Procesamiento del Lenguaje Natural, nº39 (2007), pp. 123-130 recibido 18-05-2007; aceptado 22-06-2007

Inducción de clases de comportamiento verbal a partir del corpus SENSEM

Laura Alonso Alemany
Universidad de la República, Uruguay
Universidad Nacional de Córdoba, Argentina
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 la construcción de un clasificador con el objetivo final de 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 verbales. Estos esquemas se han agrupado en clases de equivalencia 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 lexicón verbal con esquemas 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 the construction of a classifier with the final objective of 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 verbal senses have been acquired. These frames have been grouped in equivalence 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 verbal subcategorizations, parsing, syntactic classes, verb senses.

1 *Introducción*

En este artículo presentamos la construcción de un clasificador de sentidos verbales con el último fin de establecer un método para enriquecer un léxico verbal con información de subcategorización de forma semiautomática, extrapolando la información de un corpus anotado manualmente a ejemplos sin anotación.

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 grupos de verbos con comportamientos sintácticos similares, ya que en un mismo cluster se agrupan verbos que ocurren con esquemas sintácticos parecidos.

Analizamos diferentes opciones para obtener estas clases de verbos similares: diferentes subconjuntos de propiedades para describir a los verbos y diferentes técnicas de clustering. Aplicamos métricas cuantitativas y cualitativas para analizar las diferentes soluciones obtenidas, y finalmente optamos por estudiar con más detalle una solución en dos niveles que consta de 5 clases iniciales y 11 clases en un segundo nivel. Se ha evaluado la utilidad de esta solución para asignar una clase de comportamiento sintáctico a piezas verbales desconocidas con diferentes clasificadores aprendidos automáticamente.

ISSN: 1135-5948 © 2007 Sociedad Española para el Procesamiento del Lenguaje Natural

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 la forma como preparamos los datos del corpus SENSEM, los parámetros de los experimentos de clustering y las métricas para evaluarlas. En la sección 4 mostramos cómo analizamos los resultados de los experimentos, con una breve descripción de las soluciones obtenidas y una descripción más extensa de una de las soluciones. 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, contribuye a la concreción del sentido de los elementos nominales y a la determinación del sentido global de la escena. Por ejemplo, en la frase (1), el verbo *entrar* asigna papel semántico de *ruta* a "*la puerta*", por lo que se prima el sentido de "*abertura*" de la palabra *puerta*, mientras que en la frase (2) el verbo *abrir* le asigna el papel de *tema*, lo cual prima el significado de "*armazón*" para *puerta*.

(1) El viento **entró** por *la puerta*.
(2) *La puerta* **se abre** sobre una explanada.

Por otro lado, desde una perspectiva puramente sintáctica, el verbo nos informa sobre el tipo de complementos que precisa para que una frase sea *gramatical* y si este esquema alterna o no con otros complementos, es decir, sobre las diferentes configuraciones sintácticas de los argumentos. En los siguientes ejemplos observamos cómo la misma construcción sintáctica da lugar a una frase agramatical con el verbo *dormir* o *desear*, pero no con *soñar*.

(3) * *Los niños* ***duermen*** *sueños tranquilos.*
Los niños **duermen**.
(4) Los niños **desean** sueños tranquilos.
* *Los niños* ***desean***.
(5) Los niños **sueñan** sueños tranquilos.
Los niños **sueñan**.

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 el buen funcionamiento de los analizadores sintácticos automáticos, 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 se encuentran determinar de qué núcleo léxico depende un sintagma preposicional (6), la resolución de la coordinación (7) o la determinación de la función de determinados sintagmas nominales (8). A estos problemas se añaden para el español el grado de libertad en el orden de ocurrencia de los constituyentes (9), haciendo que los casos anteriores sean más difícil resolución. Así, conocer la subcategorización del verbo permite evitar la mala identificación de categorías.

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

2.1 Trabajo Relacionado

Los trabajos realizados en el área de la adquisición de subcategorización tienen como objetivo final establecer los patrones de realización para cada unidad verbal. Para ello se trabaja con grandes corpus a partir de los cuales se extrae la información relativa a las realizaciones oracionales.

La adquisición automática de dicha información ha sido tratada por diferentes autores 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, uno de los principales

problemas en esta tarea. Estos trabajos han tenido un acierto de diferente grado en diferentes lenguas. Para el español encontramos trabajos basados en las diátesis o clases verbales que aplican técnicas simlares a los anteriores (Esteve 2004, Chrupala 2004), con resultados bastante positivos

Una de las ambigüedades más difíciles de tratar es la de la adjunción de los sintagmas preposicionales. Algunos autores (Atserias 2006) proponen disponer de dos modelos, uno nominal y otro verbal para que en base a determinadas condiciones disputen por determinados argumentos en una situación ambigua.

2.2 Nuestra Aproximación

A diferencia de estos trabajos, nuestro método parte de una serie de patrones ya adquiridos y evaluados para los sentidos verbales descritos dentro del proyecto SENSEM (ver Figura 1).

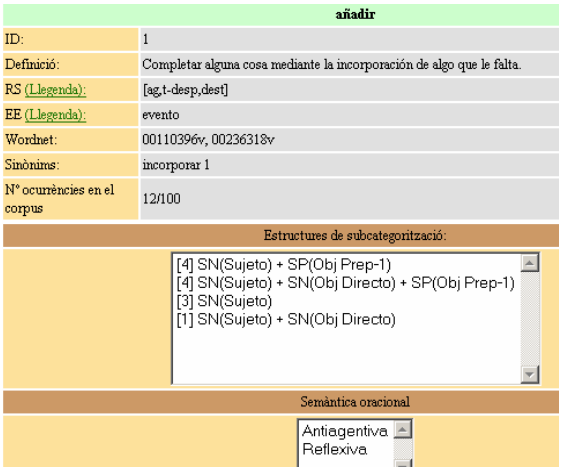

Figura 1. Esquemas de subcategorización adquiridos para el sentido *añadir_1* a partir de la base de datos verbal SENSEM.

Nuestro objetivo final consiste en asociar esquemas de subcategorización a sentidos verbales no descritos en SENSEM. Para ello procedemos en dos pasos:

1) **descubrimos** grandes clases de comportamiento sintáctico distinguible dentro de los verbos de SENSEM, y
2) **clasificamos** nuevos predicados verbales en una de esas clases.

Para llegar a este objetivo final partimos de una serie de hipótesis que creemos necesario exponer. En primer lugar, asumimos que la subcategorización es una información asociada a los sentidos verbales, no a los lemas. En algunos trabajos sobre adquisición de subcategorizaciones se ha trabajado con el lema como unidad de subcategorización (Manning 1993, Briscoe et al 1997). Así, para aplicar el clasificador sobre corpus será necesario disponer de alguna aplicación de algún tipo de desambiguación de sentidos.

Otra 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 resulta muy probable que se pueda caracterizar el comportamiento de un sentido verbal nuevo a partir de extrapolar de alguno de los verbos ya conocidos.

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 mediante clasificadores automáticos. A continuación describimos las fases del experimento: caracterización de los ejemplos, inducción de clases mediante clustering y clasificación de ejemplos no vistos.

3.1 Caracterización de los ejemplos anotados manualmente

El procedimiento que seguimos se basa en los resultados de la anotación de SENSEM. Los ejemplos del banco de datos de SENSEM son frases de corpus periodístico anotadas 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 papel 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 caracterizar el comportamiento sintáctico de los sentidos verbales debemos obtener procedemos en los siguientes pasos:

1) esquema de realización sintáctica de cada ejemplo: para cada ejemplo del corpus, se obtiene su esquema sintáctico

1.1) compactación de 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.

1.2) selección de argumentos, eliminando los constituyentes opcionales (adjuntos).

1.3) eliminación de orden de constituyentes, ordenando los constituyentes en orden alfabético.

2) comportamiento de cada sentido, caracterizado por el número de ejemplos del sentido que ocurren con cada esquema de realización sintáctica posible.

De esta forma obtenemos el equivalente empírico al esquema de subcategorización, a partir de los datos asociados a los sentidos verbales de la base de datos verbal SENSEM (Fernández et al 2004).

Hemos caracterizado los ejemplos (y por lo tanto los esquemas de subcategorización de los sentidos verbales) con diferentes subconjuntos de toda la información disponible:

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

Además, observando los resultados se evidenció que los esquemas de realización sintáctica con pocas ocurrencias en corpus introducían mucho ruido en el espacio de búsqueda, causando agrupaciones extrañas. Así decidimos caracterizar los esquemas de subcategorización utilizando como atributos sólo los esquemas de realización con más de 5 o con más de 10 ocurrencias en el corpus, lo cual redujo sensiblemente el número de atributos, como se ve en la Tabla 1.

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

Tabla 1: Número de esquemas de realización sintáctica distintos encontrados en el corpus al caracterizar los ejemplos con diferentes aproximaciones.

3.2 Inducción de clases de verbos

A partir de los esquemas de subcategorización de los sentidos presentes en el corpus, con los distintos subconjuntos de atributos descritos arriba, tratamos de descubrir clases de sentidos con esquemas semejantes. Para ello caracterizamos a cada sentido como un vector, con los esquemas de realización posibles como dimensiones y el número de ejemplos del sentido que ocurren con cada esquema de realización como valor del sentido para esa dimensión. Esto nos dá una representación de los sentidos en un espacio matemático caracterizado por los esquemas de realización, donde podemos aplicar nociones de *distancia* (o *semejanza*). Sobre este espacio aplicamos métodos de clasificación no supervisada (*clustering*) para encontrar grupos de vectores (sentidos) cercanos en el espacio, es decir, que tienden a ocurrir con los mismos esquemas sintácticos. 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).

Además, en muchas soluciones obtuvimos una clase mayoritaria que contenía verbos con muy distintos comportamientos, 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 de forma preliminar 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. En el futuro aplicaremos un algoritmo de clustering jerárquico.

4 Selección de un conjunto adecuado de clases de equivalencia de sentidos verbales

4.1 Métodos para evaluar soluciones de clustering

La gran cantidad de parámetros descritos en el apartado anterior deja entrever el gran número de experimentos que llevamos a cabo, con soluciones de clustering con diferentes métodos y diferentes subconjuntos de atributos para caracterizar a los sentidos verbales. Por lo tanto se hizo necesario establecer métodos de evaluación sistemáticos, descritos extensamente 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: un bajo índice de solapamiento indica que los sentidos de las distintas clases efectivamente ocurren con distintos esquemas.
- **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 (bonificado) o en los mismos (penalizado). Dado que una de las diferencias entre sentidos verbales puede ser su distinto comportamiento sintáctico, éste es un indicador sólo orientativo.

4.2 Descripción general de las diferentes 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.

En general, el método KMeans, que necesita un parámetro especificando el número de clases que se quieren establecer, proporcionaba peores resultados que EM, sobretodo respecto a la *distribución de la población*. 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 con más detalle las soluciones obtenidas con diferentes tipos de información.

En las soluciones con **categoría, función y papeles semánticos** se distinguen claramente clases con tipos distintos de esquemas de subcategorización, especialmente las soluciones en las que sólo se tienen en cuenta los esquemas de realización que ocurren más de 5 o 10 veces, debido a una notable reducción en la escasez de datos (*data sparseness*) cuando usamos sólo esquemas frecuentes. 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 con papel muy marcado (*origen*, *destino*), con pocas diátesis intransitivas.

En 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 papeles semánticos, 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.

Finalmente, las soluciones donde los sentidos se caracterizan únicamente mediante **categoría** 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. Resulta difícil generalizar el comportamiento de los verbos de estas clases por la gran ambigüedad de los patrones basados únicamente en categorías.

4.3 Solución seleccionada: 5 clases, función + categoría, esquemas que ocurren > 10 veces

A partir de los resultados y comparando las diferentes medidas de evaluación, 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 papeles semánticos. Por este motivo en este primer momento prescindimos de las clases obtenidas con información de papeles semánticos

Tomamos pues como punto de referencia la solución en 5 clases, obtenida con los esquemas caracterizados con función y categoría con más de 10 ocurrencias en corpus. 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. El total de clases es de 5 que se subdivide en un total de 11 clases.

La clase más grande (clase 5, 477 sentidos) está compuesta por sentidos verbales que alternan entre esquemas **transitivos e intransitivos** y en algún caso con preposicionales. 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.

En la segunda clase (clase 2, 163 sentidos) 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 ditranstivos (2.2) y la ausencia en la otra, que se caracteriza por contener esquemas con circunstanciales (2.1).

Las dos siguientes clases (clase 1, 103 sentidos, y clase 3, 68 sentidos) están caracterizadas por alternancias **transtiva – ditransitiva – intransitiva**, con omisiones de ciertos constituyentes. Estas clases no presentan subclases.

La última clase, (clase 4, 63 sentidos) 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, la clase 4.2 es totalmente preposicional y en la clase 4.3 se clasifican sentidos con esquemas transitivos alternantes con preposicionales.

Como vemos, esta solución presenta clases mixtas y algunas que contienen sentidos con comportamiento comparable a los de otras clases. Parece evidente que habrá que profundizar en el método de inducción de clases, pero los resultados hasta el momento son alentadores.

5 Evaluación para aplicación final

Hemos aprendido diversos clasificadores que, dado un sentido caracterizado como vector por sus esquemas de realización, lo asigna a una de las grandes clases de comportamiento verbal inducidas en el paso anterior. Hemos aprendido dos clasificadores bayesianos (clásico y Naive Bayes), dos basados en decisiones (J48, basado en árboles de decisión, y JRip, basado en reglas de decisión), uno basado en los *k* vecinos cercanos (IBk, con *k=1*), y una baseline, equivalente a los resultados obtenidos por casualidad (OneR). Estos clasificadores han sido evaluados mediante *ten-fold cross validation* en el corpus SENSEM.

Recordemos que el objetivo final de la nuestro trabajo es asignar una clase de subcategorización a verbos no descritos previamente, a partir de ejemplos de corpus analizados automáticamente. Para evaluar la utilidad para este objetivo de las clases de equivalencia descritas en el apartado anterior, analizamos el corpus SENSEM automáticamente con Freeling (Carreras et al 2004). La única información que utilizamos del corpus SENSEM es el alcance de los constituyentes dominados por el verbo en cada ejemplo. Hemos comparado el desempeño de los clasificadores en ejemplos caracterizados con análisis

automático y en ejemplos caracterizados con el análisis manual de SENSEM.

También hemos comparado el desempeño de los clasificadores en las grandes clases descritas en el apartado anterior (clases gruesas), y en las clases de granularidad más fina (clases finas). Los resultados pueden verse en la Tabla 2.

	clases gruesas		clases finas	
	manual	auto	manual	auto
Naive Bayes	**78**	**63**	41	25
IBk	**76**	53	**64**	24
Bayes	72	63	56	25
J48	70	52	58	26
JRip	69	60	54	31
OneR	11	19	11	8

Tabla 2. Porcentaje de **sentidos** bien clasificados mediante diferentes clasificadores, con los ejemplos anotados manualmente o automáticamente, con clases finas o gruesas (ver apartado 4.3).

Se puede observar que todos los clasificadores superan significativamente la baseline de OneR. En clases gruesas, los clasificadores simples como Naive Bayes o IBk dan los mejores resultados. Se observa un decremento de unos 10-15 puntos en el desempeño de los clasificadores cuando los ejemplos son caracterizados mediante un análisis automático, lo cual supone una importante desmejora en los resultados, que tendrá que ser mejorada en el futuro.

En clases finas el desempeño de Naive Bayes cae en picado, mientras que el del resto de clasificadores cae unos 10-15 puntos. Probablemente esta desmejora se dá porque los datos disponibles para esas clases, con menos población, son más escasos y los clasificadores no pueden generalizar adecuadamente. En los ejemplos caracterizados automáticamente, la desmejora es muy importante, y, aunque no llega a los niveles del baseline, la significatividad de la clasificación se acerca peligrosamente a los niveles de la casualidad. Habrá que estudiar detenidamente las causas de error para mejorar estos resultados en el futuro.

Por otro lado, hemos realizado otro experimento en el que hemos simulado la ausencia de un algoritmo para desambiguar sentidos. Por ese motivo, la unidad a aprender y clasificar ya no era el sentido verbal, sino que cada uno de los ejemplos era caracterizado como un vector. Estos vectores tienen una caracterización muy pobre, ya que sólo uno de los atributos tiene un valor distinto de cero, justamente, el atributo que se corresponde con el esquema de realización con el que ocurre el ejemplo en concreto. Vemos los resultados en la Tabla 3.

	clases gruesas		clases finas	
	manual	auto	manual	auto
Naive Bayes	40	30	33	22
IBk	48	32	37	23
Bayes	41	28	30	34
J48	41	31	34	24
JRip	30	27	28	22
OneR	26	26	2	2

Tabla 3. Porcentaje de **ejemplos** bien clasificados mediante diferentes clasificadores, con los ejemplos anotados manualmente o automáticamente, con clases finas o gruesas (ver apartado 4.3).

Respecto a la clasificación de ejemplos (vs. sentidos) podemos ver que, aunque los resultados son significativamente mejores que los obtenidos para la baseline en las clases finas, en las clases gruesas los resultados no difieren significativamente, especialmente si los ejemplos son caracterizados con análisis automático. Los métodos simples, especialmente el basado en distancia, IBk, siguen dando los mejores resultados. En clases finas, los resultados son equiparables en análisis manual o automático, pero los porcentajes de ejemplos bien clasificados son demasiado bajos en ambos casos.

6 Conclusiones y trabajo futuro

Hemos presentado una aproximación al enriquecimiento semiautomático de un léxico verbal con esquemas de subcategorización. La aproximación se basa en dos pasos: 1) inducción de grandes clases de comportamiento verbal a partir de ejemplos anotados manualmente, y 2) aprendizaje de clasificadores que etiquetan nuevos ejemplos con esas clases.

Presentamos un método para evaluar sistemáticamente las clases obtenidas con esta aproximación. Mostramos una aplicación preliminar de todo el proceso, con resultados prometedores pero claramente mejorables.

A nivel lingüístico, observamos que las clases de comportamiento verbal inducidas se caracterizan por comportamientos diatéticos 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.

El trabajo futuro que se presenta es mucho e interesante. En primer lugar, creemos importante experimentar más con los diferentes métodos y parámetros de clustering para poder inducir las mejores clases desde una perspectiva lingüística. En especial, nos planteamos el uso de técnicas de clustering jerárquico.

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., I. Castellón y N. Tincheva. 2007. Obtaining coarse-grained classes of subcategorization patterns for Spanish. *RANLP 2007*, Borovets, Bulgaria.

Atserias, J. 2006. Towards Robustness in Natural Language Understanding. Tesis doctoral. Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea, Donosti.

Atserias, J., B. Casas, E. Comelles, M. González, L. Padró y M. Padró (2006). FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *LREC'06*, Génova, Italia.

Brent, M. R. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19, p. 243-262.

Briscoe, T. y J. Carroll. 1997. Automatic extraction of subcategorization from corpora. *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. *LREC'04*, Lisboa, 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. *LREC'06*, Génova, Italia, p. 355-359.

Chrupala, G. (2003) *Acquiring Verb Subcategorization from Spanish Corpora*. Research project presented for the Diploma d'Estudis Avançats. Universitat de Barcelona

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. *Journal of the Royal Statistical Society*, 39.

Esteve, E. (2004) "Towards a semantic classification of Spanish verbs based on subcategorisation information" *Proceedings of the ACL 2004 workshop on Student research*. Barcelona

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, México, p. 155-163.

Hartigan, J. A. y M. A. Wong. 1979. Algorithm as136: a k-means clustering algorithm. *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. *ACL 2003*.

Manning, Ch. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. *ACL'93*, p. 235-242.

Sarkar, A. y D. Zeman. 2000. Automatic extraction of subcategorization frames for Czech. *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.