

Adquisición de subcategorizaciones verbales mediante un clasificador automático

Laura Alonso Alemany
Universidad de la República
Universidad Nacional de Córdoba
alemany@famaf.unc.edu.ar

Irene Castellón Masalles
Universidad de Barcelona
icastellon@ub.edu

Nevena Tinkova Tincheva
Universidad de Barcelona
nevenatinkova@ub.edu

Resumen: En este artículo presentamos un método para asignar automáticamente patrones de subcategorización a piezas verbales no conocidas previamente, partiendo de una generalización de patrones anotados manualmente.

A partir del banco de datos SENSEM (Fernández et al 2004) se han adquirido los esquemas de subcategorización de 1161 sentidos. Estos esquemas se han agrupado en clases mediante técnicas de clustering. Cada clase representa una generalización sobre el comportamiento sintáctico-semántico de los verbos que contiene. Nuestro objetivo final es enriquecer un lexicon verbal con marcos de subcategorización, asignando automáticamente cada pieza verbal a una de estas clases, a partir de ejemplos de corpus anotados automáticamente. Presentamos una evaluación preliminar de un clasificador que lleva a cabo esta tarea.

Palabras clave: Adquisición de subcategorización, análisis sintáctico, clases sintácticas, sentidos verbales.

Abstract: In this paper we present a method for automatically assigning subcategorization frames to previously unseen verb senses of Spanish, starting from a generalization of manually annotated frames.

Taking as a departure point the data base SENSEM (Fernández et al 2004), the subcategorization frames of 1161 senses have been acquired. These frames have been grouped in classes by clustering techniques. Each class represents a generalization over the syntactico-semantic behaviour of the verbs in it. Our final target is to enrich a verbal lexicon with subcategorization frames, automatically assigning each verbal piece to one of these classes based on examples from corpus that have been automatically analyzed. We present a preliminary evaluation of a classifier that carries out this task.

Keywords: Acquiring verb subcategorizations, parsing, syntactic classes, verb senses.

1 Introducción

En este artículo presentamos un método para enriquecer un léxico verbal con información de subcategorización de forma semiautomática, extrapolando la información de un corpus anotado a mano a ejemplos sin anotación previa.

Partimos del corpus anotado a mano SENSEM (Fernández et al 2004), y caracterizamos los verbos que en él aparecen tomando como propiedades los esquemas sintácticos en los que ocurren. Después generalizamos el comportamiento de estos verbos mediante técnicas de clustering. Así obtenemos clases de verbos con comportamientos sintácticos similares, ya que en un mismo cluster se agrupan verbos que ocurren en un conjunto de esquemas sintácticos parecido.

Analizamos diferentes opciones para obtener estas clases de verbos similares: diferentes subconjuntos de propiedades para describir a los verbos, diferentes técnicas de clustering. Aplicamos métricas cuantitativas y cualitativas para analizar las diferentes soluciones obtenidas, y finalmente optamos por estudiar con más detalle dos soluciones: la clasificación en 3 clases, a partir de categorías y funciones, y la clasificación de dos niveles que consta de 5 clases iniciales y 16 clases en un segundo nivel. Se ha evaluado la utilidad de esta solución para asignar una clase de comportamiento sintáctico a piezas verbales desconocidas mediante diferentes experimentos con clasificadores aprendidos automáticamente.

El resto del artículo está organizado de la siguiente manera. En la próxima sección se argumenta la utilidad de la información de subcategorización para la mejora del análisis

sintáctico automático, analizamos algunos trabajos relacionados y exponemos nuestra aproximación. En la sección 3 presentamos el método para caracterizar los ejemplos del corpus, los parámetros de las diferentes soluciones de clustering y las métricas para evaluarlas, así como una breve descripción de las mejores soluciones obtenidas. En la sección 4 describimos las clases de la solución que hemos elegido como óptima. En la sección 5 evaluamos la aplicación de las clases seleccionadas a ejemplos no vistos, mediante clasificadores aprendidos automáticamente. Finalmente, en la sección 6 presentamos las conclusiones de este trabajo y el esquema de trabajo futuro.

2 Motivación: la subcategorización y el análisis sintáctico

La descripción del funcionamiento de una pieza verbal tanto a nivel sintáctico como semántico es una tarea necesaria para abordar la 'comprensión' del lenguaje en el área del procesamiento del lenguaje natural. Por un lado, el verbo es el núcleo semántico de la oración, es decir, el que distribuye papeles semánticos y por lo tanto, el que da sentido concreto a los elementos nominales. Por otro, desde una perspectiva puramente sintáctica, el verbo nos informa sobre el tipo de complementos que precisa y si este esquema alterna o no con otros complementos, es decir, sobre las diferentes configuraciones sintácticas de los argumentos. De esta manera, la estructura de subcategorización se puede considerar como la información lingüística básica que posibilita la restricción del número de estructuras obtenidas en el análisis sintáctico. Esta información es crucial para dicho análisis, ya que hay problemas fundamentales para la buena resolución del análisis sintáctico cuyo comportamiento depende de la idiosincrasia de los núcleos léxicos.

Entre los casos más complejos de resolución son el establecimiento de la dependencia de un sintagma preposicional (1), la resolución de la coordinación (2) o la determinación de la función de determinados sintagmas nominales (3). A estos problemas se añaden para el español el grado de libertad de los constituyentes (4), haciendo que los casos anteriores sean más difícil resolución. Así, conocer la subcategorización del verbo permite evitar la mala identificación de categorías.

- (1) *Y lo haremos defendiendo las libertades y los derechos ciudadanos [en el combate contra sus enemigos].*
- (2) *... armaba sus modelos con pedazos de cartón, tablitas, goma, engrudo, cartulinas y lápices de colores.*
- (3) *Macri anuncia esta tarde su postulación a jefe de gobierno.*
- (4) *Papel fundamental han desempeñado en esta recuperación los evangelios llamados apócrifos, sobre todo los de carácter gnóstico.*

2.1 Trabajo Previo

Se han propuesto diversas soluciones que pasan por establecer modelos ya léxicamente determinados. Algunos autores (Atserias 2006) proponen disponer de dos modelos, uno nominal y otro verbal para que en base a determinadas condiciones disputen por el sintagma preposicional. En esta línea, una forma de solucionar este tipo de problemas es el establecimiento ya a nivel léxico de las preferencias de combinación de unidades verbales. Una de las informaciones que más se ha intentado establecer son los esquemas de subcategorización, una modelización del comportamiento sintáctico de los núcleos léxicos. La adquisición automática de dicha información ha sido tratada por diferentes autores (Korhonen 2002, Briscoe et al 1997) en general partiendo de un corpus analizado a nivel sintáctico automáticamente (Korhonen et al 2003, Briscoe et al 1997) o manualmente (Sarkar et al 2000) y aplicando determinados filtros para no contemplar información de adjuntos. Estos trabajos han tenido un acierto de diferente grado.

2.2 Nuestra Aproximación

A diferencia de estos trabajos nuestro sistema parte de una serie de patrones ya adquiridos y evaluados dentro del proyecto SENSEM (ver Figura 1). Nuestro objetivo final consiste en asociar esquemas de subcategorización a sentidos verbales que no están en SENSEM. Para ello clasificamos los nuevos predicados verbales dentro de una de las clases de subcategorización inducidas a partir de los verbos en el corpus.

El banco de datos de SENSEM está compuesto por un corpus anotado a nivel sintáctico-semántico (Castellón et al 2006). La anotación ha consistido en etiquetar en forma manual el verbo y los constituyentes directamente relacionados con él, donde cada constituyente se anota mediante: la categoría morfosintáctica (p.ej.: sintagma nominal, oración adverbial), la función sintáctica (p.ej.: sujeto, objeto preposicional), su relación con el verbo (p.ej.: argumento o adjunto), y el rol semántico (p.ej.: iniciador, tema afectado, origen, tiempo). El total de lemas tratados es de 250, seleccionados por su frecuencia en un corpus equilibrado de la lengua (Davies 2005), y el número de sentidos es de 1161.

Para llegar a este objetivo final, la adquisición de esquemas de subcategorización, partimos de una serie de presupuestos que creemos necesario exponer. En primer lugar partimos de la hipótesis de que la subcategorización es una información asociada a los sentidos verbales, no a los lemas. Aunque este presupuesto parece evidente, en los trabajos sobre adquisición de subcategorizaciones no siempre se ha asumido de esta manera (Manning 1993, Korhonen 2002). Así, para aplicar el clasificador sobre corpus será necesario disponer de alguna aplicación de algún tipo de desambiguación de sentidos. Por otro lado, representamos dichos esquemas mediante categorías sintácticas, funciones sintácticas y papeles semánticos. Por lo que hacemos uso de esta información si disponemos de ella y mejora los resultados, sin embargo no pretendemos adquirir toda esta información sino la puramente sintáctica (categorías y en algunos casos funciones). Una de nuestras hipótesis de partida es que en la base de datos sensem ya existen la mayoría de los esquemas de subcategorización existentes en español, por lo que nos parece razonable pensar que cada nuevo sentido verbal que veamos puede realizarse como alguno de los verbos ya conocidos. Además creemos que podemos adquirir la subcategorización comparando el comportamiento verbal de un sentido no conocido con el comportamiento de todos los verbos conocidos clasificados.

3 Metodología

El objetivo inicial, como hemos dicho, consiste en inducir clases de comportamiento sintáctico de los verbos a partir de la información de

SENSEM y extrapolar estos comportamientos a verbos desconocidos. A continuación describimos las fases del experimento.

3.1 Caracterización de los ejemplos

El procedimiento que seguimos se basa en los resultados de la anotación de SENSEM y de los datos que podemos etiquetar de forma automática. Por un lado, disponemos de unos 1000 sentidos anotados sintáctica y semánticamente, de los que obtenemos esquemas de realización (ejemplos) y esquemas de subcategorización (patrones). Por otro lado, disponemos del análisis sintáctico de Freeling.

En primer lugar, para cada ejemplo del corpus, se extrae el esquema sintáctico y se realiza una generalización mediante la compactación de ciertas categorías que tienen la misma distribución, como por ejemplo los pronombres relativos (de sujeto u objeto directo) o los sujetos elididos con los sintagmas nominales, entre otros. Posteriormente se procede a la eliminación de los adjuntos y por tanto la selección de los constituyentes marcados en el corpus como argumentales. Por último se compactan en uno aquellos esquemas que únicamente difieren por el orden de los constituyentes. De esta forma obtenemos el esquema de subcategorización asociado a los sentidos verbales de la base.

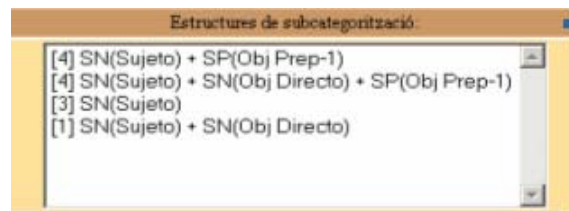


Figura 1. Esquemas de subcategorización adquiridos para el sentido *añadir_1*.

Hemos caracterizado los ejemplos con diferentes subconjuntos de datos:

- categoría morfosintáctica de argumentos;
- categoría y función sintáctica;
- categoría, función y rol semántico.

Además, observando los resultados se evidenció que los datos con pocas ocurrencias en corpus introducían mucho ruido en el espacio de búsqueda, causando agrupaciones extrañas. Así decidimos trabajar también con subconjuntos de los esquemas de subcategorización, utilizando como atributos los esquemas con más de 5 o con más de 10

ocurrencias en el corpus, como se ve en la Tabla 1.

	todos	> 5 ocs.	> 10 ocs.
cat	240	98	69
func + cat	785	213	130
rol + func + cat	2854	464	317

Tabla 1: Número de patrones sintácticos distintos encontrados en el corpus al caracterizar los ejemplos con diferentes aproximaciones.

3.2 Inducción de clases de verbos

A partir de los patrones sintácticos de los ejemplos del corpus, con sus diferentes caracterizaciones, realizamos una clasificación no supervisada (*clustering*) con el fin de inducir clases de sentidos motivadas lingüísticamente. Para ello, realizamos una evaluación de diversos métodos de clustering, utilizando diferentes caracterizaciones de los ejemplos del corpus información y número de clases, para determinar la mejor solución. Para la clasificación utilizamos los algoritmos de clustering proporcionados por Weka (Witten et al 2005). Específicamente, elegimos Simple KMeans (Hartigan et al 1979) y el clustering basado en Expectation-Maximization (EM) (Dempster et al 1977).

Los métodos para evaluar las soluciones de clustering se describen en (Alonso et al 2007). Se trata de una combinación de inspección cualitativa de las clases obtenidas y las siguientes métricas sobre las soluciones:

- Dada una lista de **parejas de verbos** muy similares creada a mano, observamos si se agrupan en las mismas clases (bonificado) o no (penalizado).
- Índice de **solapamiento de los esquemas** que caracterizan a las diferentes clases.
- **Distribución de la población** en las clases, penalizando soluciones con clases con poca población (uno o dos sentidos), ya que no generalizan comportamientos.
- Índice de **distinguibilidad de sentidos**, que indica si los distintos sentidos de un lema verbal se distribuyen en distintos clusters o en los mismos.
- Acierto de **clasificadores** aprendidos sobre las clases obtenidas en cada solución de clustering, evaluados en el mismo corpus de aprendizaje mediante *ten-fold cross validation*.

Además, en muchas soluciones obtuvimos una clase mayoritaria que contenía verbos con

muy distintos comportamiento, típicamente, verbos que comparten algún esquema de subcategorización muy frecuente. Si intentamos aumentar el número de clusters que se pedía al método de clustering (ya fuera EM o KMeans), se producía una distribución muy irregular de la población. Esto nos llevó a investigar una forma de clustering jerárquico partitivo: aplicamos clustering dentro de la población de las clases obtenidas por cada solución, para poder establecer más clases con menor población y más específicas en cuanto a los esquemas de subcategorización. Esta aproximación resultó adecuada para obtener clases con población bien distribuida.

4 Selección de una solución óptima

4.1 Descripción general de las soluciones

En esta sección describimos sucintamente las soluciones de clustering obtenidas con diferentes criterios para caracterizar los sentidos verbales, para motivar la elección final de una de ellas. Las figuras 2 y 3 proporcionan una perspectiva general de diferentes medidas de evaluación para estas soluciones.

En general, el método KMeans, que necesita un parámetro especificando el número de clases que se quieren establecer, proporcionaba peores resultados que EM. En concreto, tendía a proporcionar clases con un solo sentido verbal en las soluciones que proponían más de tres clases. En las soluciones con tres o menos clases el índice de solapamiento de esquemas y el test de parejas resultaban considerablemente peor que para EM. Por esa razón optamos por EM como método para obtener las soluciones de clustering.

Una vez decidimos que EM sería nuestro método, inspeccionamos las soluciones obtenidas con diferentes tipos de información.

Las clases en las soluciones con roles distinguen claramente tipos distintos de marcos de subcategorización, especialmente las soluciones en las que sólo se tienen en cuenta los esquemas que ocurren más de 5 o 10 veces. También es destacable que en el caso de las soluciones con roles, usar sólo patrones frecuentes contribuye sensiblemente a mejorar el aprendizaje de los clasificadores automáticos, como puede verse en la Figura 3. Esto se debe principalmente a una notable reducción en la escasez de datos (*data sparseness*) cuando usamos sólo esquemas que ocurren en un buen

número de casos. En estas soluciones encontramos siempre 4 clases, una mayoritaria donde claramente encontramos los verbos con prácticamente cualquier patrón de argumentos pero con una importante presencia de diátesis intransitivas, que se producirían por la elisión de alguno de los argumentos en los ejemplos de corpus, junto con verbos propiamente intransitivos; una segunda clase bastante grande con verbos fuertemente caracterizados como transitivos, con pocas diátesis intransitivas; y dos clases pequeñas con verbos con algún argumento de tipo *Origen*, *Destino* o similar, también con pocas diátesis intransitivas.

Por lo que respecta a las soluciones donde los verbos están caracterizados mediante categoría y función, se distingue en todos los casos una clase con más de la mitad de la población, que contiene verbos con comportamientos muy dispares, con el rasgo común de contar con alguna diátesis intransitiva, probablemente causada, como en el caso de las aproximaciones con roles, por la elisión de alguno de los argumentos. Se suele distinguir también claramente una o más clases de verbos con algún argumento preposicional o adverbial, y también una clase con verbos ditransitivos y sus diátesis transitivas e intransitivas.

El método EM encuentra 2 clases para la aproximación con todos los esquemas. Como dos clases no nos proporcionan suficiente distinguibilidad de sentidos, decidimos aplicar aumentar la granularidad de las clases de dos formas: parametrizando el número de clases manualmente para el método de clustering, y realizando clustering dentro de la población de cada uno de los clusters.

Finalmente, las soluciones donde los ejemplos están caracterizados mediante categoría únicamente tienen una tendencia a producir muchas clases, pero la población se encuentra bien distribuida en clases de tamaño mediano, excepto en la solución que tiene en cuenta todos los esquemas. En las soluciones con patrones que ocurren más de 5 y más de 10 veces, se encuentra siempre una clase con la mayor parte de la población, dos clases medianas y un número variable de clases más pequeñas con. Resulta difícil generalizar el comportamiento de los verbos de estas clases por la gran ambigüedad de los patrones basados

únicamente en categorías. A la inversa que en el caso de las soluciones con roles, los clasificadores aprenden peor cuando se descartan esquemas que ocurren menos de 5 o de 10 veces en el corpus. Se puede interpretar que, dado que esta aproximación tiene mucha menos variedad de esquemas (ver Tabla 1), eliminar alguno de ellos conlleva una importante pérdida de información.

A partir de los resultados y comparando las diferentes medidas, finalmente se optó por tomar algunas de las clases de las soluciones de clustering que utilizan información de categoría y de función sintáctica. Esta decisión vino parcialmente condicionada por la caracterización de los verbos a los que se pretende asignar una clase de forma automática en última instancia. Los ejemplos de estos verbos podrán ser analizados automáticamente a nivel sintáctico, pero no al nivel de th-roles. Por este motivo intentamos prescindir de las clases obtenidas con información de roles semánticos, a pesar de que las clases obtenidas tienen un importante interés lingüístico.

Tomamos pues como punto de referencia las soluciones con funciones y categorías en 3 clases y todos los esquemas de subcategorización y con 5 clases, obtenida con los esquemas con más de 10 ocurrencias en corpus. Estas soluciones se estudiaron en más profundidad desde el punto de vista lingüístico, tal como describimos en detalle en la siguiente sección.

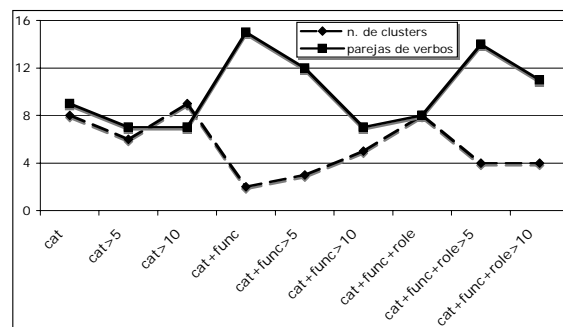


Figura 2. Número de clusters y de parejas de verbos similares clasificados en el mismo cluster en las diferentes aproximaciones de clustering, con diferentes tipos de información y filtro de argumentos.

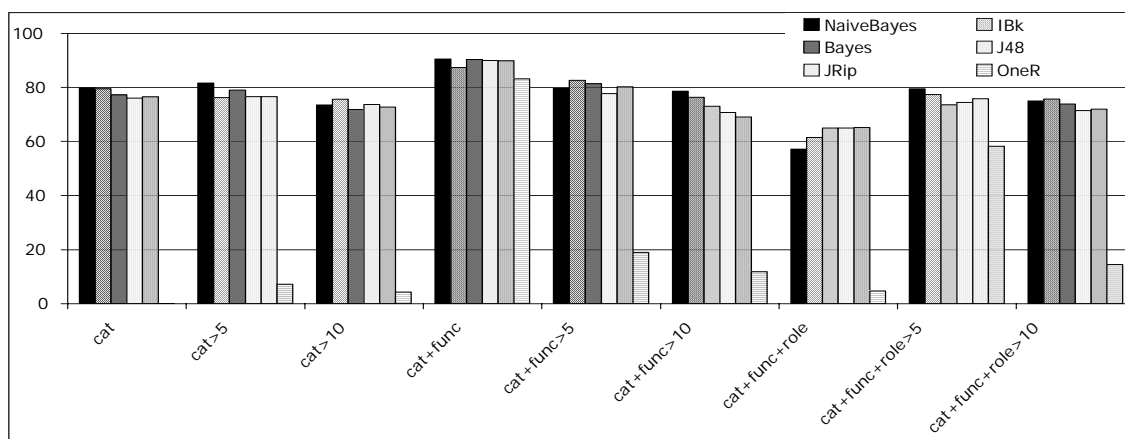


Figura 3. Funcionamiento de diferentes clasificadores a partir de clases de soluciones de clustering con diferentes subconjuntos de atributos.

4.2 Soluciones seleccionadas

4.2.1 Función + categoría en 3 clases

Clase 1: una clase mixta que contiene 704 sentidos verbales. Contiene un 83% de sentidos verbales que presentan alternancias en sus patrones. El 16% de sentidos que presentan un único patrón sin alternancia se realizan tanto de forma intransitiva, como transitiva o preposicional. En el caso de los sentidos que sí aceptan alternancias diatéticas, la omisión de complementos es la causa más frecuente de cambio de patrón, así nos encontramos con transitivos y ocurren también en su forma intransitiva, ditransitivos que alternan con esquemas transitivos o preposicionales que lo hacen con esquemas intransitivos. En definitiva, existe cierta dificultad para caracterizar esta clase por los diferentes comportamientos que comprende.

Por la naturaleza mixta de esta clase, se decidió aplicar clustering para distinguir subclases dentro de su población. El método EM proporciona un óptimo de 5 clases, en las que se vuelve a encontrar una clase mayoritaria con esquemas de subcategorización muy dispares y diversas clases más pequeñas que agrupan verbos con comportamientos mucho más definidos.

Clase 2: 153 verbos con mayoría de esquemas transitivos. La presencia alta de intransitivas y de ditransitivas se justifica por las alternancias diatéticas. En un 98% la realización de estos sentidos verbales incluye objetos directos de los cuales un 46% son oraciones completivas, alternando con omisiones de dicho complemento produciendo un patrón intransitivo. Hay una pequeña presencia de

complementos circunstanciales (12,58%) y también son frecuentes los preposicionales en un 55% de los sentidos verbales, de los cuales un 49% corresponden a ditransitivos.

Clase 3: 39 sentidos verbales, principalmente con esquemas preposicionales y verbos atributivos, donde las diátesis se caracterizan por la elisión de algunos o todos los preposicionales. En esta tercera clase nos encontramos con sentidos que tienen un alto porcentaje de realización con predicativos, circunstanciales y atributos que se realizan mediante sintagmas preposicionales, adjetivales y adverbiales. Estas realizaciones corresponden al 58% de alternancias de la clase. Otras alternancias se deben a la elisión o omisión de argumentos, concretamente de este otro grupo de alternancias, un 82% corresponde a omisiones de argumentos preposicionales.

Como podemos observar, las clases resultantes están muy delimitadas en dos casos (clase 2 y 3) pero la clase 1 es una clase mixta sin una caracterización clara, con el 70% de los sentidos tratados.

4.2.2 Función + categoría con esquemas que ocurren > 10 veces, en 5 clases

Dada la gran compacidad de esta solución, aplicamos clustering dentro de todas las clases, con ánimo de observar si era posible obtener clases más granulares dentro de la misma aproximación.

La clase más grande (clase 5) está compuesta por sentidos verbales que alternan entre esquemas transitivos e intransitivos y en algún caso con preposicionales. Observando su descripción esta clase es muy similar a la segunda clase de la solución en tres clases aunque con un mayor número de sentidos, un

total de 477. Las subclases obtenidas a partir de ésta, están mucho más caracterizadas, las clases 5.5, 5.3 y 5.2 agrupan los sentidos que alternan entre esquemas transitivos e intransitivos, las clases 5.4, 5.6, 5.7 y 5.8 se caracterizan por la alternancia intransitivo/preposicional, con alguna diferencia por la aparición de predicativos o de esquemas transitivos. A este nivel la asociación de una clase a esquemas como *sn v sn* o *sn v sp* parece bastante asumible.

La segunda clase (clase 2) está compuesta por 163 sentidos en los que predominan realizaciones preposicionales e intransitivas que se justifican por la omisión de los argumentos preposicionales. En algún caso encontramos esquemas ditransitivos alternantes con preposicionales. Las subclases obtenidas son muy similares entre ellas exceptuando la presencia en una de esquemas ditransitivos (2.2) y la ausencia en la otra, que se caracteriza por contener esquemas con circunstanciales (2.1).

Las dos siguientes clases (clase 1 y clase 3) están integradas por un total de 103 sentidos y 68, respectivamente, caracterizados por alternancias transitiva/ditransitiva/intransitiva, con omisiones de ciertos constituyentes. Estas clases no presentan subclases.

La última clase, la clase 4, contiene sentidos caracterizados por esquemas básicamente preposicionales alternantes con intransitivos y con la presencia de atributos. Las tres subclases que contiene están diferenciadas por diversos esquemas. 4.1 se caracteriza por la alternancia preposicional/intransitiva con atributos frecuentemente, la clase 4.2 es totalmente preposicional y en la clase 4.3 se clasifican sentidos con esquemas transitivos alternantes con preposicionales.

Desde una perspectiva comparada, parece que esta segunda aproximación, aunque crea subclases un poco redundantes, no tiene clases tan mixtas como la primera. Además, resulta más natural asociar un marco de subcategorización a cada una de las clases de la segunda aproximación.

5 Evaluación para aplicación final

Recordemos que el objetivo final de este trabajo es asignar una clase verbos no vistos, a partir de ejemplos de corpus analizados automáticamente. Para evaluar la utilidad de la solución de clustering seleccionada analizamos el corpus SENSEM automáticamente con

Freeling (Carreras et al 2004). En esta anotación automática no se filtran de ninguna forma los adjuntos, pero el alcance del verbo se delimita manualmente. También hemos experimentado las diferencias de distinguir sentidos, con lo que se pueden agrupar todos los ejemplos de un sentido, o bien tratar cada ejemplo individualmente, tanto para el corpus anotado manualmente como automáticamente.

Los resultados se pueden ver en las Figuras 4 y 5, para las soluciones de clustering de primer nivel y de segundo nivel, respectivamente. Vemos que los resultados para el segundo nivel son en general más bajos que para el primer nivel, probablemente porque los datos disponibles para esas clases, con menos población, son más escasos. En el primer nivel observamos que las aproximaciones con el corpus anotado automáticamente proporcionan peores clasificadores que cuando el corpus está anotado manualmente, pero observamos que la mayor diferencia se produce entre las aproximaciones que distinguen sentidos y las que se basan únicamente en ejemplos.

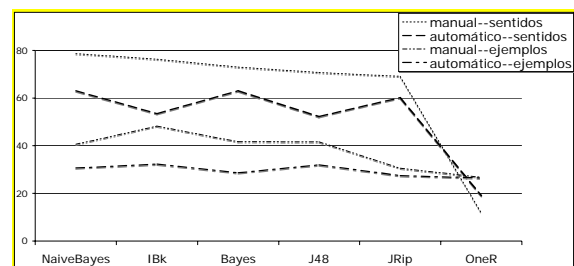


Figura 4. Porcentaje de sentidos bien clasificados en el primer nivel de clustering para la aproximación con esquemas de función y categoría que ocurren más de 10 veces.

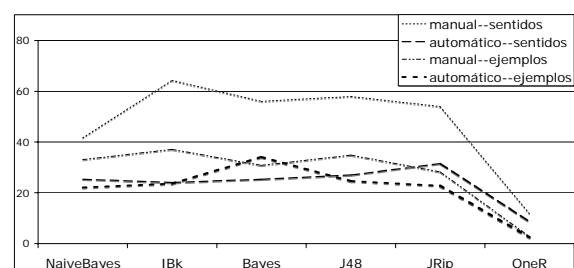


Figura 5. Porcentaje de sentidos bien clasificados en el segundo nivel de clustering para la aproximación con esquemas de función y categoría que ocurren más de 10 veces.

6 Conclusiones y trabajo futuro

Hemos presentado el trabajo realizado en torno a la adquisición de patrones de

subcategorización para sentidos verbales. La metodología empleada es la de clasificar los sentidos por comportamientos verbales extraídos de un corpus anotado manualmente (SENSEM), y a partir de las clases, clasificar sentidos no incluidos en este corpus.

Las clases extraídas presentan diferentes propiedades, nos encontramos con un conjunto importante de sentidos que se incluyen en clases mixtas pero por otro lado, parece que las clases que no son mixtas están claramente delimitadas. Además creemos muy interesante constatar que las clases se caracterizan por comportamientos diatópicos de las piezas verbales, por lo que nos anima a seguir investigando en esta línea.

Por otro lado, los resultados de la compactación y clasificación de los sentidos ya conocidos en clases, a partir del análisis sintáctico automático son muy prometedores, y aportan datos cruciales sobre la importancia de la desambiguación verbal para asignar marco de subcategorización.

En primer lugar, creemos interesante experimentar más con los diferentes métodos de clasificación para poder establecer las mejores clases desde una perspectiva lingüística.

Además, como hemos expuesto, la aplicación del procedimiento en un entorno real, requiere partir de corpus no anotados y no desambiguados semánticamente. Dada la complejidad del proceso hemos dividido la tarea en dos fases, para poder evaluar cada una de las situaciones independientemente. En una primera fase, la que hemos presentado en este artículo, utilizamos el corpus de SENSEM, donde los sentidos verbales están desambiguados, pero sin la anotación manual sintáctico- semántica. Esta experimentación requiere de un análisis morfosintáctico automático y de la aplicación del clasificador.

Una segunda fase consiste en evaluar el clasificador sobre el mismo corpus pero utilizando WSD y análisis automático, para realizar una prueba de adquisición sobre un corpus controlado. Esta fase prevé la aplicación del clasificador sobre corpus de verbos no conocidos.

Referencias

- Alonso, L. e I. Castellón. 2007. Obtaining coarse-grained classes of subcategorization patterns for Spanish. *Submitted*.
- Atserias, J. 2006. Towards Robustness in Natural Language Understanding. Tesis doctoral.

- Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea, Donosti.
- Brent, M. R. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. En *Computational Linguistics*, 19, p. 243-262.
- Briscoe, T. y J. Carroll. 1997. Automatic extraction of subcategorization from corpora. En *Proceedings of the 5th conference on Applied Natural Language Processing*, p. 356-363.
- Carreras, X., I. Chao, L. Padró y M. Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. En *LREC'04*, Portugal.
- Castellón, I., A. Fernández, G. Vázquez, L. Alonso y J. A. Capilla. 2006. The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level. En *LREC'06*, p. 355-359.
- Davies, M. 2005. *A Frequency Dictionary of Spanish*. New York and London: Routledge.
- Dempster, A., N. Laird y D. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. En *Journal of the Royal Statistical Society*, 39.
- Fernández, A., G. Vázquez e I. Castellón. 2004. Sensem: base de datos verbal del español. G. de Ita, O. Fuentes, M. Osorio (ed.), *IX Ibero-American Workshop on Artificial Intelligence, IBERAMIA*. Puebla de los Ángeles, Mexico, p. 155-163.
- Hartigan, J. A. y M. A. Wong. 1979. Algorithm as136: a k-means clustering algorithm. En *Applied Statistics*, 28, p.100-108.
- Korhonen, A. 2002. Subcategorization Acquisition. PhD thesis, *Computer Laboratory*, University of Cambridge.
- Korhonen, A. y J. Preiss. 2003. Improving subcategorization acquisition using word sense disambiguation. En *Proceedings of ACL*.
- Manning, Ch. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. En *31st Annual Meeting of the Association for Computational Linguistics*, p. 235-242.
- Sarkar, A. y D. Zeman. 2000. Automatic extraction of subcategorization frames for Czech. En *COLING'2000*.
- Witten, I. H. y E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Agradecimientos

Esta investigación ha sido posible gracias al proyecto KNOW (TIN2006-1549-C03-02) del Ministerio de Educación y Ciencia, a una beca Postdoctoral Beatriu de Pinós de la Generalitat de Catalunya otorgada a Laura Alonso y a la beca Predoctoral FI-IQUC también de la Generalitat de Catalunya, otorgada a Nevena Tinkova, con número de expediente 2004FI-IQUC1/00084.