Una Propuesta para el Etiquetado Automático de Roles Semánticos

Isabel Segura Bedmar, José L. Martínez Fernández, Paloma Martínez
Departamento de Informática
Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés, Madrid
{isegura, jlmferna, pmf}@inf.uc3m.es

Resumen: La identificación de los roles semánticos es una parte crucial en tareas que involucran tratamiento automático del lenguaje natural como la extracción y recuperación de información, sistemas de búsquedas de respuestas, generación de resúmenes, traducción automática, etc. Para el caso del español, la investigación en roles semánticos es escasa. El objetivo del actual trabajo es analizar los sistemas de etiquetado de roles semánticos y proponer un nuevo enfoque del problema para mejorar los resultados basado en un mayor explotación de los recursos léxicos.
Palabras clave: Roles Semánticos, Clasificación Argumentos Semánticos

1 Introducción

El objeto de un analizador semántico es identificar las relaciones semánticas entre las palabras y la construcción de una estructura que permite la interpretación del significado del texto (Shi y Mihalcea, 2005).

La identificación de los roles semánticos, en el sentido de conocer el papel que desempeña cada elemento de una oración en relación con otros elementos, por ejemplo, verbo y sus argumentos, nombre y sus modificadores, es una parte crucial en la interpretación de los textos (Gildea y Palmer, 2002), y por tanto es importante para tareas como extracción y recuperación de información, sistemas de búsqueda de respuestas, generación de resúmenes, interfaces en lenguaje natural, etc. (Hacioglu y Ward, 2003), (Navarro et al. 2004).

En la década de los 90, los sistemas de extracción de información se basaban en reglas que cubrían una amplia variedad de fenómenos semánticos o en modelos estadísticos o de estados finitos.. Otros sistemas comerciales han utilizado técnicas de representación del conocimiento empleados tradicionalmente en IA. En la actualidad, el reto es conseguir desarrollar sistemas independientes del dominio o, por lo menos, fácilmente ajustables a cualquier dominio semántico.

Los sistemas de búsquedas de respuestas necesitan información lingüística para afrontar con garantías la tarea de la localización de la respuesta correcta (Navarro et al. 2004) en documentos. Los roles semánticos juegan un papel fundamental, y con esta información se pueden responder a preguntas como "quién", "cuándo", "dónde" o "qué".

El objetivo de este trabajo consiste en revisar los distintos enfoques y métodos empleados por los sistemas de etiquetado de roles semánticos, detectar sus fortalezas y defectos, y proponer un nuevo enfoque que mejore los resultados. Este enfoque se basa en el uso de un segmentador sintáctico (*chunker*), incluido en la plataforma GATE (http://gate.ac.uk/) y potenciar la utilización de información semántica, como por ejemplo la clase semántica obtenida a partir de WordNet.

2 Roles Semánticos

Los roles semánticos describen la relación semántica (no gramatical) que los argumentos tienen con respecto al predicado (normalmente un verbo). Otros términos utilizados para su denominación son: roles temáticos, casos semánticos, relaciones temáticas, argumentos semánticos, roles de los participantes, etc.

Un rol semántico describe una función abstracta desempeñada por un elemento que participa en una acción. Esta función abstracta

se define independientemente de las distintas estructuras sintácticas que puede adquirir una oración. De esta forma, los roles semánticos permiten representar de forma genérica acciones, sin depender del idioma ni de los diversos recursos gramaticales que ofrece un idioma para expresar una misma acción.

En las siguientes oraciones, los roles semánticos del predicado *romper* tienen realizaciones sintácticas distintas:

[Juan $_{\text{Agent}}$] [rompió $_{\text{V}}$] [la ventana $_{\text{Object}}$] con [el martillo $_{\text{Instrument}}$]

[La ventana $_{\text{Object}}$] [fue rota $_{\text{V}}$]por [Juan $_{\text{Agent}}$]

[El martillo $_{\text{Instrument}}$] [rompió $_{\text{V}}$] [la ventana $_{\text{Object}}$]

A diferencia del nivel sintáctico, donde hay más o menos acuerdo entre la comunidad lingüística sobre los constituyentes sintácticos y su definición, con los roles semánticos no hay acuerdo alguno sobre qué roles semánticos existen ni las características de cada uno.

Los lingüistas están interesados en describir la relación entre los roles semánticos de un verbo y sus posibles realizaciones sintácticas. Por este motivo, han propuesto la mayoría de los conjuntos de roles generales como parte de la teoría gramatical *Linking*, que se ocupa de describir la relación entre sintaxis y semántica, y que defiende que la representación sintáctica de los argumentos de un predicado es predecible a partir de la semántica. Existen dos enfoques: *Mono-stratal Frameworks* (correspondencia directa sintaxis-semántica) y *Multi-stratal Frameworks* (utilizan un nivel intermedio gramatical) (Giuglea y Moschitti, 2004).

La teoría Proto-Roles es la más abstracta y consideran dos roles: Proto-Agent, Proto-Patient. En (Guitart, 1998) se propone un conjunto de roles semánticos para el español: Causa, Tema, Locus.

(Fillmore 1971) propuso una gramática de casos que clasificaba los verbos en función de los marcos de casos o roles necesarios por un verbo. Uno de los elementos esenciales de la gramática era un pequeño conjunto de roles universal, es decir, válido para todos los idiomas y lo suficientemente genérico.

Sin embargo, los conjuntos de roles más específicos han sido propuestos por los ingenieros informáticos, más interesados en la representación sintáctica de los argumentos de ciertos verbos específicos. Estos roles se utilizan en la implementación de diversos sistemas de tratamiento automático del lenguaje, como los sistemas de extracción de información. Por ejemplo, para un sistema de información de vuelos, algunos roles específicos podrían ser: FROM_AIRPORT, TO_AIRPORT, DEPART_TIME. Otro tipo de roles aún más específicos son los dependientes del verbo: EATER, EATEN.

Por último, es importante resaltar la dificultad que implica la identificación de los roles semánticos. Su principal motivo es que no siempre existe una correspondencia directa entre la sintaxis y la semántica, produciéndose ambigüedad en la determinación del rol semántico que le corresponde a un determinado componente sintáctico.

3 Revisión de los principales recursos lingüísticos con información semántica

Para llevar a cabo la implementación de cualquier sistema de etiquetado de roles semánticos es imprescindible el uso de recursos que dispongan de información lingüística. Además, los sistemas que utilizan métodos de aprendizaje supervisado necesitan ser entrenados utilizando corpus etiquetados. Los recursos léxicos pueden utilizarse en su fase de aprendizaje, porque las relaciones semánticas entre las palabras están explícitamente identificadas.

FrameNet (Baker, Fillmore, Lowe, 1998) se basa en la teoría de los *frames* semánticos, donde un frame corresponde a un escenario que implica una interacción y sus participantes (roles). El nombre del frame sirve para identificar la relación semántica que agrupa los roles semánticos. Los roles (*frame elements*) propuestos por FrameNet son específicos de cada frame.

FrameNet incluye un corpus de oraciones del idioma inglés etiquetadas con roles semánticos. La principal ventaja de FrameNet es que su corpus proporciona una evidencia empírica para la representación sintáctica de las estructuras semánticas. El corpus puede ser utilizado en el aprendizaje de la identificación de las relaciones semánticas a partir de las estructuras sintácticas. Extrayendo del corpus los elementos sintácticos y sus correspondientes roles es posible construir automáticamente un conjunto de reglas que codifiquen las posibles representaciones sintácticas de los frames semánticos.

Su principal desventaja es que no define restricciones de selección semántica para los

roles semánticos. Estas restricciones de selección indican la categoría semántica a la que debe pertenecer el núcleo del argumento que se asocia a un rol. Además, tiene una limitada cobertura del lenguaje (3040 verbos) y su escalabilidad está seriamente limitada.

PropBank (Kingsbury, Palmer y Marcus, 2002) es un corpus de artículos de Wall Street Journal etiquetados con relaciones de argumentos – predicados. La anotación de PropBank se basa en la clasificación verbal de Levin (Levin, 1993), que supone que existe una fuerte conexión entre la sintaxis y la semántica. Los verbos se agrupan según su comportamiento sintáctico. Los grupos resultantes son coherentes desde el punto de vista semántico cuando todos los verbos de la misma clase comparten los mismos roles semánticos. Los grupos están formados en el nivel gramatical según el criterio de la alternación de diátesis. Los argumentos de PropBank son específicos de cada verbo.

VerbNet (Kipper, Dang y Palmer, 2000) es un lexicón de verbos con información sintáctica y semántica explícita, basado en la clasificación verbal de Levin. Como consecuencia, la principal hipótesis de VerbNet es que los frames sintácticos de un verbo (expresados como argumentos) son un reflejo directo de la semántica subyacente.

La principal ventaja de VerbNet es que ofrece fuertes generalizaciones del comportamiento sintáctico de los verbos. VerbNet define las relaciones sintáctico-semánticas de forma más explícita, etiquetando los roles temáticos y proporcionando restricciones de selección semántica para sus argumentos. Además. también guarda una correspondencia de cada verbo con sus posibles significado en WordNet. Además, tiene una mayor cobertura que FrameNet (4159 verbos frente a 3040 verbos de FrameNet; definidos en ambos recursos 2398). Como desventaja, su conjunto de roles temáticos es demasiado genérico.

3.1 El caso del español.

El principal objetivo de FrameNet Español (Rüggeberg, 2004) consiste en identificar las clases semánticas que configuran conceptualmente el léxico de predicados del español, determinar los argumentos semánticos que definen cada una de dichas clases y anotar semántica y sintácticamente oraciones en las que aparecen predicados de dichas clases, proporcionando un diccionario *online* de las características semánticas de los predicados de léxico español.

El corpus 3LB (Navarro et al. 2004) es un corpus para los sistemas de búsqueda de respuestas, formado por tres corpus para euskera, catalán y español. Contiene información sintáctica y semántica (*sense WordNet*) para cada verbo, nombre y adjetivo.

El conjunto de roles fue establecido basándose en fundamentos teóricos y como consecuencia de las necesidades de los sistemas de búsqueda de respuestas.. Los roles principales propuestos son: Agente-Causa Tema – Paciente, Beneficiario – Receptor y los adjuntos son Tiempo, Lugar y Modo.

El proyecto ADESSE proporciona una base de datos con información sintáctico-semántica sobre los verbos del español (García-*Miguel*, Costas y Martínez, 2003). El corpus puede ser utilizado como recurso para comprobar la correspondencia entre las funciones sintácticas y los roles.

Su clasificación semántica es una jerarquía de base conceptual, en la que se han reconocido un total de siete macro-clases (*Mental, Relación, Procesos Materiales, Comunicación, Existencia, Causativo, Dispositivos*), 14 clases (*Sensación, Percepción, Cognición, Atribución, Posesión, Espacio, Cambio, Hecho, Social, Comportamiento, Comunicación, Existencia, Causativo, Dispositivo*) y 52 subclases que dan cobertura a 1642 verbos.

Un proyecto similar es SENSEM base de datos léxica de los verbos del español en la que se ha descrito el comportamiento sintáctico-semántico de aproximadamente unos 1.100 predicados de esta lengua.

4 Revisión de los sistemas de etiquetado de roles semánticos

Uno de los trabajos más relevantes es el propuesto en (Gildea y Jurafsky, 2002), que ha sido la base para el desarrollo de sistemas posteriores y que plantea una aproximación estadística basada en el corpus FrameNet. El sistema tiene una arquitectura de dos capas: la identificación de los límites de los roles semánticos en la oración y la asignación de la etiqueta correcta. Esta arquitectura se utiliza en otros trabajos: (Pradhan et al. 2005), (Johansson y Nugues, 2005), etc.

Un conjunto de elementos sintácticos y léxicos se extraen del árbol sintáctico de cada

oración del corpus, y se utiliza para estimar la probabilidad de cada rol. Los elementos propuestos en (Gildea y Jurafsky, 2002) son: tipo sintagma, función gramatical, camino del predicado al componente, posición respecto al verbo, voz del verbo, núcleo del componente.

Los experimentos demostraron que los elementos más discriminatorios en la tarea de identificación eran el camino y el núcleo. Este último también mejoraba la tarea de clasificación, aunque generaba una gran dispersión de los datos.

El principal problema de este enfoque es que no hay una forma de cuantificar los efectos de los elementos, mientras que no haya duda de que los errores introducidos por el analizador sintáctico no afectan negativamente en los resultados obtenidos (precisión 65.0%, cobertura=61.0%).

Un trabajo posterior (Gildea y Palmer, 2002) analizaba si la información facilitada por los analizadores Collins y Charniak (Charniak, 2001) contribuye a resolver el problema de etiquetado de roles semánticos. Los experimentos demostraron que del 12% al 18% de los argumentos se pierden durante el análisis sintáctico.

En el sistema (Hacioglu y Ward, 2003) sólo se utilizó información sintáctica parcial. El sistema mejoraba la precisión (un 67.6% frente al 65.0% obtenido en Gildea y Jurafsky), pero su cobertura disminuía (55.9% frente al 61.0% de Gildea y Jurafsky).

Otro trabajo que también ha servido como base a otros sistemas, es el propuesto en (Surdeanu et al., 2003). Extiende el sistema (Gildea y Jurafsky, 2002) con elementos adicionales y utiliza el clasificador de árbol de decisión C5. Entre los nuevos elementos propuestos resaltan la palabra con más información del sintagma, la categoría morfosintáctica del núcleo y las entidades**.** Los experimentos demostraron que los elementos *entidad* y *la categoría morfosintáctica del núcleo* mejoraban los resultados obtenidos en la tarea de clasificación, en particular, para los argumentos adjuntos de localización y tiempo: ARGM-LOC y ARGM-TMP.

El sistema que ha conseguido los mejores resultados hasta el 2005 es el propuesto por (Pradhan et al, 2005) basado en el corpus PropBank. Extiende los trabajos (Gildea y Jurafsky, 2002), (Surdeanu et al., 2003) y mejora los resultados utilizando nuevos elementos, por ejemplo, *la clase semántica del verbo* y el *significado del verbo*. Se realizaron varios experimentos, combinando distintos elementos. La eliminación de elementos léxicos (núcleo y otros relacionados) provocaba un mayor detrimento en los resultados del sistema.

Los experimentos (Pradhan et al, 2005) demostraron que los resultados obtenidos al utilizar información sintáctica parcial son peores que si se utilizan analizadores sintácticos.

El sistema obtuvo una precisión del 84% y una cobertura del 75%. En la tabla 2, se comparan los resultados de los métodos SVM (Support Vector Machine), Lattice backoff, C5, utilizando el mismo conjunto de elementos.

Método	A
Lattice backof (Gildea y Jurafsky, 2002)	77%
C5 (Surdeanu et al., 2003)	79%
SVM (Pradhan et al, 2005)	87%

Tabla 2: Resultados en la fase de clasificación, utilizando el mismo conjunto de elementos

En (Brharati, Venkatapathy y Reddy, 2005), también se estima la secuencia de roles más probable utilizando un modelo de máxima entropía. Además, la subcategorización verbal de PropBank se utiliza en la inferencia de los argumentos semánticos. El sistema no maneja información sobre el significado del verbo y por este motivo, se utilizan todos los marcos del predicado, pudiéndose producir ambigüedad en la clasificación de los argumentos obligatorios.

Los experimentos demostraron que la subcategorización verbal ayudaba en la predicción de los roles (68.14% frente a un 67.03% sin utilizar la subcategorización verbal).

Otras técnicas de aprendizaje automático han sido aplicadas con éxito: Generative Model (Thompson, Levy y Manning, 2003), Sparce Bayesian Classification (Johansson y Nugues, 2005), Perceptron (Carreras, Màrquez y Chrupała, 2004), etc. La mayoría de los sistemas utilizan métodos de aprendizaje supervisado.

El sistema (Swier y Stevenson, 2004) utiliza un método no supervisado basado en una estrategia de "bootstrapping" y el conjunto de roles temáticos de VerbNet. Los elementos utilizados para estimar la probabilidad de un determinado rol son: predicado, la función gramatical del componente sintáctico y su núcleo. No se estimó ni la precisión ni la cobertura del sistema, en su lugar se realizó un análisis más detallado de los resultados basado

en el porcentaje de acierto de asignación. Se realizaron dos experimentos. En el primero, los argumentos candidatos fueron identificados por manualmente y se obtuvo un 90.1% de aciertos. En el segundo experimento, el sistema era el encargado de identificar los argumentos candidatos. En este caso, se obtuvo un 87.2% de aciertos.

Los recursos léxicos son información escasa pero muy valiosa. En (Shi y Mihalcea, 2005) se propone la integración de los recursos léxicos FrameNet, VerbNet y WordNet. Cada uno de estos recursos codifica un diferente tipo de conocimiento y su combinación podría tener como resultado un recurso más rico que permita un análisis semántico más robusto y preciso.

Su propuesta consiste en vincular los componentes semánticos de cada uno de los recursos: *frames, roles semánticos, clases semánticas*.

Existen pocos métodos automáticos para la clasificación semántica, debido principalmente a la falta de recursos anotados con información semántica. Esta escasez es aún mayor en los sistemas de etiquetado semántico de la lengua española. FrameNet Español puede favorecer el uso de enfoques estadísticos para la clasificación automática semántica para el idioma español.

5 Una propuesta de un sistema de etiquetado de roles semánticos.

Los resultados de un sistema de etiquetado de roles semánticos dependen en gran medida de la precisión de los analizadores sintácticos (Gildea y Palmer, 2002). El cuello de botella de los sistemas de anotación semántica está en el análisis sintáctico. Construir un árbol sintáctico es computacionalmente más costoso que segmentar una oración en grupos nominales y verbales..

GATE es una plataforma que dispone de varios tipos de recursos, entre ellos, *Noun Phrase Chunker (Segmentador de grupos nominales)* y *ANNIE VP Chunker (segmentador de grupos verbales)* y *Garzettee,* útil en la detección de entidades.

La propuesta consiste en utilizar el segmentador de GATE, para etiquetar roles semánticos y comparar los resultados obtenidos con otros analizadores, como Collins o Charniak.

Se ha analizado el corpus PropBank con la herramienta GATE. Para la fase de aprendizaje, se proponen los siguientes elementos para cada argumento: *Etiqueta Semántica (rol), Tipo de sintagma del argumento, Posición respecto al predicado, Núcleo del sintagma* (primera palabra)*, Categoría Morfosintática del Núcleo***,** *Entidad del núcleo* (obtenida de Gazetteerr; si la palabra no es una entidad, entonces *null*), *Preposición y POS que precede al sintagma* (*null* si no existe), *Predicado (*lema del verbo)*, Voz, Frame Sintáctico*, *Clase Verbal, Sense Verb.*

La clase verbal es un elemento importante, porque los corpus no cubren la totalidad de los verbos. Si el predicado (verbo) no está presente en el corpus, el sistema no será capaz de predecir sus roles. En estas situaciones, la información sobre la clase verbal del predicado puede ser útil y por ello interesa poder establecerla automáticamente. Si además se utiliza su significado (*sense*), el número de los posibles marcos (frames obtenidos de FrameNet, PropBank o VerbNet) del predicado disminuye y también la ambigüedad en la clasificación.

El núcleo, y otros elementos léxicos relacionados, han generado buenos resultados en la tarea de clasificación (Gildea y Jurafsky, 2002), (Pradhan et al., 2005), pero se produce una gran dispersión en los datos, provocando ruido en la clasificación. Por este motivo, es importante utilizar la *clase semántica* del núcleo. En la actualidad, sólo se han recuperado las clases semántica de los nombres. Para cada nombre, se asciende en la jerarquía de WordNet. Los conceptos más abstractos de la jerarquía (*entity, psychological_feature, abstraction, state, event, act, group, possession, phenomenon*) se han considerado demasiado generales para el propósito de este trabajo. Por este motivo, se asciende hasta llegar a un concepto del nivel 1, es decir, hasta llegar a algún hipónimo (*hijo*) de los conceptos abstractos. WordNet es un recurso muy granulado (en el nivel 1 existen 137 conceptos distintos). Este número tan elevado de posibles clases candidatas podría seguir generando ruido en la clasificación.

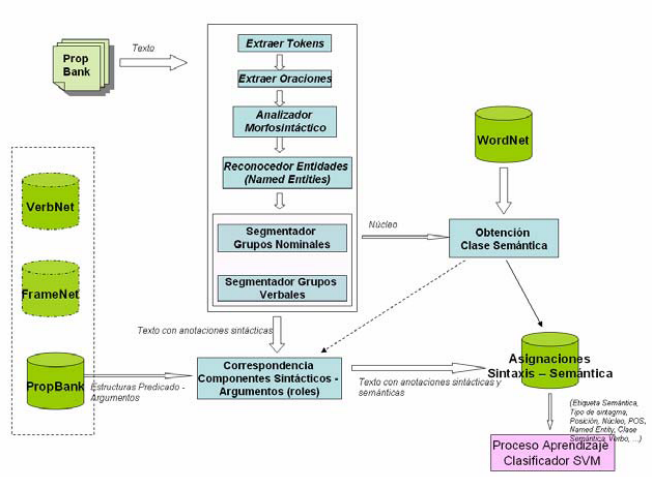

Figura 1. Arquitectura de la fase de aprendizaje del clasificador.

Para reducir el número de conceptos del nivel 1, se han implementado las reglas propuestas en el trabajo de (Mihalcea y Moldovan, 2001). Este trabajo propone una metodología para obtener una nueva versión de WordNet colapsando los conceptos (*synsets*) similares en significado y eliminando conceptos poco probables. Se ha conseguido reducir el conjunto de posibles clases (conceptos nivel 1) a 88. Aproximadamente, para el 25% de los nombres, utilizando la versión generada de WordNet, no se obtiene una clase semántica. Esto puede ser debido principalmente, a que el 73% de los nombres no clasificados son nombres propios.

En la figura 1, se muestra la arquitectura necesaria para entrenar el clasificador. El corpus de PropBank es procesado utilizando un conjunto de procesos que proporciona GATE. En primer lugar, el texto es dividido en *tokens* y son extraídas sus oraciones, se reconocen y etiquetan las entidades. A continuación, un segmentador obtiene los grupos oracionales de cada oración. Para los grupos nominales, se extrae su núcleo y su clase semántica a partir de WordNet, como se explicó anteriormente en esta sección.

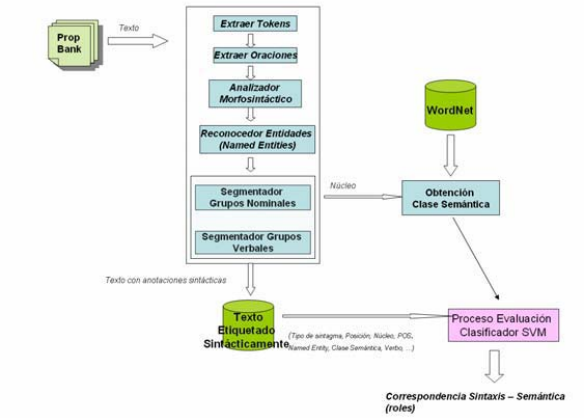

Figura 2. Arquitectura de la fase de validación del clasificador.

El siguiente paso, consiste en realizar una correspondencia entre los componentes sintácticos de las oraciones y sus etiquetas semánticas (*roles* o *argumentos*). PropBank proporciona para cada predicado (verbo) un conjunto de posibles estructuras de argumentos (*roles*) – predicado. Estas estructuras (*marcos* o *frames*), son comparadas con las oraciones anotadas con información sintáctica; como resultado a cada componente sintáctico, siempre que se corresponda con un argumento, le es asignado una etiqueta semántica (*rol*).PropBank no define restricciones de selección semántica para sus argumentos, pero éstas podrían obtenerse fácilmente, ya que PropBank y VerbNet se apoyan en la misma clasificación verbal, y ambos recursos podrían ser vinculados como se propone en el trabajo de (Giuglea y Moschitti, 2004). En este caso, la clase semántica del núcleo sería útil a la hora de determinar la correspondencia entre los componentes sintácticos y los argumentos semánticos de PropBank.

Una vez que el texto está anotado con información sintáctica y semántica, el clasificador puede ser entrenado.

En la figura 2, se muestra la arquitectura necesaria para la fase de validación. Durante esta fase, otra sección distinta del corpus es procesada. El clasificador entrenado en la fase anterior, recibe como entrada los elementos sintácticos y léxicos extraídos del texto anotado con información sintáctica y de la clase semántica obtenida a partir de WordNet, con el objeto de asignar a cada uno de los componentes sintácticos una etiqueta semántica (*argumento* o *rol*).

5.1 Ampliación de la propuesta al caso del español.

Existen pocos métodos automáticos para la clasificación semántica, debido principalmente a la falta de recursos anotados con información semántica. Esta escasez es aún mayor en los sistemas de etiquetado semántico de la lengua española.

FrameNet Español puede favorecer el uso de enfoques estadísticos para la clasificación automática semántica para el idioma español.

Los resultados podrían compararse con los obtenidos al utilizar en la fase de validación el corpus 3LB, pero en este caso sería necesario buscar alguna correspondencia entre el conjunto de roles FrameNet Español y el del corpus 3LB.

Para los elementos como la clase semántica del verbo, su sentido o su subcategorización sería necesario emplear recursos como los proporcionados por el proyecto ADESSE, la base de datos verbal SENSEM, FrameNet Español o EuroWordNet. La clase semántica de los elementos léxicos, como el núcleo del sintagma, podría obtenerse a partir de EuroWordNet.

6 Conclusión y Trabajo Futuro

La identificación de los roles semánticos es una parte crucial en la interpretación de los textos, y por tanto es importante para muchas tareas del Procesamiento del Lenguaje Natural.

Esta propuesta considera la identificación de roles semánticos como un primer paso hacía un enfoque más ambicioso, la detección de conceptos e interrelaciones entre ellos a modo de ontología que represente el contenido semántico de un documento. Por ello es necesario disponer de una plataforma que facilite la integración de distintos recursos.

En la fase de análisis sintáctico, los errores introducidos por los analizadores, Collins o Charniak, influyen negativamente en los resultados del etiquetado pues entre el 12 y el 18% de los argumentos se pierden. El uso de los segmentadores es menos costoso y además, es una opción para los idiomas que no dispongan de analizador.

Se han expuesto distintas técnicas probabilísticas y de aprendizaje automático. El método que ha obtenido los mejores resultados es SVM (Pradhan et al, 2005). Los sistemas que utilizan técnicas de aprendizaje supervisado necesitan corpus como FrameNet o PropBank. Estos recursos son escasos y costosos de producir.

El trabajo actual propone el uso de un nuevo elemento, la clase semántica del núcleo del componente, que se espera que reduzca el ruido en la fase de clasificación. Actualmente, dentro de la arquitectura de la fase de aprendizaje se ha implementado la parte del sistema que se ocupa de anotar el texto con información sintáctica, extraer la clase semántica del núcleo de los grupos oracionales, y realizar la correspondencia entre la sintaxis y la semántica, a partir, de las estructuras de argumentos – predicados proporcionadas por PropBank.

El siguiente paso en nuestro trabajo consiste en entrenar el clasificador y realizar una serie de experimentos.

También sería interesante utilizar un corpus distinto, como FrameNet, e incluso combinar varios recursos como se propone en los trabajos de (Shi y Mihalcea, 2005), (Giuglea y Moschitti, 2004).

En la lengua española, los trabajos sobre etiquetado de roles semánticos como FrameNet Español, y otros recursos semánticos, pueden favorecer el uso de enfoques estadísticos en el etiquetado de roles semánticos para el idioma español.

Bibliografía

Baker, C. F., C. J. Fillmore, J. B. Lowe. 1998. The Berkeley Framenet Project. *COLING 98,* pág. 86-90.

Brharati, A., S. Venkatapathy, P. Reddy. 2005. Inferring semantic roles using sub-categorizacion frames and maximum entropy model. *CoNLL'2005 Shared Task*.

Carreras, X., L. Màrquez, G. Chrupała. 2004. Hierarchical Recognition of Propositional Arguments with Perceptrons. *CoNLL-2004.*.

Charniak, E. 2001. Immediate-head parsing for language models. *ACL'01.*

Fernández, A. G., I. Vázquez. 2004. Sensem: base de datos verbal del español"., *IBERAMIA 2004*.

Fillmore, C. J. 1971. Some problems for case grammar. En R. J. O'Brien, editor, 22nd annual Round Table. Linguistics: developments of the sixties – viewpoints of the seventies.

García-Miguel, J. M., L. Costas, S. Martínez. 2003. Diátesis verbales y esquemas construccionales Verbos, clase semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. En *VI Congreso Internacional de Lingüística Hispánica.*.

Gildea, D. y D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computacional Linguistics*, 28(3):245-288.

Gildea, D. y M. Palmer. 2002. The Necessity of Parsing for Predicate Argument Recognition. *ACL 2002*.

Giuglea, A. M. y A. Moschitti. 2004. Knowledge Discovering using FrameNet, VerbNet and PropBank. En *Proceedings of the Workshop on Ontology and Knowledge Discovery at ECML 2004,*.

Guitart, J. M. 1998. El caso gramatical en español en la teoría de los roles semánticos. Editorial Runasimi.

Hacioglu, K. y W. Ward. 2003. Target Word detection and semantic role chunking using support vector machines. *Human Language Technology Conference 2003*

Johansson, R., P. Nugues. 2005. Sparce Bayesian Classification of Predicate Arguments. *CoNLL'2005 Shared Task*.

Kingsbury, P., M. Palmer, M. Marcus. 2002. Adding semantic annotation to the Penn Treebank. *Human Language Technology Conference 202*.

Kipper, K., H. T. Dang, M. Palmer. 2000. Class-Based Construction of a Verb Lexicon. *AAAI-2000,*.

Levin, B. 1993. English Verb Classes and Alternations: A Preliminary Investigation. The University of Chicago Press.

Mihalcea, R., y D. I. Moldovan. 2001. Automatic generation of a coarse grained WordNet. *NAACL Workshop on WordNet and Other Lexical Resources 2001*.

Navarro, B., P. Moreda, B. Fernández, R. Marcos y M. Palomar. 2004. Anotación de roles semánticos en el corpus 3LB. Herramientas y Recursos Lingüísticos para el Español y el Portugués. IBERAMIA 2004.

Pradhan, S., K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, D. Jurafsky. 2005. Support Vector Learning for Semantic Argument Clasification. *Machine Learning,* 60, 11-39.

Rüggeberg, C. S. 2004. FrameNet Español. Una red semántica de marcos conceptuales. *En Serra, E. y G. Wotjak (eds),* páginas *182-196.*http://gemini.uab.es/SFN/

Shi, L. y R. Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing, En *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, Méjico.

Surdeanu, M., S. Harabagiu, J. Williams, P. Aarseth. 2003. Using predicate-argument structures for information extration. ACL 2003.

Swier, R. S. y S. Stevenson. 2004. Unsupervised Semantic Role Labelling. *EMNLP 2004*.

Thompson, C. A., R. Levy, C. D. Manning. 2003. A Generative Model for Semantic Role Labeling. *ECML 2003*.