, se-

leccionando F18. Por otro lado, F22 y F33, representan la raíz del núcleo de los sintagmas del argumento, y la raíz del núcleo de los sintagmas del argumento y su PoS. Por las mismas razones que antes se selecciona la característica F22. En consecuencia, se descartan las combinaciones F0, F1, F2, F13, F20, F22, F30, F35, F37, F39, F43, F45, {F16|F25|F26|F27|F28} por contener la característica F30, y las combinaciones F0, F1, F2, F13, F18, F20, F33, F35, F37, F39, F43, F45, {F25|F26|F27} por contener F33. Este descarte elimina ocho de las posibles trece combinaciones. Para el siguiente paso, sólo se mantienen las combinaciones F0, F1, F2, F13, F18, F20, F22, F35, F37, F39, F43, F45, {F16|F25|F26|F27|F28} con el resto de características. Como resultado de la selección, las trece posibles combinaciones quedan reducidas a cinco.

Es importante recordar que el proceso de ajuste se afronta haciendo uso de estrategias de validación cruzada con k=3, por lo que los resultados mostrados en todos los cuadros corresponden a los valores medios obtenidos end dicho proceso de validación cruzada. Es decir, en realidad, cada fila de un cuadro corresponde a tres ejecuciones diferentes, cada una con sus propios corpus de entrenamiento y test.

Procesos similares se han llevado a cabo utilizando ME como algoritmo de aprendizaje; siguiendo una estrategia de anotación única, en lugar de por sentidos; y para cada uno de los clasificadores individuales.

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
0	34,72	34,55	34,59
1	54,49	53,96	54,23
2	43,24	42,72	42,98
3	45,36	44,91	45,13
4	45,61	45,14	45,37
5	38,13	37,78	37,95
6	44,23	43,78	44,00
7	47,23	46,77	47,00
8	38,06	37,71	37,89
9	38,65	38,28	38,46
10	38,28	37,91	38,09
11	52,75	52,27	52,51
12	53,53	53,04	53,28
13	42,39	42,08	42,23
14	46,66	46,24	46,45
15	49,56	49,10	49,34
16	44,31	43,93	44,12
17	41,06	40,75	40,91
18	40,26	39,95	40,10
19	48,59	48,22	48,40
20	41,68	41,38	41,53
21	48,61	48,25	48,43
22	46,71	46,29	46,50
23	56,64	56,19	56,42
24	53,16	52,66	52,91
25	34,69	34,44	34,56
26	34,69	34,44	34,56
27	34,74	34,51	34,62
28	44,30	43,92	44,11
29	41,01	36,96	38,68
30	40,26	39,95	40,10
31	48,28	33,46	45,52
32	48,63	48,27	48,45
33	46,72	46,29	46,51
34	43,29	42,91	43,10
35	51,36	50,93	51,14
36	51,27	50,85	51,06
37	51,56	51,12	51,34
38	51,50	51,07	51,28
39	**61,69**	**60,98**	**61,33**
40	52,00	51,29	51,64
41	58,50	57,71	58,11
42	52,17	51,45	51,81
43	52,22	51,50	51,86
44	52,24	51,54	51,89
45	59,11	58,53	58,82

Cuadro 8.1. Combinaciones con 1 característica. TiMBL. Anotación por sentidos

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
0,39	61,83	61,12	61,48
1,39	**69,83**	**68,99**	**69,41**
2,39	64,82	64,01	64,42
3,39	65,25	64,45	64,84
4,39	63,52	62,73	63,13
5,39	62,43	61,68	62,05
6,39	64,48	63,67	64,07
7,39	63,42	62,65	63,03
8,39	62,39	61,64	62,01
9,39	62,11	61,36	61,73
10,39	61,99	61,25	61,62
11,39	63,51	62,78	63,14
12,39	63,55	62,82	63,18
13,39	64,21	63,47	63,84
14,39	63,42	62,68	63,05
15,39	62,53	61,79	62,16
16,39	63,33	62,59	62,96
17,39	62,07	61,36	61,71
18,39	62,78	62,06	62,42
19,39	64,41	63,67	64,03
20,39	63,02	62,3	62,66
21,39	64,4	63,66	64,03
22,39	63,42	62,68	63,04
23,39	62,78	62,06	62,42
24,39	62,62	61,9	62,26
25,39	61,69	60,98	61,21
26,39	61,69	60,98	61,33
27,39	61,69	60,98	61,33
28,39	63,33	62,59	62,96
29,39	61,98	55,69	58,37
30,39	62,78	62,28	62,42
31,39	64,04	57,45	60,27
32,39	64,37	63,64	64,00
33,39	63,41	62,67	63,04
34,39	62,64	61,89	62,27
35,39	64,95	64,2	64,58
36,39	64,89	64,15	64,52
37,39	64,98	64,23	64,60
38,39	64,90	64,16	64,53
40,39	66,87	65,95	66,41
41,39	67,88	66,96	67,42
42,39	66,79	65,87	66,32
43,39	66,80	65,88	66,33
44,39	67,05	66,12	66,58
45,39	69,21	68,35	68,78

Cuadro 8.2. Combinaciones con 2 características. TiMBL. Anotación por sentidos

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
0,1,39	70,19	69,33	69,76
2,1,39	71,26	70,35	70,80
3,1,39	71,18	70,22	70,70
4,1,39	70,30	69,42	69,86
5,1,39	70,23	69,36	69,79
6,1,39	70,85	69,90	70,37
7,1,39	70,37	69,49	69,93
8,1,39	70,19	69,32	69,75
9,1,39	69,89	69,02	69,45
10,1,39	69,86	69,00	69,43
11,1,39	71,04	70,17	70,61
12,1,39	71,08	70,22	70,64
13,1,39	**72,31**	**71,43**	**71,87**
14,1,39	70,77	69,91	70,33
15,1,39	69,91	69,05	69,48
16,1,39	70,76	69,89	70,32
17,1,39	70,06	69,22	69,64
18,1,39	71,07	70,22	70,64
19,1,39	71,66	70,80	71,23
20,1,39	71,18	70,32	70,75
21,1,39	71,66	70,79	71,22
22,1,39	70,75	69,89	70,32
23,1,39	70,34	69,49	69,91
24,1,39	70,06	69,20	69,62
25,1,39	69,83	68,99	69,41
26,1,39	69,83	68,99	69,41
27,1,39	69,83	68,99	69,41
28,1,39	70,76	69,89	70,32
29,1,39	69,85	62,65	65,72
30,1,39	71,07	70,22	70,64
31,1,39	71,37	64,00	67,15
32,1,39	71,64	70,78	71,20
33,1,39	70,75	69,89	70,32
34,1,39	69,87	69,00	69,43
35,1,39	72,11	71,24	71,67
36,1,39	72,05	71,18	71,61
37,1,39	72,12	71,25	71,68
38,1,39	72,06	71,19	71,62
40,1,39	71,96	70,95	71,45
41,1,39	72,06	71,07	71,56
42,1,39	72,03	71,03	71,53
43,1,39	72,04	71,03	71,53
44,1,39	71,93	70,91	71,41
45,1,39	71,72	70,75	71,23

Cuadro 8.3. Combinaciones con 3 características. TiMBL. Anotación por sentidos

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
0,1,13,39	72,60	71,71	72,15
2,1,13,39	73,43	72,48	72,96
3,1,13,39	73,32	72,33	72,82
4,1,13,39	72,57	71,64	72,10
5,1,13,39	72,55	71,64	72,09
6,1,13,39	73,02	72,04	72,53
7,1,13,39	72,62	71,70	72,16
8,1,13,39	72,52	71,61	72,06
9,1,13,39	72,25	71,35	71,79
10,1,13,39	72,24	71,35	71,79
11,1,13,39	72,84	71,95	72,39
12,1,13,39	72,88	71,99	72,43
14,1,13,39	72,81	71,91	72,35
15,1,13,39	72,04	71,15	71,59
16,1,13,39	72,79	71,89	72,34
17,1,13,39	72,26	71,39	71,82
18,1,13,39	73,52	72,64	73,08
19,1,13,39	73,78	72,88	73,33
20,1,13,39	72,48	71,59	72,03
21,1,13,39	73,77	72,88	73,32
22,1,13,39	72,79	71,90	72,34
23,1,13,39	72,32	71,45	71,88
24,1,13,39	72,08	71,19	71,63
25,1,13,39	72,31	71,43	71,87
26,1,13,39	72,31	71,43	71,87
27,1,13,39	72,31	71,43	71,87
28,1,13,39	72,79	71,89	72,34
29,1,13,39	71,98	64,55	67,73
30,1,13,39	73,52	72,63	73,08
31,1,13,39	73,47	65,87	69,12
32,1,13,39	73,76	72,86	73,31
33,1,13,39	72,79	71,90	72,34
34,1,13,39	72,11	71,21	71,65
35,1,13,39	74,11	73,21	73,65
36,1,13,39	74,05	73,15	73,60
37,1,13,39	74,11	73,21	73,66
38,1,13,39	74,05	73,15	73,60
40,1,13,39	74,12	73,08	73,60
41,1,13,39	74,14	73,13	73,63
42,1,13,39	74,19	73,15	73,66
43,1,13,39	**74,19**	**73,15**	**73,67**
44,1,13,39	74,10	73,06	73,58
45,1,13,39	73,91	72,91	73,41

Cuadro 8.4. Combinaciones con 4 características. TiMBL. Anotación por sentidos

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
0,1,13,39,43	74,44	73,39	73,91
2,1,13,39,43	74,84	73,78	74,31
3,1,13,39,43	74,38	73,31	73,84
4,1,13,39,43	74,02	72,96	73,49
5,1,13,39,43	73,96	72,91	73,43
6,1,13,39,43	74,11	73,03	73,57
7,1,13,39,43	74,10	73,04	73,56
8,1,13,39,43	73,95	72,90	73,42
9,1,13,39,43	74,15	73,11	73,62
10,1,13,39,43	74,14	73,10	73,61
11,1,13,39,43	74,46	73,42	73,93
12,1,13,39,43	74,48	73,44	73,96
14,1,13,39,43	74,56	73,51	74,03
15,1,13,39,43	73,91	72,87	73,38
16,1,13,39,43	74,56	73,51	74,03
17,1,13,39,43	74,08	73,06	73,56
18,1,13,39,43	75,11	74,06	74,58
19,1,13,39,43	75,29	74,23	74,75
20,1,13,39,43	74,27	73,24	73,75
21,1,13,39,43	75,28	74,22	74,75
22,1,13,39,43	74,56	73,51	74,03
23,1,13,39,43	74,13	73,11	73,62
24,1,13,39,43	73,91	72,88	73,39
25,1,13,39,43	74,19	73,15	73,67
26,1,13,39,43	74,19	73,15	73,67
27,1,13,39,43	74,19	73,15	73,67
28,1,13,39,43	74,56	73,51	74,03
29,1,13,39,43	73,79	66,05	69,36
30,1,13,39,43	75,11	74,05	74,58
31,1,13,39,43	74,98	67,09	70,47
32,1,13,39,43	75,27	74,21	74,74
33,1,13,39,43	74,55	73,50	74,02
34,1,13,39,43	73,81	72,77	73,28
35,1,13,39,43	**75,59**	**74,52**	**75,05**
36,1,13,39,43	75,52	74,46	74,99
37,1,13,39,43	75,58	74,52	75,04
38,1,13,39,43	75,53	74,47	74,99
40,1,13,39,43	73,59	72,55	73,07
41,1,13,39,43	73,85	72,80	73,32
42,1,13,39,43	73,60	72,56	73,08
44,1,13,39,43	73,62	72,58	73,10
45,1,13,39,43	74,23	73,17	73,70

Cuadro 8.5. Combinaciones con 5 características. TiMBL. Anotación por sentidos

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
0,1,13,35,39,43	75,79	74,72	75,25
2,1,13,35,39,43	**76,15**	**75,05**	**75,60**
3,1,13,35,39,43	75,66	74,56	75,11
4,1,13,35,39,43	75,17	74,09	74,63
5,1,13,35,39,43	75,33	74,26	74,79
6,1,13,35,39,43	75,45	74,35	74,89
7,1,13,35,39,43	75,21	74,13	74,67
8,1,13,35,39,43	75,32	74,24	74,78
9,1,13,35,39,43	75,50	74,44	74,97
10,1,13,35,39,43	75,49	74,43	74,96
11,1,13,35,39,43	75,39	74,33	74,86
12,1,13,35,39,43	75,36	74,31	74,83
14,1,13,35,39,43	75,66	74,59	75,12
15,1,13,35,39,43	75,21	74,15	74,68
16,1,13,35,39,43	75,66	74,58	75,12
17,1,13,35,39,43	75,41	74,36	74,88
18,1,13,35,39,43	75,66	74,60	75,13
19,1,13,35,39,43	75,63	74,56	75,09
20,1,13,35,39,43	75,66	74,59	75,12
21,1,13,35,39,43	75,63	74,56	75,09
22,1,13,35,39,43	75,65	74,58	75,11
23,1,13,35,39,43	75,34	74,29	74,81
24,1,13,35,39,43	75,19	74,13	74,66
25,1,13,35,39,43	75,59	74,52	75,05
26,1,13,35,39,43	75,59	74,52	75,05
27,1,13,35,39,43	75,59	74,52	75,05
28,1,13,35,39,43	75,66	74,58	75,12
29,1,13,35,39,43	75,10	67,22	70,59
30,1,13,35,39,43	75,66	74,59	75,12
31,1,13,35,39,43	75,33	67,40	70,79
32,1,13,35,39,43	75,63	74,56	75,09
33,1,13,35,39,43	75,65	74,58	75,11
34,1,13,35,39,43	75,10	74,03	74,56
36,1,13,35,39,43	75,65	74,59	75,12
37,1,13,35,39,43	75,65	74,59	75,12
38,1,13,35,39,43	75,64	74,59	75,11
40,1,13,35,39,43	75,11	74,04	74,57
41,1,13,35,39,43	75,16	74,09	74,62
42,1,13,35,39,43	75,12	74,05	74,58
44,1,13,35,39,43	75,12	74,05	74,58
45,1,13,35,39,43	75,35	74,26	74,80

Cuadro 8.6. Combinaciones con 6 características. TiMBL. Anotación por sentidos

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
0,1,2,13,35,39,43	**76,33**	**75,22**	**75,77**
3,1,2,13,35,39,43	76,12	75,00	75,56
4,1,2,13,35,39,43	75,86	74,74	75,30
5,1,2,13,35,39,43	75,78	74,67	75,22
6,1,2,13,35,39,43	75,91	74,79	75,34
7,1,2,13,35,39,43	75,83	74,71	75,27
8,1,2,13,35,39,43	75,77	74,65	75,20
9,1,2,13,35,39,43	75,95	74,85	75,39
10,1,2,13,35,39,43	75,92	74,83	75,37
11,1,2,13,35,39,43	75,94	74,84	75,39
12,1,2,13,35,39,43	75,91	74,82	75,37
14,1,2,13,35,39,43	76,23	75,13	75,68
15,1,2,13,35,39,43	75,75	74,65	75,19
16,1,2,13,35,39,43	76,21	75,10	75,65
17,1,2,13,35,39,43	75,92	74,83	75,37
18,1,2,13,35,39,43	76,30	75,20	75,74
19,1,2,13,35,39,43	76,28	75,17	75,72
20,1,2,13,35,39,43	76,22	75,13	75,67
21,1,2,13,35,39,43	76,28	75,17	75,72
22,1,2,13,35,39,43	76,23	75,12	75,67
23,1,2,13,35,39,43	75,94	74,85	75,39
24,1,2,13,35,39,43	75,78	74,69	75,23
25,1,2,13,35,39,43	76,15	75,05	75,60
26,1,2,13,35,39,43	76,15	75,05	75,60
27,1,2,13,35,39,43	76,15	75,05	75,60
28,1,2,13,35,39,43	76,21	75,10	75,65
29,1,2,13,35,39,43	75,57	67,60	71,01
30,1,2,13,35,39,43	76,30	75,20	75,74
31,1,2,13,35,39,43	75,95	67,93	71,36
32,1,2,13,35,39,43	76,28	75,17	75,72
33,1,2,13,35,39,43	76,23	75,12	75,67
34,1,2,13,35,39,43	75,58	74,48	75,02
36,1,2,13,35,39,43	76,27	75,17	75,71
37,1,2,13,35,39,43	76,27	75,17	75,72
38,1,2,13,35,39,43	76,27	75,17	75,72
40,1,2,13,35,39,43	75,75	74,64	75,19
41,1,2,13,35,39,43	75,78	74,67	75,22
42,1,2,13,35,39,43	75,78	74,67	75,22
44,1,2,13,35,39,43	75,77	74,66	75,21
45,1,2,13,35,39,43	75,45	74,31	74,87

Cuadro 8.7. Combinaciones con 7 características. TiMBL. Anotación por sentidos

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,35,39,43	76,25	75,12	75,68
4,0,1,2,13,35,39,43	76,02	74,90	75,45
5,0,1,2,13,35,39,43	75,94	74,82	75,38
6,0,1,2,13,35,39,43	76,06	74,92	75,48
7,0,1,2,13,35,39,43	75,99	74,87	75,42
8,0,1,2,13,35,39,43	75,93	74,81	75,36
9,0,1,2,13,35,39,43	76,13	75,02	75,57
10,0,1,2,13,35,39,43	76,11	75,01	75,55
11,0,1,2,13,35,39,43	76,08	74,98	75,53
12,0,1,2,13,35,39,43	76,07	74,97	75,51
14,0,1,2,13,35,39,43	76,39	75,28	75,83
15,0,1,2,13,35,39,43	75,88	74,78	75,33
16,0,1,2,13,35,39,43	76,37	75,26	75,81
17,0,1,2,13,35,39,43	76,08	74,98	75,52
18,0,1,2,13,35,39,43	**76,47**	**75,36**	**75,91**
19,0,1,2,13,35,39,43	76,44	75,33	75,88
20,0,1,2,13,35,39,43	76,39	75,28	75,83
21,0,1,2,13,35,39,43	76,43	75,33	75,88
22,0,1,2,13,35,39,43	76,39	75,28	75,82
23,0,1,2,13,35,39,43	76,07	74,97	75,52
24,0,1,2,13,35,39,43	75,92	74,83	75,37
25,0,1,2,13,35,39,43	76,33	75,22	75,77
26,0,1,2,13,35,39,43	76,33	75,22	75,77
27,0,1,2,13,35,39,43	76,33	75,22	75,77
28,0,1,2,13,35,39,43	76,37	75,26	75,81
29,0,1,2,13,35,39,43	75,71	67,73	71,14
30,0,1,2,13,35,39,43	**76,47**	75,36	**75,91**
31,0,1,2,13,35,39,43	76,09	68,05	71,49
32,0,1,2,13,35,39,43	76,44	75,33	75,88
33,0,1,2,13,35,39,43	76,38	75,27	75,83
34,0,1,2,13,35,39,43	75,73	74,62	75,17
36,0,1,2,13,35,39,43	76,43	75,33	75,88
37,0,1,2,13,35,39,43	76,43	75,34	75,88
38,0,1,2,13,35,39,43	76,44	75,34	75,88
40,0,1,2,13,35,39,43	75,91	74,80	75,35
41,0,1,2,13,35,39,43	75,95	74,84	75,39
42,0,1,2,13,35,39,43	75,94	74,82	75,38
44,0,1,2,13,35,39,43	75,93	74,81	75,37
45,0,1,2,13,35,39,43	76,07	74,95	75,51

Cuadro 8.8. Combinaciones con 8 características. TiMBL. Anotación por sentidos

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,35,39,43	76,52	75,38	75,94
4,0,1,2,13,18,35,39,43	76,31	75,18	75,75
5,0,1,2,13,18,35,39,43	76,15	75,03	75,59
6,0,1,2,13,18,35,39,43	76,34	75,20	75,76
7,0,1,2,13,18,35,39,43	76,26	75,15	75,70
8,0,1,2,13,18,35,39,43	76,13	75,02	75,57
9,0,1,2,13,18,35,39,43	76,29	75,18	75,73
10,0,1,2,13,18,35,39,43	76,26	75,16	75,71
11,0,1,2,13,18,35,39,43	76,25	75,14	75,69
12,0,1,2,13,18,35,39,43	76,23	75,13	75,68
14,0,1,2,13,18,35,39,43	**76,58**	**75,47**	**76,02**
15,0,1,2,13,18,35,39,43	76,07	74,97	75,52
16,0,1,2,13,18,35,39,43	76,54	75,43	75,98
17,0,1,2,13,18,35,39,43	76,25	75,16	75,70
19,0,1,2,13,18,35,39,43	76,48	75,37	75,92
20,0,1,2,13,18,35,39,43	76,53	75,42	75,97
21,0,1,2,13,18,35,39,43	76,48	75,37	75,92
22,0,1,2,13,18,35,39,43	**76,58**	**75,47**	**76,02**
23,0,1,2,13,18,35,39,43	76,18	75,08	75,63
24,0,1,2,13,18,35,39,43	75,99	74,90	75,44
25,0,1,2,13,18,35,39,43	76,47	75,36	75,91
26,0,1,2,13,18,35,39,43	76,47	75,36	75,91
27,0,1,2,13,18,35,39,43	76,47	75,36	75,91
28,0,1,2,13,18,35,39,43	76,54	75,43	75,98
29,0,1,2,13,18,35,39,43	75,88	67,88	71,30
30,0,1,2,13,18,35,39,43	76,28	75,18	75,73
31,0,1,2,13,18,35,39,43	76,11	68,07	71,51
32,0,1,2,13,18,35,39,43	76,47	75,37	75,92
33,0,1,2,13,18,35,39,43	**76,58**	**75,46**	**76,02**
34,0,1,2,13,18,35,39,43	75,93	74,82	75,37
36,0,1,2,13,18,35,39,43	76,50	75,40	75,94
37,0,1,2,13,18,35,39,43	76,50	75,40	75,95
38,0,1,2,13,18,35,39,43	76,50	75,40	75,94
40,0,1,2,13,18,35,39,43	76,21	75,08	75,64
41,0,1,2,13,18,35,39,43	76,33	75,21	75,77
42,0,1,2,13,18,35,39,43	76,22	75,10	75,66
44,0,1,2,13,18,35,39,43	76,24	75,11	75,67
45,0,1,2,13,18,35,39,43	76,50	75,37	75,93

Cuadro 8.9. Combinaciones con 9 características. TiMBL. Anotación por sentidos (1/2)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,30,35,39,43	76,51	75,38	75,94
4,0,1,2,13,30,35,39,43	76,31	75,18	75,74
5,0,1,2,13,30,35,39,43	76,15	75,03	75,59
6,0,1,2,13,30,35,39,43	76,33	75,19	75,76
7,0,1,2,13,30,35,39,43	76,26	75,14	75,70
8,0,1,2,13,30,35,39,43	76,13	75,01	75,57
9,0,1,2,13,30,35,39,43	76,28	75,18	75,73
10,0,1,2,13,30,35,39,43	76,26	75,16	75,70
11,0,1,2,13,30,35,39,43	76,24	75,14	75,69
12,0,1,2,13,30,35,39,43	76,23	75,13	75,67
14,0,1,2,13,30,35,39,43	**76,58**	**75,46**	**76,02**
15,0,1,2,13,30,35,39,43	76,07	74,97	75,51
16,0,1,2,13,30,35,39,43	76,53	75,43	75,98
17,0,1,2,13,30,35,39,43	76,25	75,16	75,70
18,0,1,2,13,30,35,39,43	76,28	75,18	75,73
19,0,1,2,13,30,35,39,43	76,47	75,37	75,92
20,0,1,2,13,30,35,39,43	76,53	75,42	75,97
21,0,1,2,13,30,35,39,43	76,48	75,37	75,92
22,0,1,2,13,30,35,39,43	**76,57**	**75,46**	**76,02**
23,0,1,2,13,30,35,39,43	76,18	75,08	75,63
24,0,1,2,13,30,35,39,43	75,99	74,90	75,44
25,0,1,2,13,30,35,39,43	76,47	75,36	75,91
26,0,1,2,13,30,35,39,43	76,47	75,36	75,91
27,0,1,2,13,30,35,39,43	76,47	75,36	75,91
28,0,1,2,13,30,35,39,43	76,53	75,43	75,98
29,0,1,2,13,30,35,39,43	75,87	67,99	71,37
31,0,1,2,13,30,35,39,43	76,11	68,07	71,51
32,0,1,2,13,30,35,39,43	76,47	75,37	75,91
33,0,1,2,13,30,35,39,43	76,57	75,46	76,01
34,0,1,2,13,30,35,39,43	75,92	74,82	75,37
36,0,1,2,13,30,35,39,43	76,49	75,39	75,94
37,0,1,2,13,30,35,39,43	76,50	75,40	75,94
38,0,1,2,13,30,35,39,43	76,49	75,39	75,94
40,0,1,2,13,30,35,39,43	76,21	75,08	75,64
41,0,1,2,13,30,35,39,43	76,33	75,21	75,77
42,0,1,2,13,30,35,39,43	76,22	75,10	75,65
44,0,1,2,13,30,35,39,43	76,24	75,11	75,67
45,0,1,2,13,30,35,39,43	76,50	75,37	75,93

Cuadro 8.10. Combinaciones con 9 características. TiMBL. Anotación por sentidos (2/2)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,14,18,35,39,43	76,67	75,53	76,09
4,0,1,2,13,14,18,35,39,43	76,43	75,29	75,85
5,0,1,2,13,14,18,35,39,43	76,29	75,17	75,73
6,0,1,2,13,14,18,35,39,43	76,07	74,98	75,53
7,0,1,2,13,14,18,35,39,43	76,38	75,25	75,81
8,0,1,2,13,14,18,35,39,43	76,27	75,15	75,71
9,0,1,2,13,14,18,35,39,43	76,39	75,27	75,82
10,0,1,2,13,14,18,35,39,43	76,36	75,25	75,80
11,0,1,2,13,14,18,35,39,43	76,36	75,25	75,80
12,0,1,2,13,14,18,35,39,43	76,35	75,25	75,80
15,0,1,2,13,14,18,35,39,43	76,08	74,97	75,52
16,0,1,2,13,14,18,35,39,43	76,60	75,49	76,04
17,0,1,2,13,14,18,35,39,43	76,36	75,26	75,80
19,0,1,2,13,14,18,35,39,43	76,56	75,45	76,00
20,0,1,2,13,14,18,35,39,43	76,64	75,52	76,08
21,0,1,2,13,14,18,35,39,43	76,56	75,45	76,00
22,0,1,2,13,14,18,35,39,43	76,56	75,45	76,00
23,0,1,2,13,14,18,35,39,43	76,28	75,18	75,72
24,0,1,2,13,14,18,35,39,43	76,50	75,36	75,92
25,0,1,2,13,14,18,35,39,43	76,58	75,47	76,02
26,0,1,2,13,14,18,35,39,43	76,58	75,47	76,02
27,0,1,2,13,14,18,35,39,43	76,58	75,47	76,02
28,0,1,2,13,14,18,35,39,43	76,60	75,49	76,04
29,0,1,2,13,14,18,35,39,43	76,00	67,97	71,41
30,0,1,2,13,14,18,35,39,43	76,40	75,30	75,85
31,0,1,2,13,14,18,35,39,43	76,22	68,16	71,61
32,0,1,2,13,14,18,35,39,43	76,56	75,44	76,00
33,0,1,2,13,14,18,35,39,43	76,56	75,45	76,00
34,0,1,2,13,14,18,35,39,43	76,04	74,93	75,48
36,0,1,2,13,14,18,35,39,43	76,57	75,46	76,01
37,0,1,2,13,14,18,35,39,43	76,57	75,46	76,01
38,0,1,2,13,14,18,35,39,43	76,57	75,46	76,01
40,0,1,2,13,14,18,35,39,43	76,37	75,24	75,80
41,0,1,2,13,14,18,35,39,43	76,52	75,38	75,95
42,0,1,2,13,14,18,35,39,43	76,39	75,26	75,82
44,0,1,2,13,14,18,35,39,43	76,40	75,26	75,82
45,0,1,2,13,14,18,35,39,43	**76,71**	**75,57**	**76,14**

Cuadro 8.11. Combinaciones con 10 características. TiMBL. Anotación por sentidos (1/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,22,35,39,43	76,67	75,52	76,09
4,0,1,2,13,18,22,35,39,43	76,43	75,29	75,86
5,0,1,2,13,18,22,35,39,43	76,29	75,17	75,72
6,0,1,2,13,18,22,35,39,43	76,50	75,35	75,92
7,0,1,2,13,18,22,35,39,43	76,38	75,26	75,81
8,0,1,2,13,18,22,35,39,43	76,27	75,15	75,70
9,0,1,2,13,18,22,35,39,43	76,38	75,27	75,82
10,0,1,2,13,18,22,35,39,43	76,36	75,25	75,80
11,0,1,2,13,18,22,35,39,43	76,36	75,25	75,80
12,0,1,2,13,18,22,35,39,43	76,35	75,25	75,80
14,0,1,2,13,18,22,35,39,43	76,56	75,45	76,00
15,0,1,2,13,18,22,35,39,43	76,07	74,97	75,52
16,0,1,2,13,18,22,35,39,43	76,60	75,49	76,04
17,0,1,2,13,18,22,35,39,43	76,36	75,26	75,81
19,0,1,2,13,18,22,35,39,43	76,56	75,45	76,00
20,0,1,2,13,18,22,35,39,43	76,64	75,52	76,08
21,0,1,2,13,18,22,35,39,43	76,54	75,43	75,98
23,0,1,2,13,18,22,35,39,43	76,27	75,17	75,72
24,0,1,2,13,18,22,35,39,43	76,07	74,98	75,52
25,0,1,2,13,18,22,35,39,43	76,58	75,47	76,02
26,0,1,2,13,18,22,35,39,43	76,58	75,47	76,02
27,0,1,2,13,18,22,35,39,43	76,58	75,47	76,02
28,0,1,2,13,18,22,35,39,43	76,60	75,49	76,04
29,0,1,2,13,18,22,35,39,43	76,00	67,98	71,41
30,0,1,2,13,18,22,35,39,43	76,40	75,29	75,84
31,0,1,2,13,18,22,35,39,43	76,22	68,16	71,61
32,0,1,2,13,18,22,35,39,43	76,55	75,43	75,99
33,0,1,2,13,18,22,35,39,43	76,53	75,43	75,97
34,0,1,2,13,18,22,35,39,43	76,05	74,94	75,49
36,0,1,2,13,18,22,35,39,43	76,56	75,46	76,01
37,0,1,2,13,18,22,35,39,43	76,57	75,46	76,01
38,0,1,2,13,18,22,35,39,43	76,56	75,45	76,00
40,0,1,2,13,18,22,35,39,43	76,37	75,24	75,80
41,0,1,2,13,18,22,35,39,43	76,53	75,39	75,95
42,0,1,2,13,18,22,35,39,43	76,38	75,25	75,81
44,0,1,2,13,18,22,35,39,43	76,40	75,26	75,82
45,0,1,2,13,18,22,35,39,43	**76,71**	**75,58**	**76,14**

Cuadro 8.12. Combinaciones con 10 características. TiMBL. Anotación por sentidos (2/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,33,35,39,43	76,66	75,52	76,09
4,0,1,2,13,18,33,35,39,43	76,42	75,29	75,85
5,0,1,2,13,18,33,35,39,43	76,29	75,16	75,72
6,0,1,2,13,18,33,35,39,43	76,50	75,35	75,92
7,0,1,2,13,18,33,35,39,43	76,37	75,25	75,81
8,0,1,2,13,18,33,35,39,43	76,27	75,14	75,70
9,0,1,2,13,18,33,35,39,43	76,38	75,27	75,82
10,0,1,2,13,18,33,35,39,43	76,36	75,25	75,80
11,0,1,2,13,18,33,35,39,43	76,35	75,25	75,80
12,0,1,2,13,18,33,35,39,43	76,35	75,24	75,79
14,0,1,2,13,18,33,35,39,43	76,56	75,45	76,00
15,0,1,2,13,18,33,35,39,43	76,07	74,97	75,51
16,0,1,2,13,18,33,35,39,43	76,60	75,49	76,04
17,0,1,2,13,18,33,35,39,43	76,35	75,26	75,80
19,0,1,2,13,18,33,35,39,43	76,56	75,44	76,00
20,0,1,2,13,18,33,35,39,43	76,63	75,52	76,08
21,0,1,2,13,18,33,35,39,43	76,55	75,43	75,99
22,0,1,2,13,18,33,35,39,43	76,53	75,43	75,97
23,0,1,2,13,18,33,35,39,43	76,27	75,17	75,72
24,0,1,2,13,18,33,35,39,43	76,07	74,98	75,52
25,0,1,2,13,18,33,35,39,43	76,58	75,46	76,02
26,0,1,2,13,18,33,35,39,43	76,58	75,46	76,02
27,0,1,2,13,18,33,35,39,43	76,58	75,46	76,02
28,0,1,2,13,18,33,35,39,43	76,60	75,49	76,04
29,0,1,2,13,18,33,35,39,43	75,99	67,97	71,40
30,0,1,2,13,18,33,35,39,43	76,40	75,29	75,84
31,0,1,2,13,18,33,35,39,43	76,21	68,16	71,60
32,0,1,2,13,18,33,35,39,43	76,55	75,43	75,98
34,0,1,2,13,18,33,35,39,43	76,04	74,93	75,48
36,0,1,2,13,18,33,35,39,43	76,56	75,45	76,00
37,0,1,2,13,18,33,35,39,43	76,56	75,46	76,01
38,0,1,2,13,18,33,35,39,43	76,56	75,45	76,00
40,0,1,2,13,18,33,35,39,43	76,37	75,24	75,80
41,0,1,2,13,18,33,35,39,43	76,52	75,39	75,95
42,0,1,2,13,18,33,35,39,43	76,38	75,25	75,81
44,0,1,2,13,18,33,35,39,43	76,39	75,25	75,82
45,0,1,2,13,18,33,35,39,43	**76,71**	**75,58**	**76,14**

Cuadro 8.13. Combinaciones con 10 características. TiMBL. Anotación por sentidos (3/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,14,30,35,39,43	76,66	75,52	76,09
4,0,1,2,13,14,30,35,39,43	76,43	75,29	75,85
5,0,1,2,13,14,30,35,39,43	76,29	75,17	75,72
6,0,1,2,13,14,30,35,39,43	76,49	75,35	75,92
7,0,1,2,13,14,30,35,39,43	76,38	75,25	75,81
8,0,1,2,13,14,30,35,39,43	76,27	75,14	75,70
9,0,1,2,13,14,30,35,39,43	76,38	75,27	75,82
10,0,1,2,13,14,30,35,39,43	76,36	75,25	75,80
11,0,1,2,13,14,30,35,39,43	76,36	75,25	75,80
12,0,1,2,13,14,30,35,39,43	76,35	75,25	75,79
15,0,1,2,13,14,30,35,39,43	76,07	74,97	75,52
16,0,1,2,13,14,30,35,39,43	76,60	75,49	76,04
17,0,1,2,13,14,30,35,39,43	76,36	75,26	75,80
18,0,1,2,13,14,30,35,39,43	76,40	75,30	75,85
19,0,1,2,13,14,30,35,39,43	76,56	75,45	76,00
20,0,1,2,13,14,30,35,39,43	76,64	75,52	76,08
21,0,1,2,13,14,30,35,39,43	76,56	75,44	76,00
22,0,1,2,13,14,30,35,39,43	76,55	75,45	76,00
23,0,1,2,13,14,30,35,39,43	76,27	75,17	75,72
24,0,1,2,13,14,30,35,39,43	76,07	74,98	75,52
25,0,1,2,13,14,30,35,39,43	76,58	75,46	76,02
26,0,1,2,13,14,30,35,39,43	76,58	75,46	76,02
27,0,1,2,13,14,30,35,39,43	76,58	75,46	76,02
28,0,1,2,13,14,30,35,39,43	76,60	75,49	76,04
29,0,1,2,13,14,30,35,39,43	75,99	67,97	71,40
31,0,1,2,13,14,30,35,39,43	76,22	68,16	71,60
32,0,1,2,13,14,30,35,39,43	76,56	75,44	76,00
33,0,1,2,13,14,30,35,39,43	76,56	75,45	76,00
34,0,1,2,13,14,30,35,39,43	76,04	74,93	75,48
36,0,1,2,13,14,30,35,39,43	76,56	75,46	76,01
37,0,1,2,13,14,30,35,39,43	76,57	75,46	76,01
38,0,1,2,13,14,30,35,39,43	76,56	75,46	76,01
40,0,1,2,13,14,30,35,39,43	76,37	75,24	75,80
41,0,1,2,13,14,30,35,39,43	76,52	75,38	75,95
42,0,1,2,13,14,30,35,39,43	76,39	75,26	75,82
44,0,1,2,13,14,30,35,39,43	76,39	75,26	75,82
45,0,1,2,13,14,30,35,39,43	76,71	75,57	76,13

Cuadro 8.14. Combinaciones con 10 características. TiMBL. Anotación por sentidos (4/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,22,30,35,39,43	76,67	75,52	76,09
4,0,1,2,13,22,30,35,39,43	76,43	75,29	75,85
5,0,1,2,13,22,30,35,39,43	76,29	75,17	75,72
6,0,1,2,13,22,30,35,39,43	76,49	75,35	75,91
7,0,1,2,13,22,30,35,39,43	76,38	75,25	75,81
8,0,1,2,13,22,30,35,39,43	76,27	75,15	75,70
9,0,1,2,13,22,30,35,39,43	76,38	75,27	75,82
10,0,1,2,13,22,30,35,39,43	76,36	75,24	75,80
11,0,1,2,13,22,30,35,39,43	76,36	75,25	75,80
12,0,1,2,13,22,30,35,39,43	76,35	75,25	75,80
14,0,1,2,13,22,30,35,39,43	76,55	75,45	76,00
15,0,1,2,13,22,30,35,39,43	76,07	74,97	75,52
16,0,1,2,13,22,30,35,39,43	76,60	75,49	76,04
17,0,1,2,13,22,30,35,39,43	76,35	75,26	75,80
18,0,1,2,13,22,30,35,39,43	76,40	75,29	75,84
19,0,1,2,13,22,30,35,39,43	76,56	75,45	76,00
20,0,1,2,13,22,30,35,39,43	76,63	75,52	76,07
21,0,1,2,13,22,30,35,39,43	76,54	75,43	75,98
23,0,1,2,13,22,30,35,39,43	76,27	75,17	75,72
24,0,1,2,13,22,30,35,39,43	76,07	74,98	75,52
25,0,1,2,13,22,30,35,39,43	76,57	75,46	76,02
26,0,1,2,13,22,30,35,39,43	76,57	75,46	76,02
27,0,1,2,13,22,30,35,39,43	76,57	75,46	76,02
28,0,1,2,13,22,30,35,39,43	76,60	75,49	76,04
29,0,1,2,13,22,30,35,39,43	76,00	67,98	71,41
31,0,1,2,13,22,30,35,39,43	76,22	68,16	71,61
32,0,1,2,13,22,30,35,39,43	76,54	75,43	75,99
33,0,1,2,13,22,30,35,39,43	76,53	75,42	75,97
34,0,1,2,13,22,30,35,39,43	76,04	74,93	75,49
36,0,1,2,13,22,30,35,39,43	76,56	75,45	76,00
37,0,1,2,13,22,30,35,39,43	76,57	75,46	76,01
38,0,1,2,13,22,30,35,39,43	76,56	75,45	76,00
40,0,1,2,13,22,30,35,39,43	76,37	75,24	75,80
41,0,1,2,13,22,30,35,39,43	76,52	75,39	75,95
42,0,1,2,13,22,30,35,39,43	76,38	75,25	75,81
44,0,1,2,13,22,30,35,39,43	76,39	75,26	75,82
45,0,1,2,13,22,30,35,39,43	**76,71**	**75,58**	**76,14**

Cuadro 8.15. Combinaciones con 10 características. TiMBL. Anotación por sentidos (5/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,14,18,35,39,43,45	76,52	75,35	75,93
4,0,1,2,13,14,18,35,39,43,45	76,62	75,47	76,04
5,0,1,2,13,14,18,35,39,43,45	76,37	75,22	75,79
6,0,1,2,13,14,18,35,39,43,45	76,42	75,25	75,83
7,0,1,2,13,14,18,35,39,43,45	76,55	75,40	75,98
8,0,1,2,13,14,18,35,39,43,45	76,36	75,21	75,78
9,0,1,2,13,14,18,35,39,43,45	76,58	75,44	76,01
10,0,1,2,13,14,18,35,39,43,45	76,56	75,42	75,99
11,0,1,2,13,14,18,35,39,43,45	76,61	75,48	76,04
12,0,1,2,13,14,18,35,39,43,45	76,60	75,47	76,04
15,0,1,2,13,14,18,35,39,43,45	76,51	75,37	75,94
16,0,1,2,13,14,18,35,39,43,45	76,77	75,63	76,19
17,0,1,2,13,14,18,35,39,43,45	76,53	75,40	75,96
19,0,1,2,13,14,18,35,39,43,45	76,77	75,62	76,19
20,0,1,2,13,14,18,35,39,43,45	**76,82**	**75,68**	**76,24**
21,0,1,2,13,14,18,35,39,43,45	76,77	75,63	76,19
22,0,1,2,13,14,18,35,39,43,45	76,74	75,60	76,16
23,0,1,2,13,14,18,35,39,43,45	76,69	75,56	76,12
24,0,1,2,13,14,18,35,39,43,45	76,51	75,38	75,94
25,0,1,2,13,14,18,35,39,43,45	76,71	75,57	76,14
26,0,1,2,13,14,18,35,39,43,45	76,71	75,57	76,14
27,0,1,2,13,14,18,35,39,43,45	76,71	75,57	76,14
28,0,1,2,13,14,18,35,39,43,45	76,77	75,63	76,19
29,0,1,2,13,14,18,35,39,43,45	76,53	75,40	75,96
30,0,1,2,13,14,18,35,39,43,45	76,65	75,51	76,07
31,0,1,2,13,14,18,35,39,43,45	76,44	68,35	71,81
32,0,1,2,13,14,18,35,39,43,45	76,76	75,62	76,18
33,0,1,2,13,14,18,35,39,43,45	76,73	75,60	76,16
34,0,1,2,13,14,18,35,39,43,45	76,36	75,23	75,79
36,0,1,2,13,14,18,35,39,43,45	76,79	75,65	76,21
37,0,1,2,13,14,18,35,39,43,45	76,80	75,67	76,23
38,0,1,2,13,14,18,35,39,43,45	76,79	75,65	76,22
40,0,1,2,13,14,18,35,39,43,45	76,18	75,04	75,61
41,0,1,2,13,14,18,35,39,43,45	76,04	74,90	75,46
42,0,1,2,13,14,18,35,39,43,45	76,21	75,07	75,64
44,0,1,2,13,14,18,35,39,43,45	76,18	75,04	75,61

Cuadro 8.16. Combinaciones con 11 características. TiMBL. Anotación por sentidos (1/4)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,22,35,39,43,45	76,53	75,36	75,94
4,0,1,2,13,18,22,35,39,43,45	76,63	75,48	76,05
5,0,1,2,13,18,22,35,39,43,45	76,37	75,22	75,80
6,0,1,2,13,18,22,35,39,43,45	76,42	75,26	75,83
7,0,1,2,13,18,22,35,39,43,45	76,56	75,41	75,98
8,0,1,2,13,18,22,35,39,43,45	76,37	75,22	75,79
9,0,1,2,13,18,22,35,39,43,45	76,58	75,45	76,01
10,0,1,2,13,18,22,35,39,43,45	76,56	75,43	75,99
11,0,1,2,13,18,22,35,39,43,45	76,62	75,49	76,05
12,0,1,2,13,18,22,35,39,43,45	76,61	75,48	76,04
14,0,1,2,13,18,22,35,39,43,45	76,74	75,60	76,16
15,0,1,2,13,18,22,35,39,43,45	76,51	75,38	75,94
16,0,1,2,13,18,22,35,39,43,45	76,77	75,64	76,20
17,0,1,2,13,18,22,35,39,43,45	76,53	75,40	75,97
19,0,1,2,13,18,22,35,39,43,45	76,77	75,63	76,20
20,0,1,2,13,18,22,35,39,43,45	**76,83**	**75,68**	**76,25**
21,0,1,2,13,18,22,35,39,43,45	76,77	75,63	76,19
23,0,1,2,13,18,22,35,39,43,45	76,69	75,57	76,13
24,0,1,2,13,18,22,35,39,43,45	76,52	75,38	75,95
25,0,1,2,13,18,22,35,39,43,45	76,71	75,58	76,14
26,0,1,2,13,18,22,35,39,43,45	76,71	75,58	76,14
27,0,1,2,13,18,22,35,39,43,45	76,71	75,58	76,14
28,0,1,2,13,18,22,35,39,43,45	76,77	75,64	76,20
29,0,1,2,13,18,22,35,39,43,45	76,20	68,14	71,59
30,0,1,2,13,18,22,35,39,43,45	76,65	75,52	76,08
31,0,1,2,13,18,22,35,39,43,45	76,45	68,35	71,82
32,0,1,2,13,18,22,35,39,43,45	76,76	75,63	76,19
33,0,1,2,13,18,22,35,39,43,45	76,73	75,59	76,15
34,0,1,2,13,18,22,35,39,43,45	76,37	75,23	75,80
36,0,1,2,13,18,22,35,39,43,45	76,79	75,66	76,22
37,0,1,2,13,18,22,35,39,43,45	76,80	75,67	76,23
38,0,1,2,13,18,22,35,39,43,45	76,79	75,66	76,22
40,0,1,2,13,18,22,35,39,43,45	76,19	75,05	75,61
41,0,1,2,13,18,22,35,39,43,45	76,05	74,90	75,47
42,0,1,2,13,18,22,35,39,43,45	76,22	75,07	75,64
44,0,1,2,13,18,22,35,39,43,45	76,19	75,04	75,61

Cuadro 8.17. Combinaciones con 11 características. TiMBL. Anotación por sentidos (2/4)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,33,35,39,43,45	76,52	75,36	75,93
4,0,1,2,13,18,33,35,39,43,45	76,62	75,47	76,04
5,0,1,2,13,18,33,35,39,43,45	76,37	75,22	75,79
6,0,1,2,13,18,33,35,39,43,45	76,42	75,25	75,83
7,0,1,2,13,18,33,35,39,43,45	76,56	75,41	75,98
8,0,1,2,13,18,33,35,39,43,45	76,37	75,21	75,79
9,0,1,2,13,18,33,35,39,43,45	76,58	75,45	76,01
10,0,1,2,13,18,33,35,39,43,45	76,56	75,43	75,99
11,0,1,2,13,18,33,35,39,43,45	76,61	75,49	76,04
12,0,1,2,13,18,33,35,39,43,45	76,60	75,48	76,04
14,0,1,2,13,18,33,35,39,43,45	76,73	75,60	76,16
15,0,1,2,13,18,33,35,39,43,45	76,51	75,38	75,94
16,0,1,2,13,18,33,35,39,43,45	76,77	75,63	76,20
17,0,1,2,13,18,33,35,39,43,45	76,53	75,40	75,96
19,0,1,2,13,18,33,35,39,43,45	76,77	75,63	76,19
20,0,1,2,13,18,33,35,39,43,45	**76,82**	**75,68**	**76,25**
21,0,1,2,13,18,33,35,39,43,45	76,77	75,63	76,19
22,0,1,2,13,18,33,35,39,43,45	76,73	75,59	76,15
23,0,1,2,13,18,33,35,39,43,45	76,69	75,56	76,12
24,0,1,2,13,18,33,35,39,43,45	76,51	75,38	75,94
25,0,1,2,13,18,33,35,39,43,45	76,71	75,58	76,14
26,0,1,2,13,18,33,35,39,43,45	76,71	75,58	76,14
27,0,1,2,13,18,33,35,39,43,45	76,71	75,58	76,14
28,0,1,2,13,18,33,35,39,43,45	76,77	75,63	76,20
29,0,1,2,13,18,33,35,39,43,45	76,19	68,14	71,59
30,0,1,2,13,18,33,35,39,43,45	76,65	75,52	76,07
31,0,1,2,13,18,33,35,39,43,45	76,45	68,35	71,81
32,0,1,2,13,18,33,35,39,43,45	76,76	75,62	76,18
34,0,1,2,13,18,33,35,39,43,45	76,36	75,23	75,79
36,0,1,2,13,18,33,35,39,43,45	76,79	75,65	76,21
37,0,1,2,13,18,33,35,39,43,45	76,80	75,67	76,23
38,0,1,2,13,18,33,35,39,43,45	76,79	75,65	76,22
40,0,1,2,13,18,33,35,39,43,45	76,18	75,04	75,61
41,0,1,2,13,18,33,35,39,43,45	76,04	74,90	75,47
42,0,1,2,13,18,33,35,39,43,45	76,21	75,07	75,64
44,0,1,2,13,18,33,35,39,43,45	76,19	75,04	75,61

Cuadro 8.18. Combinaciones con 11 características. TiMBL. Anotación por sentidos (3/4)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,22,30,35,39,43,45	76,53	75,36	75,94
4,0,1,2,13,22,30,35,39,43,45	76,63	75,47	76,05
5,0,1,2,13,22,30,35,39,43,45	76,37	75,22	75,79
6,0,1,2,13,22,30,35,39,43,45	76,42	75,25	75,83
7,0,1,2,13,22,30,35,39,43,45	76,56	75,41	75,98
8,0,1,2,13,22,30,35,39,43,45	76,37	75,22	75,79
9,0,1,2,13,22,30,35,39,43,45	76,58	75,45	76,01
10,0,1,2,13,22,30,35,39,43,45	76,56	75,42	75,99
11,0,1,2,13,22,30,35,39,43,45	76,61	75,49	76,05
12,0,1,2,13,22,30,35,39,43,45	76,60	75,48	76,04
14,0,1,2,13,22,30,35,39,43,45	76,73	75,60	76,16
15,0,1,2,13,22,30,35,39,43,45	76,51	75,38	75,94
16,0,1,2,13,22,30,35,39,43,45	76,77	75,63	76,20
17,0,1,2,13,22,30,35,39,43,45	76,53	75,40	75,96
18,0,1,2,13,22,30,35,39,43,45	76,65	75,52	76,08
19,0,1,2,13,22,30,35,39,43,45	76,77	75,63	76,19
20,0,1,2,13,22,30,35,39,43,45	**76,82**	**75,68**	**76,25**
21,0,1,2,13,22,30,35,39,43,45	76,77	75,63	76,19
23,0,1,2,13,22,30,35,39,43,45	76,69	75,56	76,12
24,0,1,2,13,22,30,35,39,43,45	76,51	75,38	75,94
25,0,1,2,13,22,30,35,39,43,45	76,71	75,58	76,14
26,0,1,2,13,22,30,35,39,43,45	76,71	75,58	76,14
27,0,1,2,13,22,30,35,39,43,45	76,71	75,58	76,14
28,0,1,2,13,22,30,35,39,43,45	76,77	75,63	76,20
29,0,1,2,13,22,30,35,39,43,45	76,20	68,14	71,59
30,0,1,2,13,22,30,35,39,43,45	76,45	68,35	71,81
31,0,1,2,13,22,30,35,39,43,45	76,76	75,62	76,19
32,0,1,2,13,22,30,35,39,43,45	76,72	75,59	76,15
34,0,1,2,13,22,30,35,39,43,45	76,37	75,23	75,80
36,0,1,2,13,22,30,35,39,43,45	76,79	75,66	76,22
37,0,1,2,13,22,30,35,39,43,45	76,80	75,67	76,23
38,0,1,2,13,22,30,35,39,43,45	76,79	75,65	76,22
40,0,1,2,13,22,30,35,39,43,45	76,19	75,04	75,61
41,0,1,2,13,22,30,35,39,43,45	76,05	74,90	75,47
42,0,1,2,13,22,30,35,39,43,45	76,21	75,07	75,64
44,0,1,2,13,22,30,35,39,43,45	76,19	75,04	75,61

Cuadro 8.19. Combinaciones con 11 características. TiMBL. Anotación por sentidos (4/4)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,20,22,35,39,43,45	76,67	75,50	76,08
4,0,1,2,13,18,20,22,35,39,43,45	76,73	75,58	76,15
5,0,1,2,13,18,20,22,35,39,43,45	76,50	75,34	75,91
6,0,1,2,13,18,20,22,35,39,43,45	76,61	75,44	76,02
7,0,1,2,13,18,20,22,35,39,43,45	76,69	75,54	76,11
8,0,1,2,13,18,20,22,35,39,43,45	76,50	75,34	75,91
9,0,1,2,13,18,20,22,35,39,43,45	76,71	75,58	76,14
10,0,1,2,13,18,20,22,35,39,43,45	76,69	75,55	76,12
11,0,1,2,13,18,20,22,35,39,43,45	76,68	75,55	76,11
12,0,1,2,13,18,20,22,35,39,43,45	76,67	75,54	76,10
14,0,1,2,13,18,20,22,35,39,43,45	76,84	75,70	76,26
15,0,1,2,13,18,20,22,35,39,43,45	76,64	75,50	76,07
16,0,1,2,13,18,20,22,35,39,43,45	76,87	75,73	76,30
17,0,1,2,13,18,20,22,35,39,43,45	76,65	75,52	76,08
19,0,1,2,13,18,20,22,35,39,43,45	76,88	75,74	76,31
21,0,1,2,13,18,20,22,35,39,43,45	76,88	75,74	76,31
23,0,1,2,13,18,20,22,35,39,43,45	76,78	75,65	76,21
24,0,1,2,13,18,20,22,35,39,43,45	76,65	75,52	76,08
25,0,1,2,13,18,20,22,35,39,43,45	76,83	75,68	76,25
26,0,1,2,13,18,20,22,35,39,43,45	76,83	75,68	76,25
27,0,1,2,13,18,20,22,35,39,43,45	76,83	75,68	76,25
28,0,1,2,13,18,20,22,35,39,43,45	76,87	75,73	76,30
29,0,1,2,13,18,20,22,35,39,43,45	76,29	68,23	71,68
30,0,1,2,13,18,20,22,35,39,43,45	76,77	75,64	76,20
31,0,1,2,13,18,20,22,35,39,43,45	76,55	68,44	71,91
32,0,1,2,13,18,20,22,35,39,43,45	76,87	75,73	76,30
33,0,1,2,13,18,20,22,35,39,43,45	76,83	75,69	76,25
34,0,1,2,13,18,20,22,35,39,43,45	76,46	75,33	75,89
36,0,1,2,13,18,20,22,35,39,43,45	76,90	75,76	76,33
37,0,1,2,13,18,20,22,35,39,43,45	**76,91**	**75,78**	**76,34**
38,0,1,2,13,18,20,22,35,39,43,45	76,90	75,77	76,33
40,0,1,2,13,18,20,22,35,39,43,45	76,44	75,29	75,86
41,0,1,2,13,18,20,22,35,39,43,45	76,30	75,15	75,72
42,0,1,2,13,18,20,22,35,39,43,45	76,47	75,33	75,90
44,0,1,2,13,18,20,22,35,39,43,45	76,45	75,30	75,87

Cuadro 8.20. Combinaciones con 12 características. TiMBL. Anotación por sentidos (1/3)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,20,33,35,39,43,45	76,66	75,50	76,08
4,0,1,2,13,18,20,33,35,39,43,45	76,73	75,58	76,15
5,0,1,2,13,18,20,33,35,39,43,45	76,49	75,34	75,91
6,0,1,2,13,18,20,33,35,39,43,45	76,60	75,43	76,02
7,0,1,2,13,18,20,33,35,39,43,45	76,69	75,54	76,11
8,0,1,2,13,18,20,33,35,39,43,45	76,49	75,34	75,91
9,0,1,2,13,18,20,33,35,39,43,45	76,71	75,57	76,14
10,0,1,2,13,18,20,33,35,39,43,45	76,69	75,55	76,12
11,0,1,2,13,18,20,33,35,39,43,45	76,67	75,55	76,11
12,0,1,2,13,18,20,33,35,39,43,45	76,66	75,54	76,10
14,0,1,2,13,18,20,33,35,39,43,45	76,83	75,70	76,26
15,0,1,2,13,18,20,33,35,39,43,45	76,63	75,50	76,06
16,0,1,2,13,18,20,33,35,39,43,45	76,87	75,73	76,29
17,0,1,2,13,18,20,33,35,39,43,45	76,64	75,51	76,08
19,0,1,2,13,18,20,33,35,39,43,45	76,87	75,74	76,30
21,0,1,2,13,18,20,33,35,39,43,45	76,87	75,73	76,30
23,0,1,2,13,18,20,33,35,39,43,45	76,83	75,69	76,25
24,0,1,2,13,18,20,33,35,39,43,45	76,78	75,65	76,21
25,0,1,2,13,18,20,33,35,39,43,45	76,65	75,51	76,08
26,0,1,2,13,18,20,33,35,39,43,45	76,82	75,68	76,25
27,0,1,2,13,18,20,33,35,39,43,45	76,82	75,68	76,25
28,0,1,2,13,18,20,33,35,39,43,45	76,82	75,68	76,25
29,0,1,2,13,18,20,33,35,39,43,45	76,87	75,73	76,29
30,0,1,2,13,18,20,33,35,39,43,45	76,46	75,24	75,85
31,0,1,2,13,18,20,33,35,39,43,45	76,77	75,63	76,20
32,0,1,2,13,18,20,33,35,39,43,45	76,54	68,43	71,90
33,0,1,2,13,18,20,33,35,39,43,45	76,87	75,73	76,30
34,0,1,2,13,18,20,33,35,39,43,45	76,46	75,32	75,89
36,0,1,2,13,18,20,33,35,39,43,45	76,89	75,76	76,32
37,0,1,2,13,18,20,33,35,39,43,45	**76,91**	**75,77**	**76,34**
38,0,1,2,13,18,20,33,35,39,43,45	76,89	75,76	76,32
40,0,1,2,13,18,20,33,35,39,43,45	76,43	75,29	75,86
41,0,1,2,13,18,20,33,35,39,43,45	76,30	75,15	75,72
42,0,1,2,13,18,20,33,35,39,43,45	76,47	75,33	75,89
44,0,1,2,13,18,20,33,35,39,43,45	76,45	75,29	75,86

Cuadro 8.21. Combinaciones con 12 características. TiMBL. Anotación por sentidos (2/3)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,20,22,30,35,39,43,45	76,67	75,50	76,08
4,0,1,2,13,20,22,30,35,39,43,45	76,73	75,58	76,15
5,0,1,2,13,20,22,30,35,39,43,45	76,50	75,34	75,91
6,0,1,2,13,20,22,30,35,39,43,45	76,61	75,44	76,02
7,0,1,2,13,20,22,30,35,39,43,45	76,69	75,54	76,11
8,0,1,2,13,20,22,30,35,39,43,45	76,49	75,34	75,91
9,0,1,2,13,20,22,30,35,39,43,45	76,71	75,58	76,14
10,0,1,2,13,20,22,30,35,39,43,45	76,69	75,55	76,12
11,0,1,2,13,20,22,30,35,39,43,45	76,68	75,55	76,11
12,0,1,2,13,20,22,30,35,39,43,45	76,66	75,54	76,10
14,0,1,2,13,20,22,30,35,39,43,45	76,84	75,69	76,26
15,0,1,2,13,20,22,30,35,39,43,45	76,63	75,50	76,06
16,0,1,2,13,20,22,30,35,39,43,45	76,87	75,73	76,29
17,0,1,2,13,20,22,30,35,39,43,45	76,64	75,52	76,08
18,0,1,2,13,20,22,30,35,39,43,45	76,77	75,64	76,20
19,0,1,2,13,20,22,30,35,39,43,45	76,88	75,74	76,31
21,0,1,2,13,20,22,30,35,39,43,45	76,88	75,74	76,31
23,0,1,2,13,20,22,30,35,39,43,45	76,78	75,65	76,21
24,0,1,2,13,20,22,30,35,39,43,45	76,65	75,51	76,08
25,0,1,2,13,20,22,30,35,39,43,45	76,82	75,68	76,25
26,0,1,2,13,20,22,30,35,39,43,45	76,82	75,68	76,25
27,0,1,2,13,20,22,30,35,39,43,45	76,82	75,68	76,25
28,0,1,2,13,20,22,30,35,39,43,45	76,87	75,73	76,29
29,0,1,2,13,20,22,30,35,39,43,45	76,29	68,22	71,68
31,0,1,2,13,20,22,30,35,39,43,45	76,55	68,44	71,90
32,0,1,2,13,20,22,30,35,39,43,45	76,87	75,73	76,30
33,0,1,2,13,20,22,30,35,39,43,45	76,83	75,69	76,25
34,0,1,2,13,20,22,30,35,39,43,45	76,46	75,33	75,89
36,0,1,2,13,20,22,30,35,39,43,45	76,89	75,76	76,32
37,0,1,2,13,20,22,30,35,39,43,45	**76,91**	**75,78**	**76,34**
38,0,1,2,13,20,22,30,35,39,43,45	76,90	75,77	76,33
40,0,1,2,13,20,22,30,35,39,43,45	76,44	75,29	75,86
41,0,1,2,13,20,22,30,35,39,43,45	76,30	75,15	75,72
42,0,1,2,13,20,22,30,35,39,43,45	76,47	75,33	75,90
44,0,1,2,13,20,22,30,35,39,43,45	76,44	75,29	75,87

Cuadro 8.22. Combinaciones con 12 características. TiMBL. Anotación por sentidos (3/3)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,20,22,35,37,39,43,45	76,78	75,61	76,19
4,0,1,2,13,18,20,22,35,37,39,43,45	76,86	75,71	76,28
5,0,1,2,13,18,20,22,35,37,39,43,45	76,59	75,44	76,01
6,0,1,2,13,18,20,22,35,37,39,43,45	76,73	75,56	76,14
7,0,1,2,13,18,20,22,35,37,39,43,45	76,83	75,68	76,25
8,0,1,2,13,18,20,22,35,37,39,43,45	76,58	75,44	76,01
9,0,1,2,13,18,20,22,35,37,39,43,45	76,80	75,67	76,24
10,0,1,2,13,18,20,22,35,37,39,43,45	76,79	75,65	76,21
11,0,1,2,13,18,20,22,35,37,39,43,45	76,78	75,65	76,21
12,0,1,2,13,18,20,22,35,37,39,43,45	76,77	75,64	76,20
14,0,1,2,13,18,20,22,35,37,39,43,45	76,88	75,75	76,31
15,0,1,2,13,18,20,22,35,37,39,43,45	76,72	75,59	76,15
16,0,1,2,13,18,20,22,35,37,39,43,45	**76,91**	**75,77**	**76,34**
17,0,1,2,13,18,20,22,35,37,39,43,45	76,73	75,61	76,16
19,0,1,2,13,18,20,22,35,37,39,43,45	76,89	75,75	76,32
21,0,1,2,13,18,20,22,35,37,39,43,45	76,89	75,75	76,31
23,0,1,2,13,18,20,22,35,37,39,43,45	76,85	75,72	76,28
24,0,1,2,13,18,20,22,35,37,39,43,45	76,73	75,60	76,16
25,0,1,2,13,18,20,22,35,37,39,43,45	**76,91**	**75,78**	**76,34**
26,0,1,2,13,18,20,22,35,37,39,43,45	**76,91**	**75,78**	**76,34**
27,0,1,2,13,18,20,22,35,37,39,43,45	**76,91**	**75,78**	**76,34**
28,0,1,2,13,18,20,22,35,37,39,43,45	**76,91**	**75,77**	**76,34**
29,0,1,2,13,18,20,22,35,37,39,43,45	76,37	68,29	71,74
30,0,1,2,13,18,20,22,35,37,39,43,45	76,82	75,69	76,26
31,0,1,2,13,18,20,22,35,37,39,43,45	76,56	68,45	71,92
32,0,1,2,13,18,20,22,35,37,39,43,45	76,89	75,75	76,31
33,0,1,2,13,18,20,22,35,37,39,43,45	76,87	75,73	76,30
34,0,1,2,13,18,20,22,35,37,39,43,45	76,54	75,41	75,97
36,0,1,2,13,18,20,22,35,37,39,43,45	76,89	75,76	76,32
38,0,1,2,13,18,20,22,35,37,39,43,45	76,89	75,75	76,32
40,0,1,2,13,18,20,22,35,37,39,43,45	76,60	75,45	76,02
41,0,1,2,13,18,20,22,35,37,39,43,45	76,50	75,35	75,92
42,0,1,2,13,18,20,22,35,37,39,43,45	76,64	75,49	76,06
44,0,1,2,13,18,20,22,35,37,39,43,45	76,61	75,46	76,03

Cuadro 8.23. Combinaciones con 13 características. TiMBL. Anotación por sentidos (1/3)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,20,33,35,37,39,43,45	76,77	75,61	76,19
4,0,1,2,13,18,20,33,35,37,39,43,45	76,85	75,70	76,27
5,0,1,2,13,18,20,33,35,37,39,43,45	76,58	75,43	76,00
6,0,1,2,13,18,20,33,35,37,39,43,45	76,73	75,56	76,14
7,0,1,2,13,18,20,33,35,37,39,43,45	76,82	75,67	76,24
8,0,1,2,13,18,20,33,35,37,39,43,45	76,58	75,43	76,00
9,0,1,2,13,18,20,33,35,37,39,43,45	76,80	75,67	76,23
10,0,1,2,13,18,20,33,35,37,39,43,45	76,78	75,65	76,21
11,0,1,2,13,18,20,33,35,37,39,43,45	76,77	75,65	76,21
12,0,1,2,13,18,20,33,35,37,39,43,45	76,76	75,64	76,20
14,0,1,2,13,18,20,33,35,37,39,43,45	76,87	75,74	76,30
15,0,1,2,13,18,20,33,35,37,39,43,45	76,71	75,59	76,15
16,0,1,2,13,18,20,33,35,37,39,43,45	76,90	75,77	76,33
17,0,1,2,13,18,20,33,35,37,39,43,45	76,72	75,60	76,16
19,0,1,2,13,18,20,33,35,37,39,43,45	76,88	75,75	76,31
21,0,1,2,13,18,20,33,35,37,39,43,45	76,88	75,75	76,31
22,0,1,2,13,18,20,33,35,37,39,43,45	76,87	75,73	76,30
23,0,1,2,13,18,20,33,35,37,39,43,45	76,84	75,71	76,27
24,0,1,2,13,18,20,33,35,37,39,43,45	76,72	75,60	76,16
25,0,1,2,13,18,20,33,35,37,39,43,45	**76,91**	**75,77**	**76,34**
26,0,1,2,13,18,20,33,35,37,39,43,45	**76,91**	**75,77**	**76,34**
27,0,1,2,13,18,20,33,35,37,39,43,45	**76,91**	**75,77**	**76,34**
28,0,1,2,13,18,20,33,35,37,39,43,45	76,90	75,77	76,33
29,0,1,2,13,18,20,33,35,37,39,43,45	76,72	75,60	76,16
30,0,1,2,13,18,20,33,35,37,39,43,45	76,82	75,69	76,25
31,0,1,2,13,18,20,33,35,37,39,43,45	76,55	68,45	71,92
32,0,1,2,13,18,20,33,35,37,39,43,45	76,88	75,74	76,31
34,0,1,2,13,18,20,33,35,37,39,43,45	76,53	75,40	75,96
36,0,1,2,13,18,20,33,35,37,39,43,45	76,88	75,75	76,32
38,0,1,2,13,18,20,33,35,37,39,43,45	76,88	75,75	76,31
40,0,1,2,13,18,20,33,35,37,39,43,45	76,60	75,45	76,02
41,0,1,2,13,18,20,33,35,37,39,43,45	76,49	75,34	75,92
42,0,1,2,13,18,20,33,35,37,39,43,45	76,64	75,49	76,06
44,0,1,2,13,18,20,33,35,37,39,43,45	76,60	75,45	76,02

Cuadro 8.24. Combinaciones con 13 características. TiMBL. Anotación por sentidos (2/3)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,20,22,30,35,37,39,43,45	76,78	75,61	76,19
4,0,1,2,13,20,22,30,35,37,39,43,45	76,86	75,71	76,28
5,0,1,2,13,20,22,30,35,37,39,43,45	76,59	75,44	76,01
6,0,1,2,13,20,22,30,35,37,39,43,45	76,73	75,56	76,14
7,0,1,2,13,20,22,30,35,37,39,43,45	76,83	75,68	76,25
8,0,1,2,13,20,22,30,35,37,39,43,45	76,58	75,43	76,00
9,0,1,2,13,20,22,30,35,37,39,43,45	76,80	75,67	76,23
10,0,1,2,13,20,22,30,35,37,39,43,45	76,79	75,65	76,21
11,0,1,2,13,20,22,30,35,37,39,43,45	76,78	75,65	76,21
12,0,1,2,13,20,22,30,35,37,39,43,45	76,77	75,64	76,20
14,0,1,2,13,20,22,30,35,37,39,43,45	76,88	75,75	76,31
15,0,1,2,13,20,22,30,35,37,39,43,45	76,72	75,59	76,15
16,0,1,2,13,20,22,30,35,37,39,43,45	**76,91**	**75,77**	**76,34**
17,0,1,2,13,20,22,30,35,37,39,43,45	76,73	75,60	76,16
18,0,1,2,13,20,22,30,35,37,39,43,45	76,82	75,69	76,26
19,0,1,2,13,20,22,30,35,37,39,43,45	76,89	75,75	76,32
21,0,1,2,13,20,22,30,35,37,39,43,45	76,88	75,75	76,31
23,0,1,2,13,20,22,30,35,37,39,43,45	76,84	75,72	76,28
24,0,1,2,13,20,22,30,35,37,39,43,45	76,73	75,60	76,16
25,0,1,2,13,20,22,30,35,37,39,43,45	**76,91**	**75,78**	**76,34**
26,0,1,2,13,20,22,30,35,37,39,43,45	**76,91**	**75,78**	**76,34**
27,0,1,2,13,20,22,30,35,37,39,43,45	**76,91**	**75,78**	**76,34**
28,0,1,2,13,20,22,30,35,37,39,43,45	**76,91**	**75,77**	**76,34**
29,0,1,2,13,20,22,30,35,37,39,43,45	76,37	68,29	71,74
31,0,1,2,13,20,22,30,35,37,39,43,45	76,56	68,49	71,95
32,0,1,2,13,20,22,30,35,37,39,43,45	76,88	75,75	76,31
33,0,1,2,13,20,22,30,35,37,39,43,45	76,87	75,73	76,30
34,0,1,2,13,20,22,30,35,37,39,43,45	76,54	75,41	75,97
36,0,1,2,13,20,22,30,35,37,39,43,45	76,89	75,76	76,32
38,0,1,2,13,20,22,30,35,37,39,43,45	76,89	75,75	76,32
40,0,1,2,13,20,22,30,35,37,39,43,45	76,60	75,45	76,02
41,0,1,2,13,20,22,30,35,37,39,43,45	76,50	75,35	75,92
42,0,1,2,13,20,22,30,35,37,39,43,45	76,64	75,49	76,06
44,0,1,2,13,20,22,30,35,37,39,43,45	76,61	75,46	76,03

Cuadro 8.25. Combinaciones con 13 características. TiMBL. Anotación por sentidos (3/3)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,16,18,20,22,35,37,39,43,45	76,82	75,64	76,23
4,0,1,2,13,16,18,20,22,35,37,39,43,45	76,86	75,71	76,28
5,0,1,2,13,16,18,20,22,35,37,39,43,45	76,59	75,44	76,01
6,0,1,2,13,16,18,20,22,35,37,39,43,45	76,74	75,57	76,15
7,0,1,2,13,16,18,20,22,35,37,39,43,45	76,86	75,70	76,27
8,0,1,2,13,16,18,20,22,35,37,39,43,45	76,58	75,44	76,00
9,0,1,2,13,16,18,20,22,35,37,39,43,45	76,80	75,66	76,22
10,0,1,2,13,16,18,20,22,35,37,39,43,45	76,78	75,64	76,20
11,0,1,2,13,16,18,20,22,35,37,39,43,45	76,78	75,65	76,21
12,0,1,2,13,16,18,20,22,35,37,39,43,45	76,77	75,64	76,20
14,0,1,2,13,16,18,20,22,35,37,39,43,45	76,88	75,75	76,31
15,0,1,2,13,16,18,20,22,35,37,39,43,45	76,72	75,59	76,15
17,0,1,2,13,16,18,20,22,35,37,39,43,45	76,69	75,57	76,13
19,0,1,2,13,16,18,20,22,35,37,39,43,45	76,90	75,77	76,33
21,0,1,2,13,16,18,20,22,35,37,39,43,45	76,90	75,77	76,33
23,0,1,2,13,16,18,20,22,35,37,39,43,45	76,82	75,69	76,25
24,0,1,2,13,16,18,20,22,35,37,39,43,45	76,72	75,59	76,15
25,0,1,2,13,16,18,20,22,35,37,39,43,45	**76,91**	**75,77**	**76,34**
26,0,1,2,13,16,18,20,22,35,37,39,43,45	**76,91**	**75,77**	**76,34**
27,0,1,2,13,16,18,20,22,35,37,39,43,45	**76,91**	**75,77**	**76,34**
28,0,1,2,13,16,18,20,22,35,37,39,43,45	76,88	75,75	76,31
29,0,1,2,13,16,18,20,22,35,37,39,43,45	76,34	68,27	71,72
30,0,1,2,13,16,18,20,22,35,37,39,43,45	76,84	75,70	76,26
31,0,1,2,13,16,18,20,22,35,37,39,43,45	76,58	68,47	71,94
32,0,1,2,13,16,18,20,22,35,37,39,43,45	76,90	75,76	76,33
33,0,1,2,13,16,18,20,22,35,37,39,43,45	76,86	75,73	76,29
34,0,1,2,13,16,18,20,22,35,37,39,43,45	76,56	75,43	75,99
36,0,1,2,13,16,18,20,22,35,37,39,43,45	76,89	75,76	76,32
38,0,1,2,13,16,18,20,22,35,37,39,43,45	76,89	75,75	76,32
40,0,1,2,13,16,18,20,22,35,37,39,43,45	76,65	75,49	76,07
41,0,1,2,13,16,18,20,22,35,37,39,43,45	76,54	75,38	75,96
42,0,1,2,13,16,18,20,22,35,37,39,43,45	76,68	75,53	76,10
44,0,1,2,13,16,18,20,22,35,37,39,43,45	76,65	75,50	76,07

Cuadro 8.26. Combinaciones con 14 características. TiMBL. Anotación por sentidos (1/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,20,22,25,35,37,39,43,45	76,78	75,61	76,19
4,0,1,2,13,18,20,22,25,35,37,39,43,45	76,86	75,71	76,28
5,0,1,2,13,18,20,22,25,35,37,39,43,45	76,59	75,44	76,01
6,0,1,2,13,18,20,22,25,35,37,39,43,45	76,73	75,56	76,14
7,0,1,2,13,18,20,22,25,35,37,39,43,45	76,83	75,68	76,25
8,0,1,2,13,18,20,22,25,35,37,39,43,45	76,58	75,44	76,01
9,0,1,2,13,18,20,22,25,35,37,39,43,45	76,80	75,67	76,24
10,0,1,2,13,18,20,22,25,35,37,39,43,45	76,79	75,65	76,21
11,0,1,2,13,18,20,22,25,35,37,39,43,45	76,78	75,65	76,21
12,0,1,2,13,18,20,22,25,35,37,39,43,45	76,77	75,64	76,20
14,0,1,2,13,18,20,22,25,35,37,39,43,45	76,88	75,75	76,31
15,0,1,2,13,18,20,22,25,35,37,39,43,45	76,72	75,59	76,15
16,0,1,2,13,18,20,22,25,35,37,39,43,45	76,34	43,60	52,11
17,0,1,2,13,18,20,22,25,35,37,39,43,45	76,03	56,79	60,87
19,0,1,2,13,18,20,22,25,35,37,39,43,45	76,89	75,75	76,32
21,0,1,2,13,18,20,22,25,35,37,39,43,45	76,89	75,75	76,31
23,0,1,2,13,18,20,22,25,35,37,39,43,45	76,85	75,72	76,28
24,0,1,2,13,18,20,22,25,35,37,39,43,45	76,73	75,60	76,16
26,0,1,2,13,18,20,22,25,35,37,39,43,45	**76,91**	**75,78**	**76,34**
27,0,1,2,13,18,20,22,25,35,37,39,43,45	**76,91**	**75,78**	**76,34**
28,0,1,2,13,18,20,22,25,35,37,39,43,45	**76,91**	**75,77**	**76,34**
29,0,1,2,13,18,20,22,25,35,37,39,43,45	76,37	68,29	71,74
30,0,1,2,13,18,20,22,25,35,37,39,43,45	76,82	75,69	76,26
31,0,1,2,13,18,20,22,25,35,37,39,43,45	76,74	68,72	72,15
32,0,1,2,13,18,20,22,25,35,37,39,43,45	76,88	75,74	76,31
33,0,1,2,13,18,20,22,25,35,37,39,43,45	76,76	75,63	76,19
34,0,1,2,13,18,20,22,25,35,37,39,43,45	76,65	75,51	76,08
36,0,1,2,13,18,20,22,25,35,37,39,43,45	76,89	75,76	76,32
38,0,1,2,13,18,20,22,25,35,37,39,43,45	76,79	75,65	76,21
40,0,1,2,13,18,20,22,25,35,37,39,43,45	76,57	75,42	75,99
41,0,1,2,13,18,20,22,25,35,37,39,43,45	76,54	75,39	75,96
42,0,1,2,13,18,20,22,25,35,37,39,43,45	76,63	75,48	76,05
44,0,1,2,13,18,20,22,25,35,37,39,43,45	76,88	75,74	76,30

Cuadro 8.27. Combinaciones con 14 características. TiMBL. Anotación por sentidos (2/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,20,22,26,35,37,39,43,45	76,78	75,61	76,19
4,0,1,2,13,18,20,22,26,35,37,39,43,45	76,86	75,71	76,28
5,0,1,2,13,18,20,22,26,35,37,39,43,45	76,59	75,44	76,01
6,0,1,2,13,18,20,22,26,35,37,39,43,45	76,73	75,56	76,14
7,0,1,2,13,18,20,22,26,35,37,39,43,45	76,83	75,68	76,25
8,0,1,2,13,18,20,22,26,35,37,39,43,45	76,58	75,44	76,01
9,0,1,2,13,18,20,22,26,35,37,39,43,45	76,80	75,67	76,24
10,0,1,2,13,18,20,22,26,35,37,39,43,45	76,79	75,65	76,21
11,0,1,2,13,18,20,22,26,35,37,39,43,45	76,78	75,65	76,21
12,0,1,2,13,18,20,22,26,35,37,39,43,45	76,77	75,64	76,20
14,0,1,2,13,18,20,22,26,35,37,39,43,45	76,88	75,75	76,31
15,0,1,2,13,18,20,22,26,35,37,39,43,45	76,72	75,59	76,15
16,0,1,2,13,18,20,22,26,35,37,39,43,45	**76,91**	**75,77**	**76,34**
17,0,1,2,13,18,20,22,26,35,37,39,43,45	76,73	75,61	76,16
19,0,1,2,13,18,20,22,26,35,37,39,43,45	76,89	75,75	76,32
21,0,1,2,13,18,20,22,26,35,37,39,43,45	76,89	75,75	76,31
23,0,1,2,13,18,20,22,26,35,37,39,43,45	76,85	75,72	76,28
24,0,1,2,13,18,20,22,26,35,37,39,43,45	76,73	75,60	76,16
25,0,1,2,13,18,20,22,26,35,37,39,43,45	**76,91**	**75,78**	**76,34**
27,0,1,2,13,18,20,22,26,35,37,39,43,45	**76,91**	**75,78**	**76,34**
28,0,1,2,13,18,20,22,26,35,37,39,43,45	**76,91**	**75,77**	**76,34**
29,0,1,2,13,18,20,22,26,35,37,39,43,45	76,37	68,29	71,74
30,0,1,2,13,18,20,22,26,35,37,39,43,45	76,82	75,69	76,26
31,0,1,2,13,18,20,22,26,35,37,39,43,45	76,56	68,45	71,92
32,0,1,2,13,18,20,22,26,35,37,39,43,45	76,89	75,75	76,31
33,0,1,2,13,18,20,22,26,35,37,39,43,45	76,87	75,73	76,30
34,0,1,2,13,18,20,22,26,35,37,39,43,45	76,54	75,41	75,97
36,0,1,2,13,18,20,22,26,35,37,39,43,45	76,89	75,76	76,32
38,0,1,2,13,18,20,22,26,35,37,39,43,45	76,89	75,75	76,32
40,0,1,2,13,18,20,22,26,35,37,39,43,45	76,60	75,45	76,02
41,0,1,2,13,18,20,22,26,35,37,39,43,45	76,50	75,35	75,92
42,0,1,2,13,18,20,22,26,35,37,39,43,45	76,64	75,49	76,06
44,0,1,2,13,18,20,22,26,35,37,39,43,45	76,61	75,46	76,03

Cuadro 8.28. Combinaciones con 14 características. TiMBL. Anotación por sentidos (3/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,20,22,27,35,37,39,43,45	76,78	75,61	76,19
4,0,1,2,13,18,20,22,27,35,37,39,43,45	76,86	75,71	76,28
5,0,1,2,13,18,20,22,27,35,37,39,43,45	76,59	75,44	76,01
6,0,1,2,13,18,20,22,27,35,37,39,43,45	76,73	75,56	76,14
7,0,1,2,13,18,20,22,27,35,37,39,43,45	76,83	75,68	76,25
8,0,1,2,13,18,20,22,27,35,37,39,43,45	76,58	75,44	76,01
9,0,1,2,13,18,20,22,27,35,37,39,43,45	76,80	75,67	76,24
10,0,1,2,13,18,20,22,27,35,37,39,43,45	76,79	75,65	76,21
11,0,1,2,13,18,20,22,27,35,37,39,43,45	76,78	75,65	76,21
12,0,1,2,13,18,20,22,27,35,37,39,43,45	76,77	75,64	76,20
14,0,1,2,13,18,20,22,27,35,37,39,43,45	76,88	75,75	76,31
15,0,1,2,13,18,20,22,27,35,37,39,43,45	76,72	75,59	76,15
16,0,1,2,13,18,20,22,27,35,37,39,43,45	**76,91**	**75,77**	**76,34**
17,0,1,2,13,18,20,22,27,35,37,39,43,45	76,73	75,61	76,16
19,0,1,2,13,18,20,22,27,35,37,39,43,45	76,89	75,75	76,32
21,0,1,2,13,18,20,22,27,35,37,39,43,45	76,89	75,75	76,31
23,0,1,2,13,18,20,22,27,35,37,39,43,45	76,85	75,72	76,28
24,0,1,2,13,18,20,22,27,35,37,39,43,45	76,73	75,60	76,16
25,0,1,2,13,18,20,22,27,35,37,39,43,45	**76,91**	**75,78**	**76,34**
26,0,1,2,13,18,20,22,27,35,37,39,43,45	**76,91**	**75,78**	**76,34**
28,0,1,2,13,18,20,22,27,35,37,39,43,45	**76,91**	**75,77**	**76,34**
29,0,1,2,13,18,20,22,27,35,37,39,43,45	76,37	68,29	71,74
30,0,1,2,13,18,20,22,27,35,37,39,43,45	76,82	75,69	76,26
31,0,1,2,13,18,20,22,27,35,37,39,43,45	76,56	68,45	71,92
32,0,1,2,13,18,20,22,27,35,37,39,43,45	76,89	75,75	76,31
33,0,1,2,13,18,20,22,27,35,37,39,43,45	76,87	75,73	76,30
34,0,1,2,13,18,20,22,27,35,37,39,43,45	76,54	75,41	75,97
36,0,1,2,13,18,20,22,27,35,37,39,43,45	76,89	75,76	76,32
38,0,1,2,13,18,20,22,27,35,37,39,43,45	76,89	75,75	76,32
40,0,1,2,13,18,20,22,27,35,37,39,43,45	76,60	75,45	76,02
41,0,1,2,13,18,20,22,27,35,37,39,43,45	76,50	75,35	75,92
42,0,1,2,13,18,20,22,27,35,37,39,43,45	76,64	75,49	76,06
44,0,1,2,13,18,20,22,27,35,37,39,43,45	76,61	75,46	76,03

Cuadro 8.29. Combinaciones con 14 características. TiMBL. Anotación por sentidos (4/5)

Universitat d'Alacant
Universidad de Alicante

Características	P (%)	C (%)	$F_{\beta=1}$ (%)
3,0,1,2,13,18,20,22,28,35,37,39,43,45	76,82	75,64	76,23
4,0,1,2,13,18,20,22,28,35,37,39,43,45	76,86	75,71	76,28
5,0,1,2,13,18,20,22,28,35,37,39,43,45	76,59	75,44	76,01
6,0,1,2,13,18,20,22,28,35,37,39,43,45	76,74	75,57	76,15
7,0,1,2,13,18,20,22,28,35,37,39,43,45	76,86	75,70	76,27
8,0,1,2,13,18,20,22,28,35,37,39,43,45	76,58	75,44	76,00
9,0,1,2,13,18,20,22,28,35,37,39,43,45	76,80	75,66	76,22
10,0,1,2,13,18,20,22,28,35,37,39,43,45	76,78	75,64	76,20
11,0,1,2,13,18,20,22,28,35,37,39,43,45	76,78	75,65	76,21
12,0,1,2,13,18,20,22,28,35,37,39,43,45	76,77	75,64	76,20
14,0,1,2,13,18,20,22,28,35,37,39,43,45	76,88	75,75	76,31
15,0,1,2,13,18,20,22,28,35,37,39,43,45	76,72	75,59	76,15
16,0,1,2,13,18,20,22,28,35,37,39,43,45	76,88	75,75	76,31
17,0,1,2,13,18,20,22,28,35,37,39,43,45	76,69	75,57	76,13
19,0,1,2,13,18,20,22,28,35,37,39,43,45	76,90	75,77	76,33
21,0,1,2,13,18,20,22,28,35,37,39,43,45	76,90	75,77	76,33
23,0,1,2,13,18,20,22,28,35,37,39,43,45	76,82	75,69	76,25
24,0,1,2,13,18,20,22,28,35,37,39,43,45	76,72	75,59	76,15
25,0,1,2,13,18,20,22,28,35,37,39,43,45	**76,91**	**75,77**	**76,34**
26,0,1,2,13,18,20,22,28,35,37,39,43,45	**76,91**	**75,77**	**76,34**
27,0,1,2,13,18,20,22,28,35,37,39,43,45	**76,91**	**75,77**	**76,34**
29,0,1,2,13,18,20,22,28,35,37,39,43,45	76,34	68,27	71,72
30,0,1,2,13,18,20,22,28,35,37,39,43,45	76,84	75,70	76,26
31,0,1,2,13,18,20,22,28,35,37,39,43,45	76,58	68,47	71,94
32,0,1,2,13,18,20,22,28,35,37,39,43,45	76,90	75,76	76,33
33,0,1,2,13,18,20,22,28,35,37,39,43,45	76,86	75,73	76,29
34,0,1,2,13,18,20,22,28,35,37,39,43,45	76,56	75,43	75,99
36,0,1,2,13,18,20,22,28,35,37,39,43,45	76,89	75,76	76,32
38,0,1,2,13,18,20,22,28,35,37,39,43,45	76,89	75,75	76,32
40,0,1,2,13,18,20,22,28,35,37,39,43,45	76,65	75,49	76,06
41,0,1,2,13,18,20,22,28,35,37,39,43,45	76,54	75,38	75,96
42,0,1,2,13,18,20,22,28,35,37,39,43,45	76,68	75,53	76,10
44,0,1,2,13,18,20,22,28,35,37,39,43,45	76,65	75,50	76,07

Cuadro 8.30. Combinaciones con 14 características. TiMBL. Anotación por sentidos (5/5)

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Bibliografía

2000 (May). *The First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2000).*

2002 (July). *40th Annual Meeting of the Association for Computational Linguistics (ACL2002).*

2003 (July). *Conference on Empirical Methods in Natural Language Processing (EMNLP2003)).*

2003 (June). *Human Language Technology Conference (HLT-NAACL2003).*

2003 (September). *Recent Advances in Natural Language Processing (RANLP2003).*

2003 (May-June). *Seventh Conference on Natural Language Learning (CoNLL2003).*

2004 (July). *42nd Annual Meeting of the Association for Computational Linguistics (ACL2004).*

2004 (July). *Conference on Empirical Methods in Natural Language Processing (EMNLP2004)).*

2004 (May). *Eighth Conference on Natural Language Learning (CoNLL2004).*

2004 (May). *Fourth International Conference on Language Resources and Evaluation (LREC2004).*

2004 (May). *Human Language Technology Conference (HLT-NAACL2004).*

2004. *Proceedings of SENSEVAL-3.* Barcela, Spain: ACL-SIGLEX.

2004. *Proceedings of the International Conference Text Speech and Dialogue (TSD2004).* Lecture Notes in Artificial Intelligence. Brno, Czech Republic: Springer-Verlag.

2004 (August). *The 20th International Conference on Computational Linguistics (COLING2004).*

Universitat d'Alacant
Universidad de Alicante

2005 (June). *43rd Annual Meeting of the Association for Computational Linguistics (ACL2005).*

2005 (June). *Ninth Conference on Natural Language Learning (CoNLL2005).*

2006 (April). *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006).*

2006 (August). *The 5th International Conference on Natural Language Processing (FINTAL2006).*

2006 (November). *5th Mexican International Conference on Artificial Intelligence (MICAI2006).*

2006 (February). *7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006).*

2006 (May). *Fifth International Conference on Language Resources and Evaluation (LREC2006).*

2006 (May). *Human Language Technology Conference (HLT-NAACL2006).*

2006 (July). *The 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL2006).*

2007 (June). *Deep Linguistic Processing Workshop in 45th Annual Meeting of the Association for Computational Linguistics (ACL2007)).*

2007. *Proceedings of SemEval2007.* Prague, Czech Republic: Association for Computational Linguistics.

2007 (September). *Recent Advances in Natural Language Processing (RANLP2007).*

2008 (February). *9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008).*

Abney, S. 2002 (July). Bootstrapping. *In:* (acl, 2002).

Aduriz, I., Aranzabe, M., Arriola, J., Atutxa, A., de Ilarraza, A. Días, Garmendia, A., & Oronoz, M. 2003 (November). Construction of a Basque Dependency Treebank. *In: Proceedings of the Second Workshop on Treebanks and Linguistic Theories in TLT 2003.*

Agirre, E., Aldezabal, I., Etxeberria, J., & Pociello, E. 2006 (May). A Preliminary Study for Building the Basque PropBank. *In:* (lre, 2006).

Aha, D.A., & R.L.Bankert. 1994. Feature Selection for Case-Based Classification of Cloud Types: An Emprirical Comparison. *Pages 106–112 of: Working notes of the AAAI94 Workshop on Case-Based Reasoning.* Seattle, WA: AAAI

Universitat d'Alacant
Universidad de Alicante

Press.

Ahn, D., Fissaha, S., Jijkoun, V., & de Rijke, M. 2004. The University of Amsterdam at Senseval-3: Semantic Roles and Logic Forms. *In:* (sen, 2004).

Aldezabal, I. 2004 (April). *ADITZ-AZPIKATEGORIZAZIOAREN AZTERKETA SINTAXI PARTZIALETIK SINTAXI OSORAKO BIDEAN. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz.* Ph.D. thesis, Euskal Filologia Saila. Zientzia Fakultatea. Leioa. UPV/EHU.

Almuallim, H., & Dietterich, T.G. 1994. Learning Boolean Concepts in the Presence of Many Irrelevant Features. *Artificial Intelligence*, **69**(1-2), 279–305.

an dJ. Weston, I. Guyon, Barnhill, S., & Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**(1-3), 389–422.

Arcuri, A. 2006. *Semantic Role Labeling: un sistema a classificatori multipli basato su MaxEntropy.* M.Phil. thesis, Facoltá di Scienze Matematiche, Fisiche e Naturali. Univeritá di Pisa.

Atserias, ., Castellon, I., Civit, M., & Rigau, G. 2000. *Semantic Parsing based on Verbal Subcategorization.*

Atserias, J. 2006. *Towards robustness in Natural Language Understanding.* Ph.D. thesis, Universidad del País Vasco.

Baker, C., Hajic, J., Palmer, M., & Pinkal, M. 2004 (July). Beyond Syntax: Prediates, Arguments, Valency Frames and Linguistic Annotation. *In: Tutorial notes of 42nd Meeting of the Association for Computational Linguistics (ACL2004).*

Baker, C., Ellsworth, M., & Erk, K. 2007. SemEval-2007 Task 19: Frame Semantic Structure Extraction. *In:* (sem, 2007).

Baldewein, U., Erk, K., Padó, S., & Prescher, D. 2004a (May). Semantic Role Labeling With Chunk Sequences. *In:* (con, 2004).

Baldewein, U., Erk, K., Padó, S., & Prescher, D. 2004b. Semantic Role Labelling with Similarity-Based Generalization Using EM-based Clustering. *In:* (sen, 2004).

Battiti, R. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. on Neural Networks*, **5**(4), 537–550.

Bedo, J., Conrad, S., & Kowalczyk, A. 2006 (December). An Efficient Alternative to SVM Based Recursive Feature Elimination with Applications in Natural Language Processing and Bioinformatics. *Pages 170–180 of: Proceedings of the 19th Australian Joint Conference on Artificial Intelligence.*

Universitat d'Alacant
Universidad de Alicante

Bejan, C.A., & Hathaway, C. 2007. UTD-SRL: A Pipeline Architecture for Extracting Frame Semantic Structures. *In:* (sem, 2007).

Bejan, C.A., Moschitti, A., P, Morarescu, Nicolae, G., & Harabagiu, S. 2004. Semantic Parsing Based on FrameNet. *In:* (sen, 2004).

Bengio, Y., & Chapados, N. 2003. Extensions to Metric-Based Model Selection. *Journal of Machine Learning Research*, **3**(March), 1209–1227.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. 2004 (March). Automatic Extraction of Opinion Propositions and their Holders. *In: Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories an Applications (AAAI2004).*

Bi, J., Bennett, K.P., Embrechts, M., Breneman, C.M., & Song, M. 2003. Dimensionality Reduction via Sparse Support Vector Machines. *Journal of Machine Learning Research*, **3**(March), 1229–1243.

Blaheta, D., & Charniak, E. 2000 (May). Assigning Function Tags to Parsed Text. *In:* (naa, 2000).

Blum, A.L., & Langley, P. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligente*, **97**, 245–271.

Blunsom, P. 2004 (December). Maximum Entropy Markov Models for Semantic Role Labelling. *In: Tenth Australian International Conference on Speech Science & Technology.*

Bo, T.H., & Jonassen, I. 2002. New feature subset selection procedures for classification of expression profiles. *Genome Biology*, **34**, 0017.1–0017.11.

Boas, H.C. 2005. Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography*, **18**(4), 445–478.

Bradley, P.S., & Mangasarian, O.L. 1998. Feature selection via convave minimization and support vector machines. *Pages 82–90 of: Proceedings of the 15th International Conference on Machine Learning.* San Francisco: Morgan Kaufman.

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. 2002 (September). The TIGER Treebank. *In: Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002).*

Brassard, G., & Bratley, P. 1996. *Fundamentals of Algorthms.* Englewood Cliffs, New Jersey: Prentice Hall.

Breiman, L. 2001. Random forest. *Machine Learning*, **45**, 5–32.

Brill, E. 1995. Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguis-*

Universitat d'Alacant
Universidad de Alicante

tics, **21**(4), 543–565.

Brill, F.Z., Brown, D.E., & Martin, W.N. 1992. Fast genetic selection of features for neural classifiers. *IEEE Trans. on Neural Networks*, **3**(2), 324–328.

Brown, K., & Miller, J. 1991. *Syntax: A Linguistic Introduction to Sentence Structure*. Harper Collins Academic.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., & Pinkal, M. 2006 (May). The SALSA Corpus: a German Corpus Resource for Lexical Semantics. *In:* (lre, 2006).

Burges, C. J.C. 1998. A Tutorial on Support Vector-Machines for Pattern Recognition. *Data Mining and K. Discovery*, **2**, 121–167.

Busser, R. De, & Moens, M.F. 2003. *Learning generic semantic roles*. Tech. rept. ICRI. Universidad Católica de Leuven. Enviado para publicación a Journal of Machine Learning.

Canisius, S., & den Bosch, A. Van. 2007 (September). Recompiling a knowledge-based dependency parser into memory. *In:* (ran, 2007).

Cardie, C. 1993. Using decision trees to improve case-based learning. *Pages 25–32 of: Proceedings of the 10th International Conference on Machine Learning*. Morgan Kaufmann.

Cardie, C. 1996. Embedded Machine Learning Systems for Natural Language Processing: A General Framework. *In:* Riloff, E., Wermter, S., & Scheler, G. (eds), *Connectionnist, Statistical an Symbolic Approaches to Learning for Natural Language Processing*, vol. LNAI: 1040. Springer.

Cardie, C., & Howe, N. 1977. Empirical methods in information extraction. *Pages 65–79 of:* Fischer, D. (ed), *Proceedings of the 14th International Conference on Machine Learning*. Morgan Kauffman.

Carreras, X., & Màrquez, L. 2003 (September). Phrase recognition by filtering and ranking with perceptrons. *In:* (ran, 2003).

Carreras, X., & Màrquez, L. 2004 (May). Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. *In:* (con, 2004).

Carreras, X., & Màrquez, L. 2005 (June). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. *In:* (con, 2005).

Carreras, X., Màrquez, L., & Chrupala, G. 2004 (May). Hierarchical Recognition of Propositional Arguments with Perceptrons. *In:* (con, 2004).

Caruana, R., & de Sa, V.R. 2003. Benefitting from the Variables that Variable Selection Discards. *Journal of Machine Learning Research*, **3**(March), 1245–1264.

Universitat d'Alacant
Universidad de Alicante

Caruana, R., & Freitag, D. 1994. Greedy Attribute Selection. *Pages 28–36 of:* Kaufman, Morgan (ed), *Proceedings of the 11th International Conference on Machine Learning.*

Castellón, I., Fernández-Montraveta, A., Vázquez, G., Alonso, L., & Capilla, J.A. 2006 (May). The SenSem Corpus: a Corpus Annotated at the Syntactic and Semantic Level. *In:* (lre, 2006).

Celce-Murcia, M. 1972. *Paradigms for Sentence Recognition.* Tech. rept. UCLA. Departamento de Lingúística.

Celce-Murcia, M. 1976. Verb Paradigms For Sentence Recognition. *American Journal of Computational Linguistics*, **1**(83).

Čermák, F. 1995. Jazykkový korpus: Prostředek a zdorj ponání (Language Corpus: A Way and Source of Knowledge). *Slovo a slovesnost*, **56**, 119–140.

Chan, S.W.K. 2006 (February). Shalloww case role annotation using two-stage feature-enhanced string matching. *In:* (cic, 2006).

Charniak, E. 2000 (May). A maximum-entropy inspired parser. *In:* (naa, 2000).

Che, W., Liu, T., Li, S., Hu, Y., & Liu, H. 2005 (June). Semantic Role Labeling System Using Maximum Entropy Classifier. *In:* (con, 2005).

Che, W., Zhang, M., & Liu, S.L. Ting. 2006 (July). A Hybrid Convolution Tree Kernel for Semantic Role Labeling. *In:* (col, 2006).

Chen, J., & Rambow, O. 2003 (July). Use of deep linguistic features for the recognition and labeling of semantic arguments. *In:* (emn, 2003).

Chen, K., Huang, C., Chang, L., & Hsu, H. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. *Pages 167–176 of:* Park, B.-S., & Eds., J.B. Kim. (eds), *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation (PACLIC II).*

Chen, X. 2003. Gene Selection for Cancer Classification Using Bootstrapped Genetic Algortihms and Support Vector Machines. *In: Proceedings of the IEEE Computer Society Bioinformatics Conference.*

Chierchia, G., & McConell-Ginet, S. 1990. *Meaning and Grammar.* Cambridge, MA: MIT Press.

Chieu, H.L., & Ng, H.T. 2003 (May-June). Named Entity Recognition With a Maximum Entropy Approach. *In:* (con, 2003).

Church, K. W., & Hanks, P. 1989 (June). Word Association Norms, Mutual Information, and Lexicography. *In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics.*

Cinková, S. 2006 (May). From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. *In:* (lre, 2006).

Civit, M., Morante, R., Oliver, A., Castelví, J., & Aparicio, J. 2005 (July-Agost). *4LEX: a Multilingual Lexical Resource*. Cross-Language Knowledge Induction Workshop - EuroLAN 2005 Summer School. Cluj-Napoca, Romania.

Clark, P., & Niblett, T. 1989. The CN2 Induction Algorithm. *Machine Learning*, **3**, 261–284.

Cohen, W., & Singer, Y. 1996. Context-sensitive Learning methods for Text Categorization. *In: Proceedings of the 19tn Annual International ACM Conference on Research and Development in Information Retrieval.*

Cohn, T., & Blunsom, P. 2005 (June). Semantic Role Labeling with tree conditional random fields. *In:* (con, 2005).

Collins, M. 1997 (June). Three generative, lexicalised models for statistical parsing. *In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic (ACL1997).*

Collins, M., & Duffy, N. 2002 (July). New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted preceptron. *In:* (acl, 2002).

Collins, M., & Singer, Y. 1999. Unsupervised models for named entity classification. *Pages 100–110 of:* Fung, Pascale, & Zhou, Joe (eds), *Proceedings of 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Contreras, H. 1976. *A Theory of Word Order with Special Reference to Spanish.* Amsterdam: North-Holland.

Cortes, C., & Vapnik, V. 1995. Support-Vector Networks. *Machine Learning*, **20**, 273–297.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. 2002 (July). GATE: A Framework and Graphical Development Enviroment for Robust NLP Tools and Applications. *In:* (acl, 2002).

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. 2003. *TiMBL: Tilburg Memory Based Learner, version 5.0, Reference Guide.* ILK Research Group Technical Report Series 03-10. Tilburg. 56 pages.

Das, S. 2001. Filters, wrappers and boosting-based hybrid for feature selection. *In: Proceedings ICML.* Morgan Kaufmann.

Dash, M., & Liu, H. 1997. Feature selection for classification. *International Journal of Intelligent Data Analysis*, **1**(3), 131–156.

Universitat d'Alacant
Universidad de Alicante

Dash, M., Liu, H., & Motoda, H. 2000. Consistency based feature selection. *Pages 89–109 of: Proceedings of the Pacific-Asian Knowledge and Data Discovery Conference.*

Davies, S., & Russell, S. 1994. NP-Completeness of searches for smallest possible feature sets. *Pages 37–39 of: Proceedings of the AAAI Fall Symposium on Relevance.*

de Mantarás, R. López. 1991. A distance-basead attribute selection measure for decision tree induction. *Machine Learning*, **6**, 91–92.

Dennis, S., Jurafsky, D., & Cer, D. 2003. Supervised and Unsupervised Models for Propositional Analysis. *In: Proceedings of the Workshop on Syntax, Semantics and Statistics at the Neural Information Processing Society Conference.*

Diab, M., Moschitti, A., & Pighin, D. 2007a. CUNIT: A Semantic Role Labeling System for Modern Standard Arabic. *In:* (sem, 2007).

Diab, M., Alkhalifa, M., Elkateb, S., Fellbaum, C., Mansouri, A., & Palmer, M. 2007b. Semeval 2007 Task 18: Arabic Semantic Labeling. *In:* (sem, 2007).

Doak, J. 1994. *An evaluation of search algorithms for feature selection.* Tech. rept. Los Alamos National Laboratory.

Domingos, P. 1997. Context-sensitive feature selection for lazy learners. *Artificial Intelligence Review. Special Issue on Lazy Learners*, **11**, 227–253.

Dong, Z., & Dong, Q. *HowNet.* On line.

Dorr, B.J., Olsen, M., Habash, N., & Thomas, S. 2001. *LCS Verb Database, Online Software Database of Lexical Conceptual Structures and Documentation.*

Dorr, B.J., Levow, G.A., & Lin, D. 2002. Construction of a Chinese-English Verb Lexicon for Machine Translation and Embedded Multilingual Applications. *Machine Translation*, **17**, 99–137.

Dowty, D. R. 1991. Thematic Proto-Roles and Argument Selecction. *Language*, **67**, 547–619.

Draper, N.R., & Smith, H. 1981. *Applied Regresion Analysis.* 2nd edition edn. John Wiley & Sons.

Duda, R.O., & Hart, P.E. 1973. *Pattern Classification and Scene Analysis.*

Duda, R.O., Hart, P.E., & Stork, D.G. 2001. *Pattern Classification.* second edition edn. John Wiley & Sons, Inc.

Dumais, S.T., Platt, J., Heckerman, D., & Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. *Pages 148–155 of: Proceedings of CIKM1998, 7th ACM International Conference on Information*

Universitat d'Alacant
Universidad de Alicante

and Knowledge Management. Bethesda, US: ACM Press, New York, US.

Embrechts, M.J., Arciniegas, F.A., Ozdemir, M., Breneman, C.M., & Benett, K.P. 2001. Bagging Neural Network sensitivity analysis for feature reduction in QSAR problems. *Pages 2478–2482 of: Proceedings of the 2001 INNS-IEEE International Joint Conference on Neural Networks*, vol. 4. Washington, DC: IEEE Press.

Erk, K., & Padó, S. 2006 (May). Shalmaneser - A Toolchain for Shallow Semantic Parsing. *In:* (lre, 2006).

Evans, R. 2003 (September). A Framework for Named Entity Recognition in the Open Domain. *In:* (ran, 2003).

Farwell, D., Helmreich, S., Dorr, B., Habash, N., Reeder, F., Miller, K., Levin, L., Mitamura, T., Hovy, E., Rambow, O., & Siddharthan, A. 2004. Interlingual Annotation of Multilingual Text Corpora. *In: Proceedings of the Workshop in Corpus Annotation in NAACL/HLT2004.*

Fayyad, U.M., & Irani, K.B. 1992. The attribute selection problem in decision tree generation. *Pages 104–110 of: Proceedings of the 10th National Conference on Aritificial Intelligence.* San Jose, CA: MIT Press.

Ferrández, A. 2003. *Sistemas de Pregunta y Respuesta.* Tech. rept. Universidad de Alicante.

Fillmore, C.J. 1968. *The case for case.* in E. Bach and R.T. Harms (ed). Universals in Linguistic Theory. Holt, Rinehart and Winston, New York. Pages 1–88.

Fillmore, C.J. 1969. *Types of lexical information.* in F. KIEFER (ed.) Studies in syntax and semantics. Dordrecht (Holland). Pages 109–137.

Fillmore, C.J. 2002. FrameNet and the Linking between Semantic and Syntactic Relations. *Pages xxviii–xxxvi, address = of: Proceedings of the 19th International Conference on Computational Linguistics (COLING).*

Fillmore, C.J., & Baker, C.F. 2001 (June). Frame Semantics for Text Understanding. *In: Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations Workshop (NAACL2001).*

Fillmore, C.J., Johnson, C.R., & Petruck, M.R.L. 2003. Background to FrameNet. *International Journal of Lexicography*, **16**(3), 235–250.

Fleischman, M., Kwon, N., & Hovy, E. 2003a (June). A Maximum Entropy Approach to FrameNet Tagging. *In:* (hlt, 2003).

Fleischman, M., Kwon, N., & Hovy, E. 2003b (July). Maximum Entropy Models for FrameNet Classification. *In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2003).*

Fliedner, G. 2003. Tools for building a lexical semantic annotation. *Pages 5–9 of: Proceedings of the Workshop Prospects and Advances in the Syntaz/Semantics Interface.*

Fliedner, G. 2007. *Linguistically Informed Question Answering.* Saarbrücken Dissertations in Computational Linguistic and Language Technology, vol. XXIII. Saarbrücken: Universität des Saarlandes und DFKI GmbH.

Folley, W.A., & Valin, R.D. Van. 1984. *Functional syntax and universal grammar.* Cambridge University Press.

Forman, G. 2003. An Extensive Emprirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, **3**(March), 1289–1305.

Foroutan, I. 1987. Feature selection for automatic classification of non-gaussian data. *IEEE Trans. on Systems, Man and Cybernetics*, **17**(2), 187–198.

Forsyth, R. 1988. *Machine Learning. Principles and Techniques.* London, UK, UK: Chapman y H. Ltd. Chap. 1, pages 3–22.

Frank, A., Krieger, H., Xu, F., Uszkoreit, H., Crysmann, B., Jorg, B., & Schafer, U. 2007. Question answering from structured knowledge sources. *Journal of Applied Logic. Special issue on Questions and Answers: Theoretical and Applied Perspectives*, **5**(1), 20–48.

Frawley, W. 1992. *Linguistic Semantics.* Lawrence Erlbaum Associates.

Freund, Y., & Schapire, R.E. 1996. Experiments with a New Boosting Algorithm. *Pages 148–156 of: Proceedings of the 13th International Conference on Machine Learning (ICML'96).* San Francisco, CA: Morgan Kaufmann.

Freund, Y., & Schapire, R.E. 1997. A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**(1), 119–139.

Frohlich, H., Chapelle, O., & Scholkorpf, B. 2003. Feature Selection for Support Vector Machines by Means of Genetic Algorithms. *Pages 142–149 of: Proccedings of the 15th IEEE International Conference on Tools with Artificial Intelligence.*

Fung, G., & Mangasarian, O.L. 2002 (September). *A feature selection Newton method for support vector machine classification.* Technical Report 02-03. Data Mining Institute, Dept. of Computer Science, University of Wisconsin.

Fung, P., & Chen, B. 2004 (August). BiFrameNet: Bilingual Frame Semantics Resource Construction by Cross-lingual Induction. *In:* (col, 2004).

García-Miguel, J.M., & Albertuz, F.J. 2005. Verbs, semantic classes and semantic roles in the ADESSE project. *In: Proceedings of the Interdisciplinary Works-*

hop on Web Features and Verb Classes.

Gildea, D., & Hockenmaier, J. 2003 (July). Identifying semantic roles using combinatory categorial grammar. *In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2003).*

Gildea, D., & Jurafsky, D. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, **28**(3), 245–288.

Gildea, D., & Palmer, M. 2002 (July). The necessity of parsing for predicate argument recognition. *In:* (acl, 2002).

Giménez, J., & Màrquez, L. 2003 (September). Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. *In: Proceedings of Recent Advances in Natural Language Processing (RANLP2003).*

Girju, R., Giuglea, A.M., Olteanu, M., Fortu, O., Bolohan, O., & Moldovan, D. 2004 (May). Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. *In:* (hlt, 2004).

Giuglea, A., & Moschitti, A. 2004 (September). Knowledge Discovering using FramNet, VerbNet and PropBank. *In: Proceedings of the Workshop on Ontology and Knowledge Discovering at ECML 2004.*

Giuglea, A., & Moschitti, A. 2006a (July). Semantic Role Labeling via FrameNet, VerbNet and PropBank. *In:* (col, 2006).

Giuglea, A., & Moschitti, A. 2006b (August). Shallow Semantic Parsing Based on FrameNet, VerbNet and PropBank. *In: Proceedings of the 17th European Conference on Artificial Intelligence (ECAI2006).*

Giuglea, A., & Moschitti, A. 2006c (April). Towards Free-text Semantic Parsing: A Unified Framework Based on FrameNet, VerbNet and PropBank. *In: Proceedings of the Workshop on Learning Structures Information for Natural Language Aplications. Eleventh International Conference on European Association for Computational Linguistics (EACL2006).*

Gomez, F. 1998. Linking WordNet VerbClasses to Semantic Interpretation. *Pages 58–64 of: Proceedings of the COLING-ACL Workshop on the Usage of WordNet in Natural Language Processing Systems.*

Gomez, F. 2004 (July). Building Verb Predicates: A Computational View. *In:* (acl, 2004).

Gomez, F. 2007 (April). Automatic Semantic Annotation of Texts. *Pages 59–66 of:* Kunze, C., Lemnitzer, L., & Osswald, R. (eds), *Workshop on Lexical-Semantics and Ontological Resources (GLDV-2007).*

González, A., & Pérez, R. 1997. Using information measures for determining the relevance of the predictive variables in learning problems. *Pages 1423–1428*

of: Proceedings of the Congress of FUZZ-IEEE'97.

González, L. 1997. *Transitivity and Structural Case Marking in Psych Verbs. A Fragment of an HPSG Grammar of Spanish.* Ph.D. thesis, University of California, Davis.

Gordon, A., & Swanson, R. 2007 (June). Generalizing semantic role annotations across syntactically similar verbs. *In:* (acl, 2007).

Green, R. 2004. *Inducing Semantic Frames from Lexical Resources.* Ph.D. thesis, University of Maryland.

Green, R., & Dorr, B.J. 2005. Frame Semantic Enhancement of Lexical-Semantic Resources. *Pages 57–66 of: Proceedings of the Association for Computational Linguistics (ACL). Workshop on Deep Lexical Acquisition.*

Green, R., Pearl, L., Dorr, B.J., & Resnik, P. 2001 (March). *Lexical Resource Integration Across the Syntax-Semantics Interface.* Tech. rept. LAMP-TR-069,CS-TR-4231,UMIACS-TR-2001-19,CS-TR-4231. University of Maryland, College Park.

Grimshaw, J.B. 1990. *Argument Structure.* Cambridge: MIT Press.

Gruber, J.S. 1965. *Studies in lexical relations.* Ph.D. thesis, Massachusetts Institute of Technology.

Guerra-Salcedo, C., Chen, S., Whitley, D., & Smith, S. 1999. Fast and accurate feature selection using hybrid genetic strategies. *Pages 177–184 of:* P.J. Angeline, Z. Michalewicz, M Schoenauer X. Yao, & Zalzala, A. (eds), *Proceedings of the Congress on Evolutionary Computation*, vol. 1. IEEE Press.

Guitar, J.M. 1998. *El caso gramatical en español en la teoría de los roles semánticos.* Lima (Perú): Editorial Runasimi.

Guyon, I., & Wlisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**(March), 1157–1182.

Hacioglu, K. 2004a (May). A Lightweight Semantic Chunking Model Based On Tagging. *In:* (hlt, 2004).

Hacioglu, K. 2004b (August). Semantic Role Labeling USing Dependency Trees. *In:* (col, 2004).

Hacioglu, K., & Ward, W. 2003 (June). Target Word Detection and Semantic Role Chunking Using Support Vector Machines. *In:* (hlt, 2003).

Hacioglu, K., Pradhan, S., Word, W., Martin, J.H., & Jurafasky, D. 2003 (June). *Shallow Semanctic Parsing Using Support Vector Machines.* Tech. rept. CSLR-2003-1. Center for Spoken Language Research. University of Colorado at Boulder, Boulder, Colorado.

Hacioglu, K., Pradhan, S., Ward, W., Martin, J.H., & Jurafsky, D. 2004 (May). Semantic Role Labeling by Tagging Syntactic Chunks. *In:* (con, 2004).

Haegeman, L. 1991. *Introduction to Government and Binding Theory (Second Edition).* Blackwell.

Haghighi, A., Toutanova, K., & Manning, C. 2005 (June). A Joint Model for Semantic Role Labeling. *In:* (con, 2005).

Hajič, J. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. Bratislava, Slovakia: Jazykovedný ústav Ľ. Štúra, SAV.

Hajič, J., Hajičová, E., Hlaváčová, J., Klimeš, V., Mírovský, J., Pajas, P., Štěpánek, J., Hladká, B.V., & Žabokrtský, Z. 2006 (June). *PDT 2.0 - Guide.* Tech. rept.

Hajičová, E., & Kučerová, I. 2002. Argument-Valency Structure in PropBank, LCS Database and Prague Dependency Treebank: A Comparative Pilot Study. *In:* (Rodríguez & Araujo, 2002).

Hall, M.A., & Holmes, G. 2000. *Benchmarking Attribute Selection Techniques for Data Mining.* Tech. rept. Working Paper 00/10. Department of Computer Science, University of Waikato, New Zealand.

Halliday, M.A.K. 1994. *An introduction to Functional Grammar.* 2nd edition edn. A Hodder Arnold Publication.

Henderson, J. 2003 (June). Inducing history representations for broad-coverage statistical parsing. *In:* (hlt, 2003).

Hensman, S., & Dunnion, J. 2004. Using Linguistic Resources to Construct Conceptual Graph Representation of Texts. *In:* (tsd, 2004).

Hermes, L., & Buhmann, J.M. 2000. Feature Selection for Support Vector Machines. *Pages 716–719 of: Proceedings of the International Conference on Pattern Recognition (ICPR00)*, vol. 2.

Hguyen, H., nd S. Ohn, T. Vua, Park, Y., Han, M.Y., & Kim, Ch.W. 2006 (November). Feature Elimination Approach Based on Random Forest for Cancer Diagnosis. *In:* (mic, 2006).

Higgins, D. 2004 (May). A transformation-based approach to argument labeling. *In:* (con, 2004).

Hirst, G. 1998. Semantic Interpretation and Ambiguity. *Artificial Intelligence*, **34**(2), 131–177.

Hlaváčková, D., & Horák, A. 2006. VerbaLex - New comprehensive lexicon of verb valencies for Czech. *Pages 107–115 of: In Proceedings of the Computer Treatment of Slavic and East European Languages.*

Hochenmaier, J., & Steedman, M. 2002 (July). Generative models for statistical parsing with Combinatory Categorial Grammar. *In:* (acl, 2002).

Holte, R.C. 1993. Very simple classification rules perform well on most commonly use datasets. *Machine Learning*, **11**, 63–91.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. 2006 (May). OntoNotes: 90 % Solution. *In:* (hlt, 2006).

Huang, C., Chen, F., Chen, K., Gao, Z., & Chen, K. 2000 (October). Sinica Treebank: Design Criteria, Annotation Guidelines and On-line Interface. *In: Proceedings of the 2nd Chinese Language Processing Workshop. Held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000).*

Huang, T.M., & Kecman, V. 2005. Gene Extraction for cancer diagnosis by support vector machines - an improvement. *Artificial Intelligence in Medicine*, **35**, 185–194.

Inza, I., Larranaga, P., Etxeberria, R., & Sierra, B. 2000. Feature Subset Selection by Bayesian network-based optimization. *Artificial Intelligence*, **123**(1-2), 157–184.

Jackendoff, R. 1990. *Semantics Structures.* Cambridge, MA: MIT Press.

Jain, A.K., Murty, M.N., & Flynn, P.J. 1999. Data Clustering: A Review. *ACM Computing Surveys*, **31**(3), 264–323.

Jain, A.N. 1990. Parsing complex sentences with structured connectionist networks. *Neural Computation*, **3**, 110–120.

Jebara, T., & Jaakkola, T. 2000. Feature Selection and dualities in maximum entropy discrimination. *In: Proceedings of the International Conference on Uncertainity in Artificial Intelligence.*

Johansson, R., & Nugues, P. 2005a (June). Sparse Bayesian classification of predicate arguments. *In:* (con, 2005).

Johansson, R., & Nugues, P. 2005b. Using Parallel Corpora for Cross-Language Projection of FrameNet Annotation. *In: Proceedings of thw 1st Romance FrameNet Workshop.*

Johansson, R., & Nugues, P. 2006a (May). Construction of a FrameNet Labeler for Swedish Text. *In:* (lre, 2006).

Johansson, R., & Nugues, P. 2006b (July). A FrameNet-based Semantic Role Labeler for Swedish. *In:* (col, 2006).

Johansson, R., & Nugues, P. 2007. LTH: Semantic Structure Extraction using Nonprojective Dependency Trees. *In:* (sem, 2007).

Universitat d'Alacant
Universidad de Alicante

John, G.H., Kohavi, R., & Pfleger, K. 1994. Irrelevant Features and the Subset Selection Problem. *Pages 121–129 of: Machine Learning: Proceedings of the Eleventh International Conference.* San Francisco, CA: Morgan Kaufmann.

John, M.F. St., & McClelland, J.L. 1990. Learning and Applying Contextual Constraints in Sentence Comprehension. *Artificial Intelligence*, **46**, 217–258.

Johnson, C.R., Fillmore, C.J., Petruck, M.R.L., Baker, C.F., Ellsworth, M., Ruppenhofer, J., & Wood, E.J. 2002. *FrameNet: Theroy and Practice.* http://gemini.uab.es/SFN/.

Jurafsky, D., & Martin, J.H. 2000a. *Representing Meaning.* Pages 501–543.

Jurafsky, D., & Martin, J.H. 2000b. *Semantic Analysis.* Pages 545–587.

Kailuweit, R. 2006. *Cursos de Lingüística española.*

Kaisser, M. 2007 (June). Question Answering based on Semantic Roles. *In:* (acl, 2007).

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics. Oxford University Press*, **19**(1), i180–i182.

Kingsbury, P., Palmer, M., & Marcus, M. 2002 (March). Adding Semantic Annotation to the Penn TreeBank. *In: Proceedings of the Human Language Technology Conference (HLT2002).*

Kiparsky, P. 2002. *On the Architecture of Panini's Grammar.* Three lectures delivered at the Hyderabad Conference on the architecture of grammar, Jan. 2002, and at UCLA March 2002.

Kipper, K. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon.* PhD, University of Pensilvania.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. 2006a (May). Extending VerbNet with Novel Verb Classes. *In:* (lre, 2006).

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. 2006b (September). A large-scale extension of VerbNet with novel verb classes. *In: Proceedings of the EURALEX 2006.*

Kira, K., & Rendell, L.A. 1992. The feature selection problem: traditional methods and a new algorithm. *Pages 129–134 of: Proceedings of the 10th National Conference on Artificial Intelligence.*

Kodratoff, Y. 1988. *Introduction to Machine Learning.* Pitman, London.

Kohavi, B., & Frasca, B. 1994. Useful feature subsets and rough set reducts. *Pages 310–317 of: Proceedings of the Third International Workshop on Rough Set*

Universitat d'Alacant
Universidad de Alicante

and Soft Computing (RCSSC-94).

Kohavi, R., & John, G.H. 1997. Wrappers for feature selection. *Artificial Intelligence*, **97**, 273–324.

Koller, D., & Sahami, M. 1996. Toward optimal feature selection. *Pages 284–292 of: Proceedings of the Thirteenth International Conference on Machine Learning.*

Kononenko, I. 1994. Estimating Attributes: Analysis and Extensions of Relief. *In: Proceedings of the European Conference on Machine Learning.*

Koppel, M., & Schler, J. 2004. Authorship verification as a one-class classification problem. *In: Proceedings of the 21st International Conference on Machine Learning (ICML).*

Korhonen, A., & Briscoe, T. 2004. Extended Lexical-Semantic Classification of English Verbs. *In: Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics.*

Kouchnir, B. 2004 (May). A Memory-based Approach for Semantic Role Labeling. *In:* (con, 2004).

Kuroda, K., Utiyama, M., & Isahara, H. 2006 (May). Getting Deeper Semantics than Berkeley FrameNet with MSFA. *In:* (lre, 2006).

Kurohashi, S., & Nagao, M. 2003. *Treebanks: Building and Using Parsed Corpora.* Kluwer Academic, Dordrecht/Boston/London. Chap. Building a Japanese parsed corpus: While improving the parsing system, pages 249–260.

Kwon, M., Fleischman, M., & Hovy, E. 2004. SENSEVAL Automatic Labeling of Semantic Roles using Maximum Entropy Models. *In:* (sen, 2004).

Kwon, N., & Hovy, E. 2006 (February). Integrating Semantic Frames from Multiple Sources. *In:* (cic, 2006).

Lallich, S., & Rakotomalala, R. 2000. Fast feature selection using partial correlation for multivalued attributes. *Pages 221–231 of: Proceedings of the 4th European Conference on Knowledge Discovery in Databases (PKDD2000).*

Langley, P. 1996. *Elements of Machine Learning.* San Francisco: Morgan Kaufmann.

Langley, P., & Sage, S. 1994. Oblivious decision trees and abstract cases. *In: Working Notes of the AAAI-94 Workshop on Case-Based Reasoning.* Seattle, WA: AAAI Press.

Law, Martin H.C., Figueiredo, Mario A.T., & Jain, A.K. 2004. Simultaneous Feature Selection and Clustering Using Mixture Models. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, **26**(9), 1154–1166.

Universitat d'Alacant
Universidad de Alicante

Lee, H.D., Monard, M.C., & Wu, F.Ch. 2006 (October). A Fractal Dimension Based Filter Algorithm to Select Features for Supervised Learning. *Pages 462–472 of: Proceedings of the Joint Conference IBERAMIA/SBIA/SBRN.*

Legrand, G., & Nicolayannis, N. 2005 (July). Feature Selection Method Using Preferences Aggregation. *Pages 203–217 of: Proceedings of the International Conference on Machine Learning and Data Minning (MLDM2005).*

Leite, D. Saraiva, & Rino, L.H. Machado. 2006 (October). Selecting a Feature Set to Summarize Texts in Brazilian Portuguese. *Pages 462–472 of: Proceedings of the Joint Conference IBERAMIA/SBIA/SBRN.*

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., & Villegas, M. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, **13**(4).

Levin, B. 1993. English Verb Classes and Verb Alternations: A Preliminary Investigation. *University of Chicago Press.*

Li, D., & Hu, W. 2006 (December). Feature Selection with RVM and Its Application to Prediction Modeling. *Pages 1140–1144 of: Proceedings of the 19th Australian Joint Conference on Artificial Intelligence.*

Li, G., Yang, J., Liu, G., & Xue, L. 2004 (August). Feature Selection for Multi-Class Problems Using Support Vector Machines. *Pages 292–300 of: Proccedings of Trends in Artificial Intelligence, 8th Pacific Rim International Conference on Ariticial Intelligence (PRICAI2004).*

Li, Guo-Zheng, & Liu, Tian-Yu. 2006. Feature Selection for Bagging of Support Vector Machines. *Pages 271–277 of: Proceedings of the Ninth Pacific Rim International Conference on AI (PRICAI2006).*

Li, X., & Roth, D. 2002 (August). Learning Question Classifiers. *In: Proceedings of the 19th International Conference on Computational Linguistics (COLING2002).*

Lim, J., Hwang, Y., Park, S., & Rim, H. 2004 (May). Semantic Role Labeling using Maximum Entropy Model. *In:* (con, 2004).

Lin, Ch., & Smith, T.C. 2005 (June). Semantic Role Labeling via consensus in pattern-matching. *In:* (con, 2005).

Litkowski, K. 2004. Senseval-3 task: Automatic Labeling of Semantic Roles. *In:* (sen, 2004).

Litkowski, K. 2007. CLR: Integration of FrameNet in a Text Representation Systems. *In:* (sem, 2007).

Liu, H., & Setiono, R. 1995. Chi2: Feature selection and discretization of numeric attributes. *In: Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence.*

Liu, H., & Setiono, R. 1996a. Feature selection and classification. *Pages 419–424 of: Proceedings of the 9th International Conference on Industrial & Engineering Applications of AI & Expert Systems.*

Liu, H., & Setiono, R. 1996b. A Probabilistic Approach to Feature Selection - A Filter Solution. *Pages 319–327 of: International Conference on Machine Learning.*

Liu, H., & Setiono, R. 1998a. Incremental feature selection. *Applied Intelligence*, **9**(3), 217–230.

Liu, H., & Setiono, R. 1998b. Some issues on scalable feature selection. *Expert Systems with Application*, **15**, 333–339.

Liu, H., Motoda, H., & Dash, M. 1998. A monotonic measure for optimal feature selection. *Pages 101–106 of:* Nédellec, C., & Rouveirol, C. (eds), *Proceedings of the 10th European Conference on Machine Learning (ECML-98).*

Liu, Y., & Sarkar, A. 2006 (July). Using LTAG-Based Features for Semantic Role Labeling. *In: Proceedings of the Eighth Workshop on Tree Adjoining Grammars and Related Formalisms: TAG+8. Poster Track. COLING-ACL2006.*

Lo, K.K., & Lam, W. 2006. Using Semantic Relations with World Knowledge for Question Answering. *In: Proceedings of The Fifteenth Text Retrieval Conference (TREC2006).*

Lopatková, M. 2003. Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics*, **79-80**, 37–59.

Loper, E., Yi, S., & Palmer, M. 2007 (January). Combining Lexical Resources: Mapping Between PropBank and VerbNet. *In: Proceedings of The 7th International Workshop on Computational Semantics (IWCS-7).*

Lorenzo, J., Hernández, M., & Méndez, J. 1997 (November). Selección de atributos mediante una medida basada en Información Mutua. *Pages 469–478 of: Proceedings of the VII Conferencia de la Asociación española para la inteligencia artificial (CAEPIA1997).*

Lorenzo, J.J. 2001 (May). *Selección de atributos en aprendizaje automático basada en teoría de la información.* Ph.D. thesis, Universidad de las Palmas de Gran Canaria, Las Palmas de Gran Canaria.

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. *In:* Nikkhou, M. (ed), *Proceedings of the International Conference on Arabic Language Resources and Tools (NEMLAR2004).*

Universitat d'Alacant
Universidad de Alicante

Machine Learning. 2002 (April). *Special issue: Unsupervised learning. 47 (1).*

Machobane, M.M. 1989. *Some Restrictions on the Sesotho Transitivizing Morphemes.* Ph.D. thesis, McGill University, Montreal, Quebec.

Maejía-Lavalle, M., & Arroyo-Figueroa, G. 2006 (November). Power System Database Feature Selection Using a Relaxed Perceptron Paradigm. *In:* (mic, 2006).

Magnini, B., Negri, M., Prevete, R., & Tanev, H. 2002 (July). Is It the Right Answer? Exploiting Web Redundancy for Answer Validation. *In:* (acl, 2002).

Malouf, R. 2002 (June). A Comparison of Algorithms for Maximum Entropy Parameter Estimation. *In: Proceedings of the Sixth Conference on Natural Language Learning (CoNLL2002).*

Manning, C.D., & Schütze, H. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: The MIT Press.

Marcus, M. 1994. The Penn Treebank: A Revised Corpus Design for Extracting Predicate Argument Structure. *In:* Morgan-Kaufman (ed), *Procedings of the ARPA Human Language Technology Workshop.*

Marcus, M.P., Santorini, B., & Marcinkiewicz, M.A. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

Màrquez, L. 2002. *Tratamiento del lenguaje natural.* Edicions Universitat de Barcelona. Chap. Aprendizaje automático y procsamiento del lenguaje natural, pages 133–138.

Màrquez, L., Comas, P., Gimènez, J., & Català, N. 2005 (June). Semantic role labeling as sequential tagging. *In:* (con, 2005).

Màrquez, L., Villarejo, L., Martí, A., & Taulé, M. 2007a. SemEval-2007 Task 09: Multilevel Semantic Annotation of Catalan and Spanish. *In:* (sem, 2007).

Màrquez, L., Padró, L., Surdeanu, M., & Villarejo, L. 2007b. UPC: Experiments with Joint Learning within SemEval Task 9. *In:* (sem, 2007).

Martí, M.A., & Llisterri, J. 2002. *Tratamiento del lenguaje natural.*

Martí, M.A., Alonso, J.A., Badia, T., Campàs, J., Gómez, X., Llisterri, J., Rafel, J., Rodríguez, H., Soler, J., & Verdejo, M.F. 2003. *Tecnologías del lenguaje.*

McClelland, J.L., & Kawamoto, A.H. 1986. *Parallel Distributed Processing.* Vol. 2. A Bradfort Book, MIT Press. Chap. 19, pages 272–325.

Melli, G., Wang, Y., Liu, Y., Kashani, M.M., Shi, Z., Gu, B., Sarkar, A., & Popowich, F. 2006 (June). Description of SQUASH, the SFU Question Answering

Summary Handler for the DUC-2005 Summarization Task. *In: Proceedings of the Document Understanding Conference 2006 (DUC2006).*

Meyers, A., Macleod, C., Yangarber, R., Grishman, R., Barrett, L., & Reeves, R. 1998 (August). Using NOMLEX to Produce Nominalization Patterns for Information Extraction. *In:* Boitet, Christian, & Whitelock, Pete (eds), *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistic and 17th International Conference on Computational Linguistics (COLING-ACL1998).*

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., & Grishman, R. 2004a (May). Annotating Noun Argument Structure for NomBank. *In:* (lre, 2004).

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., & Grishman, R. 2004b (May). The NomBank Project: An Interim Reprot. *In:* (hlt, 2004).

Michalski, R.S. 1983. A Theory and Methodology of Inductive Learning. *Artificial Intelligence*, **20**(2), 111–161.

Miikkulainen, R. 1990. A PDP architecture for processing sentences with relative clauses. *Pages 201–206 of:* Karlgren, H., editor (ed), *Proceedings of the 13th International Conference on Computational Linguistics.*

Miikkulainen, R. 1996. Subsymbilic Case-Role Analysis of Sentences with Embedded Clauses. *Cognitive Science*, **20**, 47–73.

Miikkulainen, R., & Dyer, M.G. 1991. Natural language processing with modular neural networks and distributed lexicon. *Cognitive Science*, **15**, 343–399.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. 1990. *Five Papers on WordNet. CSL Report 43.* Tech. rept. Cognitive Science Laboratory, Princeton University.

Mitkov, R. 2002. *Anaphora resolution.* Longman.

Mitkov, R., Evans, R., Orasan, C., Ha, L.A., & Pekar, V. 2007. Anaphora Resolution: To What Extent Does It Help NLP Applications? *Pages 179–190 of: Proceddings of DAARC.*

Mitsumori, T., Murata, M., Fukuda, Y., Doi, K., & Doi, H. 2005 (June). Semantic role labeling using support vector machines. *In:* (con, 2005).

Modrzejewski, M. 1993. Feature selection using rough sets theory. *Pages 213–226 of: Brazdil . P.B., ed., Proceedings of the European Conference on Machine Learning.*

Mohit, B., & Narayanan, S. 2003 (June). Semantic Extraction with Wide-Coverage Lexical Resources. *In:* (hlt, 2003).

Universitat d'Alacant
Universidad de Alicante

Moldovan, D., Girju, R., Olteanu, M., & Fortu, O. 2004. SVM Classification of FrameNet Semantic Roles. *In:* (sen, 2004).

Molina, L., Belanche, L., & Nebot, A. 2002. FS Algorithms, a survey and experimental evaluation. *Pages 3–14 of: IEEE International Conference on Data Mining.*

Mollá, D. 2003. AnswerFinder in TREC 2003. *In: Proceedings of The 12th Text Retrieval Conference (TREC2003).*

Mollá, D. 2006 (June). *Sistemas de Búsqueda de Respuestas.* Tech. rept. Centre for Language Technology. Division of Information and Communication Sciences.

Montoyo, A., Suárez, A., Rigau, G., & Palomar, M. 2005. Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods. *Journal of Artificial Intelligence Research*, **23**, 299–330.

Mooney, R.J. 1995. Encouraging Experimental Results on Learning CNF. *Machine Learning*, **19**(1), 79–92.

Moore, A.W., & Lee, M.S. 1994. Efficient Algorithms for Minimizing Cross Validation Error. *Pages 190–198 of:* Cohen, W.W., & Hirsh, H. (eds), *Proceedings of the 11th International Confonference on Machine Learning.* Morgan Kaufmann.

Mora, J.P. 2001. Directed motion in English and Spanish. *Estudios de Lingüística Española*, **11**. Capítulo 5. Lexical Semantics of Directed Motion.

Morante, R., & Busser, B. 2007. ILK2: Semanti Role Labelling for Catalan and Spanish using TiMBL. *In:* (sem, 2007).

Morante, R., & van den Bosch, A. 2007 (September). Memory-Based Semantic Role Labelling of Catalan and Spanish. *In:* (ran, 2007).

Moreda, P., & Palomar, M. 2005 (September). Selecting Features for Semantic Roles in QA Systems. *In: Proceedings of Recent Advances in Natural Language Processing (RANLP2005).*

Moreda, P., & Palomar, M. 2006 (August). The Role of Verb Sense Disambiguation in Semantic Role Labeling. *In:* (fin, 2006).

Moreda, P., Palomar, M., & Suárez, A. 2004a (November). Assignment of Semantic Roles based on Word Sense Disambiguation. *In: Proceedings of the 9TH Ibero-American Conference on AI (Iberamia2004).*

Moreda, P., Palomar, M., & Suárez, A. 2004b. Identifying Semantic Roles Using Maximum Entropy Models. *In:* (tsd, 2004).

Moreda, P., Palomar, M., & Suárez, A. 2004c (October). SemRol: Recognition of Semantic Roles. *In: Proceedings of España for Natural Language Processing*

Universitat d'Alacant
Universidad de Alicante

(EsTAL2004).

Moreda, P., Navarro, B., & Palomar, M. 2005 (June). Using Semantic Roles in Information Retrieval Systems. *In: Proceedings of 10th International Conference on Natural Language Processing and Information Systems (NLDB2005).*

Moreda, P., Navarro, B., & Palomar, M. 2007. Corpus-based semantic role approach in information retrieval. *Data and Knowledge Engineering*, **61**(3), 467–483.

Moreda, P., Llorens, H., Saquete, E., & Palomar, M. 2008a (September). The influence of semantic roles in QA: A comparative analysis. *In: Proceedings of the XXIV edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural 2008 (SEPLN´ 08).* Submitted.

Moreda, P., Llorens, H., Saquete, E., & Palomar, M. 2008b (August). Two Proposals of a QA answer extraction module based on semantic roles. *In: Proceedings of the 6th International Conference on Natural Language Processing,GoTAL 2008.* Submitted.

Moreno, L., Palomar, M., Molina, A., & Ferrández, A. 1999a. *Interpretación semántica.* Publicaciones de la Universidad de Alicante. Pages 139–196.

Moreno, L., Palomar, M., Molina, A., & Ferrández, A. 1999b. *Interpretación semántica.* Publicaciones de la Universidad de Alicante. Pages 139–196.

Moreno, L., Palomar, M., Molina, A., & Ferrández, A. 1999c. *Introducción al Procesamiento del Lenguaje Natural.* Publicaciones de la Universidad de Alicante.

Moschitti, A. 2004 (July). A Study on Convolution Kernels for Shallow Semantic Parsing. *In:* (acl, 2004).

Moschitti, A. 2006a (April). Making Tree Kernels Practical for Natural Language Learning. *In:* (eac, 2006).

Moschitti, A. 2006b (May). Syntactic Kernels for Natural Language Learning: the Semantic Role Labeling Case. *In:* (hlt, 2006).

Moschitti, A., Giuglea, A., Coppola, B., & Basili, R. 2005 (June). Hierarchical semantic role labeling. *In:* (con, 2005).

Moschitti, A., Pighin, D., & Basili, R. 2006a (June). Semantic Role Labeling via Tree Kernel Joint Inference. *In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X).*

Moschitti, A., Pighin, D., & Basili, R. 2006b (September). Tree Kernel Engineering for Proposition Re-ranking. *In: In Proceedings of Mining and Learning with Graphs (MLG 2006), Workshop held with ECML/PKDD 2006.*

Moschitti, A., Quarteroni, S., Basili, R., & Manandhar, S. 2007 (June). Exploiting Syntactic and Shallow Semantic Kernels for Question Answer Classification.

In: (acl, 2007).

Moschitti, A., Pighin, D., & Basili, R. 2008. Tree Kernels for Semantic Role Labeling. *Special Issue on Semantic Role Labeling at Computational Linguistics*, **34**(2).

Mucciardi, A.N., & Gose, E.E. 1971. A comparison of seven techniques for choosing subsets of pattern recognition. *IEEE Transactions on Computers*, **20**(September), 1023–1031.

Musillo, G., & Merlo, P. 2006 (May). Accurate Parsing of the Proposition Bank. *In:* (hlt, 2006).

Narayanan, S., & Harabagiu, S. 2004 (August). Question Answering Based on Semantic Structures. *In:* (col, 2004).

Narayanan, S., Fillmore, C.J., Baker, C.F., & Petruck, M.R.L. 2002. FrameNet Meets the Semantic Web: a DAML+OIL Frame Representation. *In: Proceedings of the Eighteenth National Conference on Artificial Intelligence*. Eighteenth National Conference on Artificial Intelligence, Edmonton, Canada.

Narendra, P., & Fukunaga, K. 1977. A branch and bound algorithm for feature selection. *IEEE Trans. on Computers*, **26**, 917–922.

Navarro, B., Moreda, P., Fernández, B., Marcos, R., & Palomar, M. 2004 (November). Anotación de roles semánticos en el corpus 3LB. *In: Proceedings of the Workshop Herramientas y Recursos Linguísticos para el Español y el Portugués*. Workshop Herramientas y Recursos Linguísticos para el Español y el Portugués. The 9TH Ibero-American Conference on Artificial Intelligence (IBERAMIA 2004), Tonantzintla, México.

Neal, R.M. 1998. *Assesing relevance determination methods using DELVE*. Neural Networks and Machine Learning, pages 97-129.

Neter, J., Wasserman, W., & Kutner, M.H. 1990. *Applied Linear Statistical Models*. 3rd edition edn. Irwin: Homewood, IL.

Neuman, J., Schorr, C., & Steidl, G. 2005. Combined SVM-Based Feature Selection and Classification. *Machine Learning*, **61**(1-3), 129–150.

Ng, A.Y. 1998. On feature selection: learning with exponentially many irrelevant features as training examples. *In: Proceedings of the Fifteenth International Conference on Machine Learning*.

Ngai, G., Wu, D., Carpuat, M., Wang, C.S., & Wang, C.Y. 2004. Semantic Role Labeling with Boosting, SVMs, Maximum Etropy, SNoW and Decision Lists. *In:* (sen, 2004).

Nielsen, R.D., & Pradhan, S. 2004 (July). Mixing Weak Learners in Semantic Parsing. *In:* (emn, 2004).

Nigam, K., & Ghani, R. 2000. Understanding the behavior of co-training. *Pages 105–106 of: Procedings of the Workshop on Text Mining at the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Ofoghi, B., Yearwood, J., & Ghosh, R. 2006 (December). A Hybrid Question Answering Schema Using Encapsulated Semantics in Lexical Resources. *Pages 1276–1280 of: Advances in Artificial Intelligence, 19th Australian Joint Conference on Artificial Intelligence.*

Ohara, K.H., Fuji, S., & Saito, H. 2003 (August). The Japanese FrameNet project: A preliminary report. *Pages 249–254 of: Proceedings of Pacific Association for Computational Linguistics (PACLING2003).*

Ohara, K.H., Fuji, S., Ohori, T., Suzuki, R., Saito, H., & Ishizaki, S. 2004 (May). The Japanese FrameNet Project: An Introduction. *In:* (lre, 2004).

Ohara, T., & Wiebe, J. 2002 (Decembre). *Classifying Preposition Semantic Roles using Class-based Lexical Associations.* Tech. rept. NMSU-CS-2002-13. Computer Science Department, New Mexico State University.

Ohara, T., & Wiebe, J. 2003 (May-June). Preposition Semantic Classification via Penn Treebank and FrameNet. *In:* (con, 2003).

Ozgencil, N.E., & McCracken, N. 2005 (June). Semantic role labeling using libSVM. *In:* (con, 2005).

Padó, U., crocker, M., & Keller, F. 2006 (April). Modelling Semantic Role Plausibility in Human Sentence Processing. *In:* (eac, 2006).

Pado, S., & Boleda, G. 2004a (July). The Influence of Argument Structure on Semantic Role Assignment. *In:* (emn, 2004).

Pado, S., & Boleda, G. 2004b (August). Towards Better Understanding of Automatic Semantic Role Assignment. *In:* (col, 2004).

Pado, S., & Lapata, M. 2005 (October). Cross-linguistic Projection of Role-Semantic Information. *In: Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP2005).*

Palmer, F.R. 1994. *Grammatical Roles and Relations.* Cambridge: Cambridge UP.

Palmer, M., Rosenzweig, J., & Cotton, S. 2001 (March). Automatic Predicate Argument Analysis of the Penn TreeBank. *In: Proceedings of the Human Language Technology Conference (HLT2001).*

Palmer, M., Gildea, D., & Kingsbury, P. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, **31**(1), 71–106.

Palomar, M., Civit, M., Díaz, A., Moreno, L., Bisbal, E., Aranzabe, M., Ageno, A., Martí, M.A., & Navarro, B. 2004. 3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano. *Procesamiento del Lenguaje Natural.*

Park, K., Hwang, Y., & Rim, H. 2004 (May). Two-Phase Semantic Role Labeling bsed on Support Vector Machines. *In:* (con, 2004).

Park, K., Hwang, Y., & Rim, H. 2005 (June). Maximum Entropy based Sematnic Role Labeling. *In:* (con, 2005).

Pazienza, M.T., Pennacchiotti, M., & Zanotto, F.M. 2006 (May). Mixing WordNet, VerbNet and PropBank for studying verb relations. *In:* (lre, 2006).

Perkins, S., Lacker, K., & Theiler, J. 2003. Grafting: Fast, Incremental Feature Selection by Gracient Descent im Function Space. *Journal of Machine Learning Research*, **3**(March), 1333–1356.

Philpot, A., Hovy, E., & Pantel, P. 2005 (October). The Omega Ontology. *In: Proceedings of the Ontologies and Lexical Resources Workshop (ONTOLEX) at IJCNLP.*

Pighin, D., & Moschitti, A. 2007 (September). A Tree Kernel-Based Shallow Semantic Parser for Thematic Role Extraction. *Pages 350–361 of:* Basili, Roberto, & Pazienza, Maria Teresa (eds), *In proceedings of AI*IA 2007: Artificial Intelligence and Human-Oriented Computing, 10th Congress of the Italian Association for Artificial Intelligence.* Lecture Notes in Computer Science, vol. 4733.

Ping, J. Zheng. 2005 (April). *Semantic Role Labeling.* Graduate Research Paper. Department of Computer Science, School of Computing, National University of Singapore.

Piramuthu, S. 1998. Evaluaitng feature selection methods for learning in data mining applications. *Pages 294–301 of: Proceedings of the 31st annual Hawaii International Conference on system sciences.*

Pizzato, L.A. Sangoi, & Mollá-Aliod, D. 2005 (December). Extracting Exact Answers using a Meta Question answering System. *In: Proceedings of the Australasian Language Technology Workshop 2005 (ALTW05).*

Pollard, C., & Sag, I.A. 1988. *Information-based syntax and semantics: Vol. 1: fundamentals.* Stanford, CA, USA: Csli Lecture Notes; Vol. 13. Center for the Study of Language and Information.

Ponzetto, S.P., & Strube, M. 2005 (June). Semantic role labeling using lexical statistical information. *In:* (con, 2005).

Pradhan, S., Hacioglu, K., Ward, W., Martin, J.H., & D.Jurafsky. 2003 (November). Semantic role parsing: Adding semantic structure to unstructured text. *In: Proceedings of the Third IEEE International Conference on Data Mining*

(ICDM2003).

Pradhan, S., Sun, H., Ward, W., Martin, J.H., & D.Jurafsky. 2004a (May). Parsing Arguments of Nominalizations in English and Chinese. *In:* (hlt, 2004).

Pradhan, S., Ward, W., Hacioglu, K., Martin, J.H., & D.Jurafsky. 2004b (July). Semantic Role Labeling Using Different Syntactic Views. *In:* (acl, 2004).

Pradhan, S., Ward, W., Hacioglu, K., Martin, J.H., & D.Jurafsky. 2004c (May). Shallow Semantic Parsing using Support Vector Machines. *In:* (hlt, 2004).

Pradhan, S., Hacioglu, K., Ward, W., Martin, J.H., & D.Jurafsky. 2005a (June). Semantic role chunking combining complementary syntactic views. *In:* (con, 2005).

Pradhan, S., Ward, W., Hacioglu, K., Martin, J.H., & D.Jurafsky. 2005b (June). Semantic role labeling using different syntactic views. *In:* (acl, 2005).

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J., & Jurafsky, D. 2005c. Support Vector Learning for Semantic Argument Classification. *Machine Learning*, **60**(1-3), 11–39.

Pradhan, S., e. Loper, Dligach, D., & Palmer, M. 2007. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. *In:* (sem, 2007).

Pradhan, S., Ward, W., & Martin, J.H. 2008. Towards Robust Semantic Role Labeling. *Computational Linguistics. Special issue on Semantic Roles*, **34**(2).

Punyakanok, V., Roth, D., Yih, W., Zimak, D., & Tu, Y. 2004 (May). Semantic Role Labeling Via Integer Linear Programming Inference. *In:* (con, 2004).

Punyakanok, V., Roth, D., Yih, W., & Zimak, D. 2005a (June). Generalized inference with multiple semantic role labeling systems. *In:* (con, 2005).

Punyakanok, V., Roth, D., & Yih, W. 2005b (August). The Necessity of Syntactic Parsing for Semantic Role Labeling. *Pages 1117–1123 of: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI2005).*

Punyakanok, V., Roth, D., & Yih, W. 2008. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics. Special issue on Semantic Roles*, **34**(2).

Quinlan, J.R. 1990. Learning Logical Definitions from Relations. *Machine Learning*, **5**(3), 239–266.

Quinlan, J.R. 1993. *C4.5: Programs of Machine Learning.* Los Altos, California: Morgan Kauffman.

Quixtiano-Xicohténcatl, R., Reyes-Galaviz, O.F., Flores-Pulido, L., & Reyes-García, C.A. 2006 (November). Hybrid Algorithm Applied to Feature Se-

Universitat d'Alacant
Universidad de Alicante

lection for Speaker Authentication. *In:* (mic, 2006).

Rabiner, L.R. 1990. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *In:* A. Waibel, K.F. Lee (ed), *Reafings in Speech Recognition.* San Mateo, CA: Morgan Kaufmann Publishers, Inc.

Rakotomamonjy, A. 2003. Variable Selection Using SVM-based Criteria. *Journal of Machine Learning Research*, **3**(March), 1357–1370.

Rambow, O., Dorr, B., Kipper, K., Kučerová, I., & Palmer, M. 2003. Automatically Deriving Tectogrammatical Labels from Other Resources. *The Prague Bulletin of Mathematical Linguistics*, **79-80**, 23–35.

Ratnaparkhi, A. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution.* Ph.D. thesis, University of Pennsylvania.

R.D. Van Valin, Jr. 2005. *A Summary of Role and Reference Grammar.* http://linguistics.buffalo.edu/research/rrg/RRGsummary.pdf.

Reeder, F., Dorr, B., Farwell, D., Nabash, N., Helmreich, S., Hovy, E., Levin, L., Mitamura, T., Miller, K., Rambow, O., & Siddharthan, A. 2004. Interlingual Annotation for MT Development. *In: Proceedings of the AMTA.*

Richardson, S.D., Dolan, W.B., & Vanderwende, L. 1998. MindNet: acquiring and structing semantic information from text. *In: Proceedings of the The Twelth International Conference on Computational Linguistics (COLING1998).*

Riloff, E., & Schmelzenbach, M. 1998. An Emprirical Approach to Conceptual Case Frame Acquisition. *In: Proceedings of Sixth Workshop on Very Large Corpora.* Montreal, Quebec, Canada: August.

Rodríguez, R. M., & Araujo, C. Paz Suárez (eds). 2002. *Third International Conference on Language Resources and Evaluation (LREC2002).* Vol. 5. Las Palmas, España: European Language Resources Association.

Rosa, J.L. Garcia. 2001 (October). HTRP II: Learning thematic relations from semantically sound sentences. *Pages 488–493 of: Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics (SMC2001).*

Rosa, J.L. Garcia. 2007 (June). A Connectionist Thematic Grid Predictor for Preparsed Natural Language Sentences. *Pages 825–834 of: Advances in Neural Networks. International Symposium on Neural Networks.*

Rosenblatt, F. 1959. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, **65**, 386–408.

Roth, D. 1998 (July). Learning to Resolve Natural Language Ambiguities: A Unified Approach. *Pages 806–813 of:* Press, MIT (ed), *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI1998).*

Ruimy, N., o. Corazzari, Gola, O., Spanu, E., Calzolari, N., & Zampolli, A. 1998. The European LE-PAROLE Project: The Italian Syntactic Lexicon. *In: Proceedings of the first International Conference on Language Resources and Evaluation (LREC1998)*. Granada, España: European Language Resources Association.

Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Uliveri, M., Calzolari, N., & Zampolli, A. 2002. Clips, a Multi-level Italian Computational Lexicon: a Glimpse to Data. *In:* (Rodríguez & Araujo, 2002).

S. Wen-tau Yih and K. Toutanova. 2006 (May). Automatic Semantic Role Labeling. *In:* (hlt, 2006). Tutorial.

Saggion, H., & Gaizauskas, R. J. 2006 (August). Experiments in Passage Selection and Answer Identification for Question Answering. *In:* (fin, 2006).

Samuel, K. 1998. Lazy Transformation-Based Learning. *Pages 235–239 of: Proceedings of the 11th International Florida Artificial Intelligence Research Symposium Conference.*

Sang, E.F.Tjong Kim, S.Canisius, & van den Bosch adn T. Bogers, A. 2005 (June). Applying spelling error correction techniques for improving Semantic Role Labeling. *In:* (con, 2005).

Schank, R.C. 1972. Conceptual Dependency: A Theory of Natural Language processing. *Cognitive Psychology*, **3**(4), 552–631.

Schapire, R.E., & Singer, Y. 1999. Improved Boosting Algorithms Using Confidence-rated Predictions. *Machine Learning*, **37**(3), 297–336.

Scherf, M., & Brauer, W. 1997. Improving RBF networks by the feature selection approach EUBAFES. *Pages 391–396 of: Proceedings of the 7th International Conference on Artificial Neurol Networks (ICANN97).*

Schmidt, T. 2006. *Kicktionary.* http://www.kicktionary.de/Introduction.html.

Semecký, J., & Cinková, S. 2006. Constructing and English Valency Lexicon. *Pages 111–113 of: Proceedings of Frontiers in Linguistically Annotated Corpora.* Sydney, Australia: The Association for Computational Linguistics.

Setiono, R., & Liu, H. 1996. Improving backpropagation learning with feature selection. *Applied Intelligence*, **6**, 129–139.

Setiono, R., & Liu, H. 1997. Neural-network feature selector. *IEEE Trans. on Neural Networks*, **8**(3), 654–662.

Sgall, P. 2001. *Three Chapters on English Syntax.* http://ufal.mff.cuni.cz/publications/year2000/esynt.doc.

Universitat d'Alacant

Sgall, P., Hajičová, E., & Panevová, J. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Prague, Czech Republic/Dordrecht, Netherlands: Academia/Reidel Publishing Company.

Sgall, P., Žabokrtský, Z., & Džeroski, S. 2002. A Machine Learning Approach to Automatic Functor Assignment in the Prague Dependency Treebank. *In:* (Rodríguez & Araujo, 2002).

Sheinvald, J., Dom, B., & Nibalck, W. 1990. A modelling approach to feature selection. *Pages 535–539 of: Proceedings of the Tenth International Conference on Pattern Recognition*, vol. 1.

Shen, D., Wiegand, M., Merkel, A., Kazalski, S., Hunsicker, S., Leidner, J.L., & Klakow, D. 2007. The Alyssa System at TREC QA 2007: Do We Need Blog06? *In: Proceedings of The Sixteenth Text Retrieval Conference (TREC2007)*.

Shi, L., & Mihalcea, R. 2004 (May). Open Text Semantic Parsing Using FrameNet and WordNet. *In:* (hlt, 2004).

Shi, L., & Mihalcea, R. 2005 (February). Putting Pieces Toghether: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. *Pages 100–111 of: Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*.

Siedlecki, W., & Skalansky, J. 1989. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, **10**, 335–347.

Skalak, D.B. 1994. Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms. *Pages 293–301 of: Proceedings of the Eleventh International Machine Learning Conference*.

Sowa, J.F. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Addison Wesley.

Stallard, D. 2000. Talk'n'Travel: A conversational system for air travel planning. *In: Proceedings of the 6th Applied Natural Language Processing Conference (ANLP2000)*.

Stenchikova, S., Hakkani-Tur, D., & Tur, G. 2006 (September). QASR: Question Answering Using Semantic Role for Speech Interface. *In: Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP))*.

Stevens, G. 2007. XARA: An XML- and rule-based semantic role labeler. *Pages 113–116 of: Proceedings of the Linguistic Annotation Workshop*. Prague, Czech Republic: Association for Computational Linguistics.

Stoppiglia, H., Dreyfus, G., Dubois, R., & Oussar, Y. 2003. Ranking a Random Feature for Variable and Feature Selection. *Journal of Machine Learning Research*, **3**(March), 1399–1414.

Universitat d'Alacant
Universidad de Alicante

Subirats, C. 2006. *FrameNet Español: un análisis cognitivo del léxico del español.* In Amparo Alcina, ed.

Subirats, C., & Petruck, M.R.L. 2003. Surprise: Spanish FrameNet. *In: Proceedings of the Workshop on Frame Semantics at eh XVII. International Congress of Linguistics.*

Sun, H., & Jurafsky, D. 2004 (May). Shallow Semantic Parsing of Chinese. *In:* (hlt, 2004).

Sun, R., Jiang, J., Tan, Y.F., Cui, H., Chua, T., & Kan, M. 2005. Using Syntactic and Semantic Relation Analysis in Question Answering. *In: Proceedings of The Fourteenth Text Retrieval Conference (TREC2005).*

Surdeanu, M., & Turmo, J. 2005 (June). Semantic Role Labeling using complete syntactic analysis. *In:* (con, 2005).

Surdeanu, M., & Turmo, J. 2008 (February). Analysis of Joint Inference Strategies for the Semantic Role Labeling of Spanish and Catalan. *In:* (cic, 2008).

Surdeanu, M., Harabagiu, S., Williams, J., & Aarseth, P. 2003 (July). Using predicate-argument structures for information extraction. *In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003).*

Surdeanu, M., Màrquez, L., Carreras, X., & Comas, P.R. 2007. Combination Strategies for Semantic Role Labeling. *Journal of Artificial Intelligence Research (JAIR)*, **29**, 105–151.

Suárez, A. 2004. *Resolución de la ambigüedad semántica de las palabras mediante modelos de probabilidad de máxima entropía.* Ph.D. thesis, Universidad de Alicante.

Suárez, A., Palomar, M., & Rigau, G. 2005. Reentrenamiento: Aprendizaje Semisupervisado de los Sentidos de las Palabras. *Procesamiento del Lenguaje Natural*, **34**, 299–330.

Sutton, Ch., & McCallum, A. 2005 (June). Joint parsing and Semantic Role Labeling. *In:* (con, 2005).

Swier, R.S., & Stevenson, S. 2004 (July). Unsupervised Semantic Role Labelling. *In:* (emn, 2004).

Taulé, M., Castellví, J., Martí, M.A., & Aparicio, J. 2006. Fundamentos teóricos y metodológicos para el etiquetado semántico de CESS-CAT y CESS-ESP. *Procesamiento del Lenguaje Natural*, 75–82.

Thompson, A., Levy, R., & Manning, C.D. 2003 (September). A generative model for semantic role labeling. *In: Proceedings of the 14th European Conference on Machine Learning (ECML2003).*

Universitat d'Alacant
Universidad de Alicante

Thompson, A., Patwardhan, S., & Arnold, C. 2004. Generative models for semantic role labeling. *In:* (sen, 2004).

Torkkola, K., Venkatesan, S., & Huan, L. 2004. Sensor selection for maneuver classification. *Pages 636–641 of: Proccedings of the 7th International IEEE Conference on Intelligent Transportation Systems.*

Toutanova, K., Haghighi, A., & Manning, C.D. 2005 (June). Joint Learning Improves Semantic Role Labeling. *In:* (acl, 2005).

Trandabăţ, M.D. 2007. Semantic Frames in Romanian Natural Language Processing Systems. *Pages 29–32 of: Proceedings of the NAACL-HLT 2007 Doctoral Consortium.* Rochester, New York: Association for Computational Linguistics.

Tsai, R.T.-H., Chou, W.-Ch., Lin, Y.-Ch., Sung, Ch.-L., Ku, W., Su, Y.-S., Sung, T.-Y., & Hsu, W.-L. 2006 (June). BIOSMILE: Adapting Semantic Role Labeling for Biomedical Verbs: An Exponential Model Coupled with Automatically Generated Template Features. *Pages 57–64 of: In Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 2006.*

Tsai, T., Wu, C., Lin, Y., & Hsu, W. 2005 (June). Exploiting full parsing information to label semantic roles using an ensemble of me and svm via integer linear programming. *In:* (con, 2005).

Tsamardinos, I., Brown, L.E., & Aliferis, C.F. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, **65**(1), 31–78.

Vafaie, H., & Imam, I.F. 1994. Feature Selection methods: Genetic algorithms vs. greedy-like search. *In: Proceedings of the 3rd International Conference on Fuzzy Systems and Intelligence Control.*

Vafaie, H., & Jong, K. De. 1993. Robust feature selection algorithms. *Pages 356–363 of: Proceedings of the 5th IEEE International Conference on Tools for Artificial Intelligence.* IEEE Press.

Valin, R.D. Van, & Polla, R. La. 1997. *Syntax, Structure, Meaning and Function.* Cambridge University Press.

van den Bosch, A., Canisius, S., Hendricks, I., Daelemans, W., & Sang, E.T.K. 2004 (May). Memory-based semantic role labeling: Optimizing features, algorithm and output. *In:* (con, 2004).

van den Bosch, A., Busser, G.J., Canisius, S., & Daelemans, W. 2007. An efficient memory-based morpho-syntactic tagger and parser for Dutch. *Pages 99–114 of:* P. Dirix, I. Schuurman, V. Vandeghinste, & Eynde, F. Van (eds), *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting.*

Universitat d'Alacant
Universidad de Alicante

Venkatapathy, S., Bharati, A., & Reddy, P. 2005 (June). Inferring semantic roles using subcategorization frames and maximum entropy model. *In:* (con, 2005).

Vossen, P. 1998. EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newsletter*, **3**(1).

Žabokrtský, Z. 2000 (September). Automatic Functor Assignment in the Prague Dependency Treebank. *Pages 45–50 of: Proceedings of the Third International Workshop on Text, Speech and Dialogue.*

Vázquez, G., Fernández, A., & Martí, M. A. 2000. *Clasificación Verbal: Alternancias de Diátesis.* Universitat de Lleida.

Wagner, A. 2005. *Learning Thematic Role Relations for Lexical Semantic Nets.* Ph.D. thesis, University of Tubingen.

Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C., & Doddington, G. 2003. *Multiple-Translation Arabic (MTA) Part 1.* Linguistic Data Consortium (LDC) catalog number LDC2003T18.

Wang, H., Bell, D., & Murtagh, F. 1999. Axiomatic approach to feature subset selection based on relevance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **21**(3), 271–277.

Wasow, T. 2003. *Reviews of Form and meaning in language, Vol. 1.*

Wechsler, S. 1995. *The Semantic Basis of Argument Structure.* CSLI Publications, Stanford.

Weston, J. ., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. 2001. Feature selection for svms. *Pages 668–674 of: Neural Information Processing Systems.* Cambridge, MA: MIT Press.

Weston, J., Elisseff, A., Scholkopf, B., & tipping, M. 2003. Use of the Zero-Norm with Linear Models and Kernel Methods. *Journal of Machine Learning Research*, **3**(March), 1439–1461.

White, J., & O'Connell, T. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. *In: Proceedings of the 1994 Conference, Association for Machine Translation in the Americas.*

Williams, K., Dozier, C., & McCulloh, A. 2004 (May). Learning Transformation Rules for Semantic Role Labeling. *In:* (con, 2004).

Wu, Y., & Zhang, A. 2004. Feature selection for classifying high-dimensional numerical data. *Pages 251–258 of: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2.

Xing, E., Jordan, M., & Carp, R. 2001. Feature selection for highdimensional genomic microarray data. *In: Proccedings of the 18th ICML.*

Xue, N., & Palmer, M. 2003. Annotating the Propositions in the Penn Chinnese Treebank. *In: Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing.*

Xue, N., & Palmer, M. 2004 (July). Calibrating Features for Semantic Role Labeling. *In:* (emn, 2004).

Yan, G., Li, Z., & Yuan, L. 2006 (November). On Combining Fractal Dimension with GA for Feature Subset Selecting. *In:* (mic, 2006).

Yang, H.H., & Moody, J. 1999 (June). Feature selection based on joint mutual information. *In: Advances in Intelligent Data Analysis (AIDA), Computational Intelligence Methods and Applications (CIMA), International Computer Science Conventions.*

Yang, J., & Honavar, V. 1998. Feature Subset Selection Using a Genetic Algorithm. *IEEE Intelligent Systems*, **13**, 44–49.

Ye, P., & Baldwin, T. 2005 (October). Semantic Role Labelling of Prepositional Phrases. *Pages 779–791 of: Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP2005).*

Yi, S., & Palmer, M. 2005 (June). The integration of syntactic parsing and semantic role labeling. *In:* (con, 2005).

You, J., & Chen, K. 2004. Automatic Semantic Role Assignment for a Tree Structure. *In: Proceedings of SIGHAN Workshop.*

Yousefi, J., & Kosseim, L. 2006 (May). Using Semantic Constraints to Improve Question Answering. *Pages 118–128 of: Proceedings of 11th International Conference on Natural Language Processing and Information Systems (NLDB2006).*

Zapirain, B., Aguirre, E., & Màrquez, L. 2008 (February). A Prelimnary Study on the Robutness and Generalization of Role Sets for Semantic Role Labeling. *In:* (cic, 2008).

Zhang, Ch., Liang, Y., Xiong, W., & Ge, H. 2006a (December). Selection for Feature Gene Subset in Microarray Expression Profiles Based on an Improved Genetic Algorithm. *Pages 161–169 of: Proceedings of the 19th Australian Joint Conference on Artificial Intelligence.*

Zhang, H., Yu, Ch., & Singer, B. 2003. Cell and tumor classification using genetic expression data: Construction forest. *Pages 4168–4172 of: Proceedings of the National Academy of Sciences of the United States of America*, vol. 100.

Zhang, Q., Weng, F., & Feng, Z. 2006b (July). A Progressive Feature Selection Algorithm for Ultra Large Feature Spaces. *Pages 561–568 of: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL2006).*

Zhou, Y., Weng, F., Wu, L., & Schmidt, H. 2003 (July). A fast Algorithm for Feature Selection in Conditional Maximum Entropy Modeling. *In:* (emn, 2003).

Zhu, J., Rosset, S., Hastie, T., & Tibshirani, R. 2004. 1-norm Support Vector Machines. *In:* S. Thrun, L. Saul, & Scholkopf, B. (eds), *Advances in Neural Information Processing Systems*, vol. 16. Cambridge, MA, USA: MIT Press.

Universitat d'Alacant
Universidad de Alicante