Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 23-31 recibido 1-05-2009; aceptado 5-06-2009

Hacia un sistema de clasificación automática de sustantivos deverbales

Towards an automatic classification system of deverbal nouns

Aina Peris, Mariona Taulé
CLiC-Universitat de Barcelona
Gran Via de les Corts Catalanes, 585,
08007 Barcelona
{aina.peris,mtaule}@ub.edu

Horacio Rodríguez
TALP-Universitat Politècnica de Catalunya
Jordi Girona Salgado 1-3
08034 Barcelona
horacio@lsi.upc.edu

Resumen: Las nominalizaciones deverbales constituyen una rica fuente de información semántica que puede ser muy útil tener reconocida para diferentes tareas de PLN. En este artículo se presenta el léxico nominal AnCora-Nom, que contiene 817 entradas léxicas de sustantivos deverbales, y una serie de experimentos basados en técnicas de aprendizaje automático que nos han permitido evaluar, positivamente, la consistencia de los datos anotados en AnCora-Nom y detectar los rasgos más relevantes para la distinción denotativa de evento y resultado. Con estos experimentos, además, se sientan las bases para la construcción de un sistema automático de clasificación de los sustantivos deverbales según su denotación.
Palabras clave: nominalización deverbal, aprendizaje automático, clasificador.

Abstract: Deverbal nominalizations constitute a rich source of semantic information. To have it recognized may be very useful for different NLP tasks. In this paper we present the nominal lexicon AnCora-Nom, which consists of 817 lexical entries of deverbal nouns, and a series of experiments based on machine-learning techniques. These experiments allow us to evaluate positively the consistency of annotated data in AnCora-Nom, and to detect the most relevant features for the denotative distinction between event and result nominalizations. Furthermore, with these experiments the foundations of an automatic classification system of the deverbal nominalizations according to their denotation are laid.
Keywords: deverbal nominalization, machine-learning techniques, classifier.

1 Introducción

El creciente interés en el análisis y tratamiento de las nominalizaciones está motivado por la rica información semántica que los SNs que las contienen expresan. De hecho, estos SNs pueden entenderse como paráfrasis de las estructuras verbales correspondientes y, al igual que éstas, expresan relaciones semánticas de tipo argumental y temático (agente, paciente, causa, etc.). Tener identificadas dichas relaciones puede ser muy útil para cualquier tarea o aplicación de PLN (extracción y recuperación de información, sistemas de búsqueda de respuestas, etc.).

De los diferentes tipos de nominalizaciones que se dan en la lengua -deverbales, deadjetivales y denominales-, este trabajo se centra en las deverbales, aquellas derivadas de verbos, ya que asumimos la hipótesis de que los nombres deverbales heredan la estructura argumental del verbo correspondiente. En concreto, nos hemos centrado en la diferencia denotativa que la mayoría de autores (Grimshaw, 1990; Pustejovsky, 1995; Picallo, 1999; Badia, 2002) les atribuye, es decir si el nombre deverbal se interpreta como evento (1a)[1] o resultado (1b).

(1a) **La *aprobación* en febrero de este año de la supresión** de 20 asientos de representación proporcional.

(1b) Contar con **la *aprobación* del Consejo_de_Ministros**.

[1] Todos los ejemplos del artículo se han extraído del corpus AnCora-Es (http://clic.ub.edu/ancora/).

ISSN: 1135-5948

Con el objetivo de determinar los rasgos lingüísticos para establecer la distinción entre ambas denotaciones, en (Peris y Taulé, 2009) se analizaron los diferentes criterios lingüísticos de la bibliografía y se establecieron cuáles eran los más relevantes. A partir de este estudio se ha creado el léxico nominal AnCora-Nom, que será la base para la anotación semiautomática de los SNs en el corpus AnCora-Es (Taulé, Martí y Recasens, 2008).

En este artículo se presentan una serie de experimentos basados en técnicas de aprendizaje automático cuyo objetivo es evaluar la consistencia de los datos anotados en AnCora-Nom, analizar la relevancia de los atributos utilizados en la representación léxico-semántica y poder inferir nuevos rasgos que puedan ser útiles para distinguir entre la denotación eventiva y resultativa en los nombres deverbales. Estos experimentos se llevan a cabo además con la finalidad de sentar las bases para la construcción de un sistema de clasificación automática de los sustantivos deverbales según su denotación.

Un sistema de dichas características puede ser útil en un sistema automático de resolución de la correferencia ya que sustantivos eventivos y resultativos se distinguen por ser antecedentes de distintos tipos de pronombres anafóricos: por ejemplo, el pronombre 'ho' del catalán se usa para correferir sustantivos de evento (2a), mientras que 'el/la' serían los pronombres anafóricos de sustantivos de resultado (2b).

(2a) **La matriculació dels 200 alumnes** ha estat existosa. **Ho** ha permès el nou entorn informàtic.[2]

(2b) **La traducció del llibre** no **l'**ha realitzada un expert.[3]

El artículo se organiza de la siguiente manera: en primer lugar (sección 2) se revisa brevemente el estudio preliminar que nos ha permitido establecer los criterios lingüísticos más relevantes para la distinción de evento y resultado. A continuación (sección 3) se describe el léxico nominal AnCora-Nom, cómo se ha obtenido y el contenido de las entradas. En el siguiente apartado (sección 4) se detallan los experimentos que se han seguido en el proceso de aprendizaje automático, para posteriormente (sección 5) evaluar los resultados obtenidos. Finalmente, se presentan las conclusiones del estudio (sección 6).

2 Estudio Preliminar

Una de las maneras más comunes de interpretar y representar los sustantivos deverbales es en función de su denotación, es decir, si el nombre se refiere a un evento o a un resultado (Grimshaw, 1990; Pustejovsky, 1995; Picallo, 1999; Badia, 2002). Con el objetivo de analizar dicha distinción en Peris y Taulé (2009) se llevó a cabo un análisis de los criterios lingüísticos propuestos en la bibliografía que están en la base de esta distinción. En concreto, se evaluaron un total de siete criterios, mediante la contrastación de los mismos en el comportamiento de 817 nombres deverbales del español (en 3077 ocurrencias) extraídos del corpus AnCora-Es (Taulé, Martí y Recasens, 2008).

La evaluación de estos criterios, de naturaleza morfosintáctica y sintáctico-semántica, ha sido posible gracias a que el corpus con el que trabajamos, AnCora-Es, está anotado a diferentes niveles lingüísticos (morfológico, sintáctico y semántico).

Los resultados obtenidos apuntan que no todos los criterios de la bibliografía parecen confirmarse totalmente y que no siempre es posible distinguir entre ambas denotaciones ya que la información sintáctico-semántica necesaria no siempre está disponible en el contexto, de ahí que en nuestra propuesta se incluya un tercer tipo denotativo, que hemos denominado subespecificado (Véase sección 3).

De los criterios evaluados, se confirman como más relevantes para distinguir entre nombres eventivos y resultativos los siguientes: la clase verbal de la que deriva el sustantivo, la capacidad de pluralización, el tipo de determinante, la preposición que introduce el complemento agentivo y la obligatoriedad del argumento interno, siendo éstos los rasgos que se representan como atributos en las entradas léxicas nominales.

3 AnCora-Nom

El léxico AnCora-Nom consta actualmente de 817 entradas de sustantivos deverbales que se corresponden con 1.121 sentidos. Estos sustantivos se obtuvieron semiautomáticamente

[2] 'La matriculación de los 200 alumnos ha sido exitosa. Lo ha permitido el nuevo entorno informático.'

[3] 'La traducción del libro no la ha realizado un experto.'

de un subconjunto de 100.000 palabras del corpus AnCora-Es a partir de una lista predefinida de sufijos (Santiago Lacuesta y Bustos Gisbert, 1999) que aportan un significado de evento y resultado y que toman verbos como base del proceso de derivación[4].

Cada entrada léxica se organiza en sentidos diferentes que se establecen en función del *synset* de WordNet[5] y de su denotación ("evento", "resultado" y "subespecificado"). En la figura 1, se observa que el lema "reclutamiento" tiene dos sentidos nominales con el mismo *synset* de WordNet pero diferente tipo denotativo.

Cada uno de los sentidos nominales contiene una serie de atributos que se obtienen a partir de la información morfosintáctica y semántica especificada en el corpus AnCora-Es y que presentamos en distintos grupos por razones de claridad expositiva.

3.1 Tipo de nombre

Los atributos que aportan información relacionada con el nombre son: el **lema** (*lemma*), obtenido automáticamente de AnCora-Es ("reclutamiento" en la figura 1); el **sentido de WordNet** (*sense*), con su correspondiente *synset* ("16.00820277n")[6]; el **tipo de nombre** (*noun-type*), cuyo valor puede ser "deverbal", "deadjetival", en función de la palabra de la que deriva (verbo o adjetivo, respectivamente) o "relacional"[7]; y el **subtipo de nombre** (*noun-subtype*) que indica el tipo de denotación, es decir, si se trata de un evento ("event"), un resultado ("result"), un tipo subespecificado ("underspecified"), cuando no es posible discriminar entre ambas lecturas, o negativo ("-") en el caso de las lexías no-nominales. En la figura 1, "reclutamiento" es un nombre de tipo deverbal porque deriva del verbo "reclutar" y puede denotar tanto un evento como un resultado por lo que los valores en los distintos sentidos son "event" y "result"[8].

lemma = reclutamiento
sense = 16.00820277n
noun-type = deverbal
noun-subtype = event
source-pos = v
source = reclutar
source-els = A2
sp = CN-Arg1-de-PAT
espec = def
plural = -

ex = La misión también habló con la secretaria de la Liga_Nacional_para_la_Democracia , Aung_San_Suu_Kyi , quien insistió en " la gravedad persistente del trabajo forzoso , en_particular debido_a su uso por militares y sobre las formas extremas que podía adoptar con **el reclutamiento de niños** " .

lemma = reclutamiento
sense = 16.00820277n
noun-type = deverbal
noun-subtype = result
source-pos = v
source = reclutar
source-els = A2
espec = pos
plural = -

ex = (Fisher ha fracasado lamentablemente en todo **su reclutamiento** desde los - Fab_Five - , y ahora vemos que el solitario ex_compañero de Webber , Juwan_Howard , es un buen jugador ... que no pondría un tapón ni a su hermanita menor).

Figura 1: Entrada léxica de "reclutamiento"[9]

3.2 Palabra Origen

Los siguientes tres atributos hacen referencia a la palabra de la que deriva el nombre, por lo tanto, son específicos de los sustantivos deverbales y deadjetivales.

Palabra origen (*source*): el valor de este atributo es la palabra de la que deriva el sustantivo (en la figura 1, "reclutar").

[4] Los sufijos considerados son: -a, -aje, -ión/-ción/-sión/-ón, -da/-do, -dura/-ura, -e, -ido, -miento/-mento, -ncia/-nza, -o/-eo.

[5] La versión utilizada es el WordNet 1.6 del español: http://garraf.epsevg.upc.es/cgibin/wei4/public/wei.consult.perl

[6] En casos especiales en los que el nombre no está representado en WordNet o forma parte de una multipalabra o de una *named entity*, etc., se les asignan etiquetas especiales, documentadas en http://clic.ub.edu/ancora/lng/en/wordnet_tagset.pdf.

[7] Hasta el momento, este atributo sólo recibe el valor deverbal ya que nos hemos centrado en este tipo de nombres.

[8] En un futuro, queremos incorporar al léxico los sustantivos deverbales agentivos, es decir, aquellos que se refieren al agente de una acción. Ej: "**El *violador* de Mariluz** estaba en libertad".

[9] Se ha destacado en negrita el SN del ejemplo en el que aparece el sustantivo deverbal.

Categoría sintáctica (*source-pos*): este atributo indica la categoría sintáctica de la palabra de la que deriva el sustantivo cuyos valores pueden ser: verbo ("v") para nombres deverbales y adjetivo ("adj") para nombres deadjetivales. En la figura 1, el valor de source-pos es "v" porque la categoría sintáctica de "reclutar" es verbo.

Clase semántica de la palabra origen (*source-els*): si la palabra de la que deriva el nombre es un verbo, con este atributo se especifica la estructura léxico-semántica (els) del verbo en cuestión, es decir, su clase semántica. Los posibles valores de este atributo son todas las clases semánticas (els) especificadas en el léxico verbal AnCora-Verb (Aparicio, Taulé y Martí, 2008)[10]. Este es un rasgo que se ha incorporado en el léxico porque en el estudio preliminar se confirmó como uno de los más concluyentes para establecer la distinción entre evento y resultado.
En la figura 1, el valor de este atributo es "A2", que corresponde a la clase verbal agentiva-transitiva de AnCora-Verb.

3.3 Complementación Nominal

Para cada sentido nominal se incluyen los constituyentes sintácticos que complementan al sustantivo, es decir, sintagmas preposicionales (**sp**), adjetivales (**s.a**), adverbiales (**sadv**), nominales (**sn**) y oraciones (**S**), generados automáticamente a partir de los ejemplos encontrados en AnCora-Es. En el caso de los *sp* y *s.a* se les asocia como valor la función sintáctica de complemento del nombre (CN) y el número de argumento y el papel temático correspondientes[11]. En los *sp* se especifica también la preposición. En cuanto a los *sadv*, los *sn* y las oraciones, sólo se les asigna función sintáctica (CN) porque estos complementos no son argumentos del nombre. Existen casos de determinantes posesivos ("**espec-dp**")[12] y pronombres relativos ("**relatiu**")[13] que pueden interpretarse como argumentos, y a los que se les asigna número de argumento y papel temático.

En la figura 1, el valor del *sp* (en el ejemplo, el reclutamiento *de niños*) es "CN-Arg1-de-PAT", es decir, un complemento del nombre introducido por la preposición "de" interpretado semánticamente como un argumento-1 paciente.

3.4 Rasgos Morfosintácticos

El atributo "**espec**" sirve para indicar el tipo de determinante que precede al nombre y el atributo "**plural**" señala si el nombre aparece o no en plural en el corpus. Se trata de dos rasgos morfosintácticos que se han mostrado muy relevantes para la diferencia denotativa entre evento y resultado. Los posibles valores de "espec" son: "def" (artículo definido), "indef" (artículo indefinido), "dem" (determinante demostrativo), "quant" (determinante numeral o cuantitativo), "pos" (determinante posesivo) y "-" para aquellos casos en los que el sustantivo aparece sin especificar[14]. En la figura 1, los valores de estos atributos son "def" en el primer sentido y "pos" en el segundo.

El valor de "plural" es *booleano*, en función de si el nombre aparece en plural (+) o no (-). En la figura 1, el valor de "plural" de "reclutamiento" es negativo "-" porque en los ejemplos del corpus siempre aparece en singular.

3.5 Lexías y ejemplos

Se ha añadido el atributo **lexía** en aquellos sentidos en los que el sustantivo deverbal forma parte de una unidad léxica compleja (por ejemplo, "clavar_la_*mirada*", "en_*busca*_de, etc.). Los posibles valores son: "verbal",

[10] Las clases verbales de AnCora-Verb se corresponden con los cuatro tipos aspectuales básicos: realizaciones (clase A), logros (clase B), estados (clase C) y actividades (clase D).

[11] Los argumentos que el sustantivo puede seleccionar están numerados incrementalmente –Arg0, Arg1, Arg2, Arg3, Arg4– expresando así su grado de proximidad con el núcleo. Los adjuntos se etiquetan como ArgM. La lista de papeles temáticos agrupa las siguientes etiquetas: AGT (Agente), CAU (Causa), EXP (Experimentador), SCR (Fuente), PAT (Paciente), TEM (Tema), ATR (Atributo), BEN (Beneficiario), EXT (Extensión), INS (Instrumento), LOC (Locativo), TMP (Temporal), MNR (Manera), ORI (Origen), DES (Destino), FIN (Finalidad), EIN (Estado Inicial), EFI (Estado Final) y ADV (Adverbial).

[12] En '[...] conducirá a polímeros más estables y, por_consiguiente, **a su posible *comercialización*** [...]', el determinante posesivo "su" recibe la interpretación de Arg1-PAT.

[13] En '[...] hemos de reconocer un don o talento natural **cuya *carencia*** ninguna educación puede suplir [...]', el relativo "cuya" se interpreta como Arg1-PAT.

[14] Estos valores pueden aparecer también combinados, es decir, "def/indef", "def/pos", etc.

"nominal", "adjetival", "preposicional", "adverbial" y "conjuntiva".

Por último, las entradas léxicas incluyen también los ejemplos del corpus en los que se observa el comportamiento morfosintáctico y semántico codificado.

4 Análisis empírico de los factores que inciden en la clasificación

El objetivo de nuestros experimentos es doble: por una parte, se pretende disponer de un marco que nos permita refrendar empíricamente nuestras hipótesis e intuiciones y evaluar cuantitativamente la importancia de los diferentes factores que consideramos pertinentes (tanto individualmente como combinados); por otra parte, se quieren sentar las bases para la construcción de un clasificador automático que nos permita clasificar un candidato -un nombre susceptible de constituir una nominalización deverbal- como evento o resultado en su contexto de aparición.

4.1 Marco de evaluación

Se han utilizado técnicas de aprendizaje automático para llevar a cabo tanto el análisis de los rasgos (*features*) como la construcción del clasificador. Como herramienta de aprendizaje se utiliza el conocido paquete Weka[15], (Witten y Frank, 2005). El tipo de aprendizaje será supervisado ya que disponemos del corpus de entrenamiento etiquetado manualmente. La evaluación se ha llevado a cabo utilizando *Cross Validation* (con 10 *folds*). De entre los clasificadores que Weka ofrece, inicialmente se ha seleccionado J48.Part, la versión en reglas del clasificador de árboles de decisión C4.5 (Quinlan, 1993). Dicha elección está fundamentada por dos motivos: i) un análisis inicial con otros clasificadores más potentes (o al menos más robustos) como los SVM o el Adaboost no parece dar resultados significativamente mejores; y ii) el clasificador aprendido consiste en una secuencia de reglas simbólicas cuya interpretación por el lingüista es más sencilla.

4.2 Selección de los rasgos

Se han considerado tres niveles de información en el material de que disponemos para llevar a cabo el aprendizaje: lema, sentido y ejemplo (oración). La información disponible es de 817 lemas (sustantivos deverbales) que corresponden a 1.121 sentidos determinados manualmente. En la parte considerada del corpus AnCora-Es existen 3.077 oraciones que contienen estos lemas. Manualmente se han asignado los ejemplos a los sentidos correctos.

Se planificó llevar a cabo dos series de experimentos, la primera a nivel de sentido y la segunda a nivel de oración. La primera serie ha sido realizada completamente y, básicamente, es la que se describe en este artículo. En esta primera serie se han utilizado como rasgos las propiedades contenidas en las entradas léxicas de AnCora-Nom, descritas en la sección 3. Disponemos en este caso de 1.121 ejemplos[16] etiquetados para llevar a cabo el aprendizaje. En la segunda serie de experimentos se añaden rasgos morfosintácticos procedentes del contexto oracional en el que aparece el nombre deverbal (estas oraciones se corresponden con los ejemplos de las entradas léxicas). En este caso, el número de ejemplos disponibles es de 3.077. Se han realizado algunos experimentos iniciales en esta línea, descritos en la sección 4.4 pero son necesarios experimentos adicionales.

4.3 Experimentos a nivel de sentido

En la tabla 1 se recogen los rasgos utilizados en el aprendizaje. En la columna 1 se indica el rasgo tal como aparece definido en la sección 3. La columna 2 indica el número de valores del rango (conjunto de valores posibles) de cada uno de los rasgos. En algunos casos el valor de un rasgo está indefinido, por ello se ha añadido el valor *nil* a cada uno de los rangos. En algunos casos, debido a la excesiva dispersión de los valores posibles, se ha añadido la posibilidad de agrupar algunos de estos valores para facilitar el aprendizaje. La columna 3 presenta el tamaño del rango para los valores agrupados. El caso más interesante de esta agrupación es el del rasgo *sp* en el que, sin agrupar, el número de valores posibles es de 101, demasiado elevado para los 1.121 ejemplos de aprendizaje disponibles. En este caso, se han considerado dos agrupaciones: una a nivel de número de argumento (Arg0, Arg1, Arg2, Arg3, Arg4, ArgM, además del valor no argumental, CN, que proporciona, pues, 7

[15] http://www.cs.waikato.ac.nz/ml/ weka/

[16] Estos ejemplos se corresponden con los 1.121 sentidos nominales representados en AnCora-Nom.

valores posibles) y otra más fina que incorpora al código la preposición involucrada (Arg0-con, Arg0-de, etc. dando lugar a 60 valores posibles). Para cada uno de los rasgos se ha realizado también una descomposición binarizada, es decir, se ha añadido para cada valor posible del rango un rasgo binario que indicara cuando el valor correspondía a dicho rasgo. En general la inclusión de rasgos binarizados ha resultado beneficiosa tal como indica la tabla 2.

Rasgo	***Rango***	***Rango agrupado***
noun_type	4	-
source_els	14	12
sp	101	60-7*
espec	74	15
plural	2	-
lexia	6	-
s.a	9	5
sn	2	-
espec_dp	5	2
S	1	-
relatiu	4	3
sadv	2	-

Tabla 1: Rasgos utilizados en los experimentos a nivel sentido (* en el caso del rasgo *sp* se han utilizado dos agrupaciones diferentes)[17].

4.4 Experimentos a nivel oración

La extracción y codificación de los rasgos a nivel de sentido es trivial a partir de la información contenida en las entradas léxicas. La extracción de los rasgos a nivel oracional, en cambio, presenta más problemas ya que debe llevarse a cabo a partir de los árboles del corpus AnCora-Es. Para esta tarea se utiliza la herramienta Tgrep2[18], que permite la manipulación e inspección de árboles de análisis en formato *treebank* de forma simple y eficiente. Se han llevado a cabo experimentos iniciales utilizando rasgos contextuales obtenidos a partir del árbol de análisis. A continuación, se incluye un ejemplo para ilustrar el proceso.

Un rasgo que hemos considerado interesante incluir es la aparición del nombre deverbal en posición de sujeto o complemento directo. El siguiente patrón de Tgrep2 ejemplifica la extracción del nombre "construcción" en posición de sujeto:

'sn < ("grup.nom" < (n < /construcción/)) < /func\:suj/'

Se puede parafrasear este patrón como "búsqueda de un SN que i) domine inmediatamente un grupo nominal que a su vez domine el nombre que contiene el sustantivo deverbal *construcción* y ii) que, además, tenga la función de sujeto*"*.

De momento, sólo se han incorporado una veintena de rasgos al proceso de aprendizaje. Aparte de los rasgos asociados a las funciones sintácticas de sujeto o complemento directo, se han incluido también nuevos rasgos asociados a otras funciones sintácticas (CC, CPRED, CI, etc.). La mayoría de ellos han resultado irrelevantes. Otros rasgos que se podrían extraer e incorporar al clasificador son el tiempo verbal del verbo con el que se combina la nominalización, la inclusión del nombre en una entidad nombrada, etc.

5 Evaluación de resultados

5.1 Experimentos a nivel sentido

La tabla 2 recoge los resultados obtenidos. El *baseline* se limita a devolver la clase más frecuente. El caso *simples* utiliza los rasgos de la Tabla 1 en su versión escalar. El caso *binarized* usa los mismos rasgos añadiendo ahora los correspondientes binarizados (en general se ha adoptado el criterio de no eliminar los anteriores al refinar los rasgos). Los siguientes casos van incorporando los rasgos agrupados de forma incremental.

En general se puede observar que la utilización de los rasgos aunque sea a nivel simple produce un incremento notable de la precisión del clasificador (del 72% al 82%). También la binarización de los rasgos supone una mejora significativa (hasta el 83,2%). La inclusión de rasgos agrupados es siempre positiva aunque no todas las agrupaciones contribuyen igualmente y no siempre su combinación supone una mejora. Además, las diferencias entre ellas no son estadísticamente significativas en todos los casos. De todas formas, sería deseable un proceso más elaborado de selección de rasgos que incluimos como trabajo futuro.

[17] Se han mantenido las etiquetas sintácticas del corpus AnCora-Es.

[18] http://tedlab.mit.edu/~dr/TGrep2/

Rasgos	*Núm. rasgos*	*Núm. reglas*	*% corrección*
baseline		1	71,9893
simples	12	24	82,0696
binarized	12	32	83,2293
+noun_type	19	27	83,4077
+source_els	34	40	84,0321
+sp (1)	134	30	84,0321
+sp (2)	211	33	83,7645
+espec	214	40	84,5674
+s.a	221	40	84,4781
+espec_dp	231	38	84,4781
+relatiu	247	30	84,5674

Tabla 2: Resultados de los experimentos a nivel de sentido

5.2 Experimentos a nivel oración

Los resultados obtenidos a nivel oracional se presentan en la tabla 3.

Rasgos	*Núm. rasgos*	*Núm. reglas*	*% corrección*
baseline		1	82,1254
full_lemmas	251	61	93,5652
+ suj+cd	258	59	93,6302
+other syntactic functions	258	52	93,4027

Tabla 3: Resultados de los experimentos a nivel de oración

Estos son aparentemente muy superiores pero en realidad la mejora es pequeña. El *baseline* en este caso es del 82% frente al 72% anterior. Esto se debe a que cuando se pasa del marco lema al marco oración, el número de ejemplos para el aprendizaje se incrementa de 1.121 a 3.077, y además, la proporción real (la que existe en el *treebank*) de ocurrencias de nombres resultativos también aumenta. Por ello, el 93,56% de corrección atribuido a la clase "full lemmas", correspondiente a la última columna de la tabla 3, debe considerarse en relación al *baseline*. A nivel de lema la mejora sobre el *baseline* era de un 13% y ahora es de un 14%, es decir, la mejora no es muy relevante. En la fila 3, se introducen los rasgos correspondientes a la aparición del nombre en función de sujeto y complemento directo, en este caso la mejora no es estadísticamente significativa. Tampoco lo es la introducción (fila 4) del resto de funciones sintácticas, que además suponen un descenso en el nivel de corrección.

5.3 Análisis de errores

Se ha llevado a cabo un análisis de los errores para los experimentos a nivel de sentido. En la tabla 4 se presentan los resultados obtenidos para cada clase en cuanto a la precisión, la cobertura y el F-Measure (como valor de la F-Measure hemos ponderado igualmente precisión y cobertura). De estos resultados, cabe destacar que el sistema clasifica mucho mejor los sustantivos resultativos (92,7% de F-Measure) que los eventivos (62,7%) y subespecificados (34,5%). Esto se debe a que existen más rasgos que permiten identificar la clase de resultativos (pluralización, tipo de determinante, clase verbal, adjetivos relacionales). En cambio, en el caso de los subespecificados, como no se dispone de ningún rasgo particular que los identifique (de ahí su clasificación como subespecificados), el sistema no consigue una clasificación óptima. Como se observa en la matriz de confusión de la Tabla 5[19], de los 131 sentidos subespecificados sólo 34 se clasifican correctamente, mientras que el resto se reparte entre la clase de resultativos (55) y la de eventivos (42).

Entre los clasificados como resultativos, el 24,3% corresponde a errores de la clasificación manual en el léxico, por lo que podríamos considerar que este porcentaje en realidad está bien clasificado automáticamente. El 40,5% de los casos se explican porque se trata de sentidos que, o bien no tienen complementos asociados en la entrada, o bien estos complementos no son argumentales (por ejemplo, s.a = CN, sp = CN, S= CN), y esta casuística aparece mayoritariamente en sentidos resultativos. En cuanto al 35,2% restante, son casos cuyos atributos no representan mayoritariamente la clase de subespecificados, sino que se trata de rasgos coincidentes con la clase de resultativos. De ahí que se clasifiquen como resultativos cuando son subespecificados. Lo mismo ocurre en los 22 casos de resultativos clasificados incorrectamente como subespecificados.

[19] Entrando por columnas encontramos los valores predichos por el clasificador, y por filas los valores correctos.

Este mismo argumento, que dos clases compartan la misma casuística de atributos, es válido tanto para los 42 casos subespecificados clasificados como eventivos, como para los 10 casos eventivos clasificados como subespecificados[20].

En el caso de los eventivos, el índice de acierto es menor que en el caso de los resultativos porque también es menor el número de rasgos identificativos de esta clase de nombres (por-SP, posesivo argumental). En concreto, los 23 casos erróneamente clasificados como resultativos aparecen con un sp= Arg1 y con complementos no argumentales, característica compartida mayoritariamente por la clase de resultativos, y de ahí su incorrecta clasificación. Finalmente, los 20 casos de resultativos clasificados como eventivos aparecen con un único sp = Arg1, mayoritariamente representativo de la clase de eventivos.

El F-Measure más alto (99,3%) lo presenta la clase "-" que se corresponde a las denominadas lexías no-nominales marcadas explícitamente en AnCora-Nom, de ahí el alto porcentaje de acierto.

Precision	***Recall***	***FMeasure***	***Class***
0.906	0.948	0.927	***result***
0.515	0.26	0.345	***underspecified***
0.563	0.708	0.627	***event***
1	0.986	0.993	***-***
0	0	0	***nil***

Tabla 4: Resultado de la clasificación según la clase resultante

a	***b***	***c***	***d***	***e***	***Class***
765	22	20	0	0	***a=result***
55	34	42	0	0	***b=underspecified***
23	10	80	0	0	***c=event***
1	0	0	69	0	***d= -***
0	0	0	0	0	***e= nil***

Tabla 5: Matriz de confusión

6 Conclusiones

Los resultados obtenidos en estos experimentos son positivos (un acierto del 84,56%, es decir, un 12,58% de mejora respecto al *baseline* en los experimentos a nivel de lemas y una proporción similar a nivel oracional) en tanto que permiten detectar los rasgos más relevantes para la distinción entre la lectura eventiva o resultativa de los nombres deverbales, corroborando la consistencia de los datos anotados en AnCora-Nom. Además, con estos experimentos se pueden observar también aquellos rasgos más conflictivos, y por lo tanto, incidir en ellos.

Este trabajo permite sentar las bases para la creación de un clasificador automático de sustantivos deverbales según su denotación.

Por supuesto la construcción y uso de un clasificador automático tendrá una penalización a nivel de corrección. El valor de corrección de un 93,56% puede ser considerado como una cota superior del valor obtenible por el sistema automático. Algunos de los rasgos utilizados deberían ser no proporcionados sino calculables automáticamente. El cálculo de buena parte de los rasgos es sencillo a partir de la información morfológica, léxica o sintáctica pero para otros rasgos el cálculo no lo es tanto. Por ejemplo, el clasificador actual supone que los sentidos del nombre deverbal han sido desambiguados, un clasificador automático debería partir no del sentido sino del lema (o, más aún, de la forma de la palabra). Por supuesto, el uso de un sistema de desambiguación de sentidos tiene un coste (en términos de tasa de errores) que se debe asumir, y que difícilmente bajará de un 30%. Los atributos de lexía y algunos otros tampoco pueden obtenerse fácilmente. En cualquier caso, la construcción y evaluación del clasificador automático son un punto importante de nuestro trabajo futuro.

Al margen del clasificador, las líneas futuras de trabajo se centran, por un lado, en la compleción de los experimentos a nivel de oración y, por otro, en la incorporación de un sistema de selección de atributos. Otra línea interesante de experimentación sería la de aplicar una arquitectura de clasificación multiclase con decisiones binarias. Es decir, la combinación de decisiones binarias sobre cada clase (resultativa o no, eventiva o no) que pueden resultar en la asignación de los objetos a más de una clase (eventiva, resultativa) para los casos subespecificados (Boleda, 2007). Otra extensión a abordar en un futuro próximo es la aplicación al catalán de la aproximación aquí propuesta, aprovechando la existencia de AnCora-Ca.

[20] El atributo sp = Arg1 puede darse en ambas clases: subespecificados y eventivos.

Agradecimientos

Este trabajo ha sido posible gracias a los proyectos Lang2World (TIN2006-15265-C06) y AnCora-Nom (FFI2008-02691-E) del Ministerio de Ciencia e Innovación y la beca FPU AP2007-01028 del Ministerio de Educación, Política Social y Deporte.

Bibliografía

Aparicio, J., Taulé, M. y Martí M.A. 2008. AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. En *Proceedings of 6th International Conference on Language Resources and Evaluation.* Marrakesh (Morocco).

Badia, T. 2002. Els complements nominals. En Solà, J. (Ed.) *Gramàtica del Català Contemporani.* Barcelona: Empúries.

Boleda, G., Schulte im Walde, S., Badia, T. 2007. Modelling Polysemy in Adjective Classes by Multi-Label Classification. En *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 171-180.

Grimshaw, J. 1990. *Argument Structure*. Cambridge, Massachussets: The MIT Press.

Peris, A. y Taulé, M. 2009. Evaluación de los criterios lingüísticos para la distinción evento y resultado en los sustantivos deverbales. En *Proceedings of the 1st International Conference on Corpus Linguistics* (CILC-09), Murcia, Spain.

Picallo, C. 1999. La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales. En Bosque y Demonte (Eds.) *Gramática descriptiva de la lengua española.* Madrid: Real Academia Española / Espasa Calpe, 363-393.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge. MIT Press.

Quinlan J. R. 1993. *C4.5: Programs for Machine Learning*. San Francisco. Morgan Kaufmann.

Taulé, M, Martí, M.A y Recasens, M. 2008. Ancora: Multilevel Annotated Corpora for Catalan and Spanish. En *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh (Morocco).

Santiago Lacuesta, R. y Bustos Gisbert, E. 1999. La Derivación Nominal. En Bosque y Demonte (Eds.) *Gramática descriptiva de la lengua española*. Madrid: Real Academia Española / Espasa Calpe, 4505-4594.

Witten, I.H. y Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Franscisco, second edition.