Procesamiento del Lenguaje Natural, Revista nº 35, septiembre de 2005 ISSN: 1135-5948

Sociedad Española para el Procesamiento del Lenguaje Natural

Artículos

© 2005 Sociedad Española para el Procesamiento del Lenguaje Natural

© 2005 Sociedad Española para el Procesamiento del Lenguaje Natural

EDITADO POR:

Ramón López-Cózar Delgado (Universidad de Granada)
L. Alfonso Ureña López (Universidad de Jaén)

COMITÉ CIENTÍFICO:

Presidentes:

Ramón López-Cózar Delgado (Universidad de Granada)
L. Alfonso Ureña López (Universidad de Jaén)

Miembros:

Prof. José Gabriel Amores Carredano (Universidad de Sevilla)
Prof. Toni Badia i Cardús (Universitat Pompeu Fabra)
Prof. Manuel de Buenaga Rodríguez (Universidad Europea de Madrid)
Prof.ª Irene Castellón Masalles (Universitat de Barcelona)
Prof.ª Arantza Díaz de Ilarraza (Euskal Herriko Unibertsitatea)
Prof. Antonio Ferrández Rodríguez (Universitat d'Alacant)
Prof. Mikel Forcada Zubizarreta (Universitat d'Alacant)
Prof.ª Ana María García Serrano (Universidad Politécnica de Madrid)
Prof. Koldo Gojenola Galletebeitia (Euskal Herriko Unibertsitatea)
Prof. Xavier Gómez Guinovart (Universidade de Vigo)
Prof. Julio Gonzalo Arroyo (Universidad Nacional de Educación a Distancia)
Prof. José Miguel Goñi Menoyo (Universidad Politécnica de Madrid)
Prof. Joaquim Llisterri (Universitat Autònoma de Barcelona)
Prof. Ramón López-Cózar Delgado (Universidad de Granada)
Prof. Javier Macías Guarasa (Universidad Politécnica de Madrid)
Prof. José B. Mariño Acebal (Universitat Politècnica de Catalunya)
Prof.ª M. Antonia Martí Antonín (Universitat de Barcelona)
Prof. Lluis Padró (Universitat Politècnica de Catalunya)
Prof. Manuel Palomar Sanz (Universitat d'Alacant)
Prof. José Manuel Pardo Muñoz (Universidad Politécnica de Madrid)
Prof. Germán Rigau (Euskal Herriko Unibertsitatea)
Prof. Horacio Rodríguez Hontoria (Universitat Politècnica de Catalunya)
Prof. Kepa Sarasola Gabiola (Euskal Herriko Unibertsitatea)
Prof. L. Alfonso Ureña López (Universidad de Jaén)
Prof.ª Mª Felisa Verdejo Maillo (Universidad Nacional de Educación a Distancia)
Prof. Manuel Vilares Ferro (Universidade de Vigo)

REVISORES EXTERNOS:

Alicia Ageno, Pablo Daniel Agüero, Iñaki Alegria, Francesc Alías, Laura Alonso, Javier Artiles, Xabier Artola, Zoraida Callejas, Rafael C. Carrasco, Montserrat Civit, Adrià de Gispert, Helenca Duxans, Nerea Ezeiza, David Farwell, Izaskun Fernández, Miguel Ángel García, Carmen García, Manuel García, José María Gómez, José Carlos González, Luis Hernández, Mikel Lersundi, Fernando Llopis, Fernando López, Pilar Manchón, Manuel J. Maña, Montserrat Marimon, María Teresa Martín, Patricio Martínez-Barco, José Luis Martínez, David Martínez, Fernando Martínez, Juan Manuel

Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 245-252 recibido 29-04-2005; aceptado 01-06-2005

Identificación de formas lógicas en el caso del español: propuesta de un modelo basado en reglas y aprendizaje automático. *

Fernando Martínez-Santiago
Universidad de Jaén
Campus Las Lagunillas
dofer@ujaen.es

Miguel Ángel García Cumbreras
Universidad de Jaén
Campus Las Lagunillas
magc@ujaen.es

Resumen: En este trabajo se presenta un modelo mixto para la identificación de formas lógicas para el idioma español. Para aquellos constituyentes de la frase que no involucran un verbo, el sistema utiliza reglas derivadas del árbol sintáctico de la frase. Por otra parte, la obtención de la forma lógica del verbo requiere reconocer los argumentos de éste en la frase. Es por ello que se han aplicado técnicas propias de identificación de roles semánticos, basadas en aprendizaje automático. El modelo se ha evaluado básicamente atendiendo a la precisión alcanzada en la identificación de los argumentos del verbo. En esta tarea se ha obtenido una precisión próxima al 85 %. La identificación completa de formas lógicas se ha llevado a cabo sobre un reducido grupo de frases de sintaxis muy sencilla, ya que el analizador sintáctico disponible en el momento de realizar los experimentos sólo permite un análisis superficial de la sintaxis de la frase.
Palabras clave: identificación de formas lógicas, roles semánticos, análisis sintáctico, aprendizaje automático.

Abstract: We present a mixed model to Logic Form Identification (LFI) for Spanish. The proposed system uses machine learning to decide every argument of the verb. This task is accomplished such as semantic role labeling does. A rule-based system is used in order to obtain the LFI of the rest of the phrase. The rules are obtained by exploring the syntactic tree of the phrase and encoding the syntactic production rules. The verb argument labeling task achieves a precision of 85 %. The LFI has been evaluated by using shallow parsing a some straightforward Spanish phrases.
Keywords: logic form identification, semantic roles, syntactic parsing, machine learning.

1. *Introducción*

En los últimos años la tarea conocida como identificación de formas lógicas o IFL ha despertado un creciente interés, debido a que permite expresar textos en lenguaje natural con un grado de formalismo que mantiene un buen equilibrio entre la complejidad del modelo y la expresividad del mismo. Usualmente estos sistemas están basados en reglas derivadas del análisis sintáctico del texto. Por ello, estos sistemas están limitados, por un lado, por la base de conocimiento codificada en tales reglas, y por otro, por situaciones de ambigüedad sintáctica. Además, codificar estas reglas es una tarea manual que debe ser realizada por un experto humano. Todos estos problemas se ven agravados cuando consideramos un idioma con una complejidad sintáctica elevada, como es el caso del español. Por todo esto en este artículo se propone un modelo que permite la inclusión de técnicas basadas en aprendizaje automático, inspirado en los sistemas de detección y clasificación de roles semánticos. Se consigue disminuir así el número de reglas codificadas y la integración de diversas fuentes de información más allá del análisis sintáctico. En concreto, se propone un modelo mixto para el caso español, y se aplica aprendizaje automático para la obtención de la forma lógica de los verbos. El resto del artículo está estructurado como sigue: en la segunda sección se introduce la tarea de identificación de formas lógicas. A continuación se presenta someramente un

* Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología mediante el proyecto TIC2003-07158-C04-04

ISSN: 1135-5948 © 2005 Sociedad Española para el Procesamiento del Lenguaje Natural

Cuadro 1: Algunas transformaciones a formas lógicas

Regla	Transformación	Aplicable en	Resultado
SP→P SN	prep/P nombre/SN → prep(-,x) nombre(x)	de verano	de(_,x_1) verano(x_1)
SN→SN SP	nombre/SN prep/SP(_,x_1) → nombre(x_2) prep($x2$,$x1$)	noche de verano	noche(x_2) de(x_2, x_1) verano(x_1)

modelo basado en reglas, y los problemas que éste presenta al aplicarse sobre un idioma de tanta complejidad gramatical como es el español. En la sección 4 se describe la tarea de identificación de roles semánticos y su aplicación a IFL. En la quinta sección se describe el marco de experimentación así como los resultados obtenidos. Finalmente se presentan las conclusiones obtenidas y el trabajo futuro.

2. *Identificación de Formas Lógicas*

La identificación de formas lógicas (Rus, 2002) es un formalismo basado en lógica de predicados que pretende obtener una representación del lenguaje natural situada entre el nivel sintáctico y el semántico, a partir de un texto expresado en lenguaje natural. La base de tal formalismo es la lógica de predicados de primer orden, de tal manera que a cada palabra presente en el texto se le asigna un predicado. A su vez cada predicado puede tener varios argumentos que representan la relación de ese predicado con otros elementos de la frase. Además, para mantener la sencillez del modelo se realizan ciertas simplificaciones sobre lo que se puede representar. De esta forma declinaciones de género, número, tiempo, determinantes, la negación y verbos auxiliares son ignorados. A modo de ejemplo, la frase "Juan cuenta cuentos a su hijo cada noche de verano" se representaría como "Juan_N(x_1) contar_V(e_1, x_1, x_2, x_3) cuento_N(x_2) a_P($e1$, x_3) su_P(x_3) hijo_N(x_3) noche_N(x_4) de_P(x_4, x_5) verano_N(x_5)".

Una propiedad relevante de este modelo es que los argumentos siempre aparecen en el mismo orden, si están presentes en la frase, facilitando de esta forma la aplicación de algoritmos de inferencia sobre la base de conocimiento obtenida. Así, en el caso del verbo, el primer argumento representa al verbo en sí mismo, el segundo es el sujeto de la acción, luego sigue el objeto directo, objeto indirecto y finalmente otras funciones sintácticas (complementos circunstanciales, predicativos, etc.).

3. *IFL basado en reglas*

Uno de los modelos más exitosos para IFL es el basado en reglas, el cual obtiene la forma lógica a partir del análisis sintáctico del texto, aplicando ciertas transformaciones sobre las reglas sintácticas que se derivan del árbol. Algunas de estas reglas se ilustran en el cuadro 1. La aplicación de un sistema basado en reglas presenta ciertas dificultades adicionales si se intenta aplicar al español, algunas de las cuales son las siguientes:

- Los sistemas basados en reglas requieren que cada posible producción sintáctica debe tener la correspondiente regla escrita por un experto humano. El número necesario de reglas no sólo dependerá de la cobertura que se desea alcanzar, sino también de la complejidad sintáctica del idioma. Por ejemplo, para un idioma con una sintaxis relativamente sencilla como es el inglés bastan diez reglas para cubrir el 90 % de las glosas presentes en WordNet (Moldovan y Rus, 2001).
- Mientras que en inglés los distintos constituyentes de la frase tienen una posición más o menos fija dependiendo del tipo sintáctico, esto no ocurre con el español, donde tan válida es la frase "Ana lee muchos libros" como "muchos libros lee Ana".
- El sujeto elíptico, si el analizador sintáctico no crea la etiqueta correspondiente, requiere ser tratado como una excepción. En el modelo aquí propuesto es detectado mediante algunas heurísticas sencillas, tales como la ausencia de un sintagma nominal a la izquierda del verbo, para el caso de las frases en forma activa.
- El uso de los pronombres presenta una gran complejidad. Distinguir entre un pronombre objetivo directo y un pronombre objetivo indirecto puede resultar confuso para un sistema basado en reglas cuya única fuente de información es

la obtenida por un analizador sintáctico. Tal es el caso del pronombre *te* en las frases "Ya te veo" y "Ana te envía un mail"

- Finalmente, estos sistemas requieren analizadores sintácticos precisos que no siempre están disponibles para otros idiomas distintos al inglés.

4. *Aplicación de técnicas de identificación de roles semánticos a la IFL*

La finalidad última de este trabajo es complementar un sistema de identificación de formas lógicas basado en reglas con un módulo basado en aprendizaje. Por un lado, el sistema basado en reglas se utilizan para derivar la forma lógica de cada constituyente no verbal de la frase. Por otro, el módulo basado en aprendizaje extrae ciertas propiedades sintácticas de los constituyentes de la frase, y predice en base a ello qué rol o argumento ocupa tal constituyente en relación al verbo. Nótese el alto grado de similitud entre el módulo basado en aprendizaje y la tarea de identificación de roles semánticos. Esta tarea consiste en la detección y clasificación de la relación semántica existente entre el verbo y los constituyentes de la oración, estableciéndose relaciones entre ambos tales como *agente, paciente, experimentador, benefactivo,* etc. (Gildea y Jurafsky, 2002) distinguen entre la detección del rol y su clasificación[1]. En la forma más sencilla de esta tarea la detección ya es dada, quedando los roles marcados previamente, preocupándose el sistema tan sólo de decidir a qué tipo pertenece tal rol. (Gildea y Jurafsky, 2002) proponen un modelo basado en aprendizaje en el cual se extraen seis propiedades para cada rol, esperando que en tales propiedades se encuentre la ligazón entre la sintaxis y la semántica de tal rol.

Si bien existe cierta similitud entre la identificación de roles semánticos y los argumentos de la forma lógica del verbo también existen sustanciales diferencias:

- El número de categorías semánticas depende del verbo, y es muy superior al número de argumentos que puede recibir la forma lógica de un verbo. Además, estos argumentos son siempre los mismos: sujeto, objeto directo, objeto indirecto y otras funciones sintácticas. Es por ello necesario agrupar diversos roles con el mismo argumento de la forma lógica del verbo. En cierta forma la detección de estos argumentos es un simplificación de la tarea de clasificación de los roles semánticos.
- La detección de los límites dentro de la frase de cada rol semántico es un problema más duro que la posterior clasificación. Sin embargo, la detección de los argumentos de la forma lógica del verbo viene dada implícitamente, mediante la aplicación de las reglas sintácticas codificadas a priori. Por ello, en este trabajo se equipara la subtarea de clasificación de roles semánticos con la identificación de los argumentos de la forma lógica del verbo, obviándose la detección del rol, que viene ya dado por la aplicación de reglas sintácticas.

El modelo de IFL propuesto aplica las reglas sintácticas para obtener la forma lógica de aquellos constituyentes no verbales de la oración. Llegado este punto, cada uno de esos constituyentes se clasifica como si de clasificar un rol semántico se tratara, sólo que en vez de tratar de averiguar de qué rol se trata, tan sólo se infiere qué argumento es de la forma lógica del verbo. El modo en que se implementa esta tarea sigue en gran medida lo propuesto en (Gildea y Jurafsky, 2002), adaptado al español y a los recursos disponibles. Se extraen siete características del constituyente, que son utilizadas para alimentar un sistema de aprendizaje automático. Estas siete características son las siguientes:

1. **Tipo de sintagma**, que puede ser *S*, *SP*, *SADJ* o *SADV*.
2. **Categoría gobernante**, que puede ser *S*, o *SV*. En nuestros experimentos, todo aquello que no pertenece al sujeto, se marca *SV* como categoría gobernante.
3. **Camino en el árbol sintáctico**, determina el camino que hay que recorrer en el árbol sintáctico para ir del constituyente al verbo.
4. **Posición** del constituyente respecto del

[1]Nótese que un sistema real que pretenda identificar los roles semánticos de un verbo deberá también desambigüar éste, ya que según el sentido del verbo, así se espera que aparezcan unos roles u otros

verbo, según quede a la izquierda o derecha de éste.

5. **Voz del verbo**, que puede ser activa o pasiva.
6. **Palabra principal** del constituyente. Esta palabra es marcada por el *parser* atendiendo a un grupo de reglas heurísticas (Collins, 1999).
7. **La concatenación de la etiqueta de parte de la oración de cada palabra del constituyente.** El sintagma "el año pasado" toma en esta característica el valor $\{DNJ\}$.

Las seis primeras características coinciden con las propuestas en (Gildea y Jurafsky, 2002). La concatenación de la etiqueta de parte de la oración es frecuente encontrarla en diversos sistemas de identificación de roles semánticos (Carreras, Màrquez, y Chrupala, 2004; Punyakanok et al., 2004).

5. *Experimentos y resultados*

5.1. Marco de experimentación: recursos utilizados

Para la elaboración de nuestros experimentos hemos utilizado el paquete de análisis lingüístico FreeLing (Carreras et al., 2004)[2]; el corpus anotado semánticamente 3LB-Cast (Palomar et al., 2004) y el software de aprendizaje automático TimBL (Daelemans et al., 2004) [3]. El conjunto de frases y su correspondiente forma lógica utilizado para la evaluación del sistema completo ha sido traducido a partir de la colección de prueba de la tarea de identificación de formas lógicas llevada a cabo en Senseval-3 (Rus, 2004):

- FreeLing es un paquete de análisis lingüístico disponible para el catalán, castellano e inglés, y que realiza, entre otras cosas un etiquetado de la parte de la oración de cada palabra, un análisis morfosintáctico y un análisis sintáctico superficial. Nótese que tanto las reglas sintácticas como las características extraídas de cada constituyente se derivan del árbol sintáctico generado por *FreeLing*. Al tratarse de un análisis sintáctico superficial, esto ha impuesto ciertas limitaciones en los experimentos en los que se ha utilizado este paquete:
  - El árbol sintáctico siempre tiene una única etiqueta sintáctica "S" que representa toda la frase, aunque tal frase esté constituida por más de una oración. Esto impide analizar correctamente frases compuestas, por lo que el conjunto de sentencias analizado ha sido limitado a frases con un único sintagma verbal.
  - Analizadores sintácticos como (Collins, 1999) marcan la palabra principal o *head word* de cada sintagma. Lamentablemente los analizadores sintácticos disponibles para el español no facilitan este dato. Sin embargo, la *head word* sí está anotada, al menos indirectamente, en el corpus 3LB-Cast utilizado en los experimentos. En este trabajo la *head word* es el núcleo del sintagma, el hijo directo del sintagma, no otro sintagma anidado (que actuaría como complemento)[4].
  - Al tratarse de un analizador sintáctico superficial los árboles sintácticos suelen ser muy planos. En consecuencia el número de reglas sintácticas necesario es más alto de lo que sería con un árbol más profundo.
- 3LB-Cast es un corpus en castellano con anotaciones sintácticas, semánticas y pragmáticas. En este corpus los roles semánticos no están anotados tal como en otros corpus como TreeBank o las glosas que se encuentran en FrameNet. Aún así la información anotada es adecuada para que un sistema aprenda con qué argumento de la forma lógica del verbo se corresponde un constituyente dado. Para ello se ha asociado cada función sintáctica con un argumento de la forma lógica del verbo según se muestra en el cuadro 2.
- Timbl es una herramienta que implementa diversos algoritmos de aprendizaje supervisado basados en memoria. Este paquete presenta algunas cualidades,

[2] disponible en http://garraf.epsevg.upc.es/freeling [28/3/2005]

[3] disponible en http://ilk.uvt.nl/software [28/3/2005]

[4] gracias a Borja Navarro, que me ha sugerido esta aproximación sobre el corpus 3LB-Cast

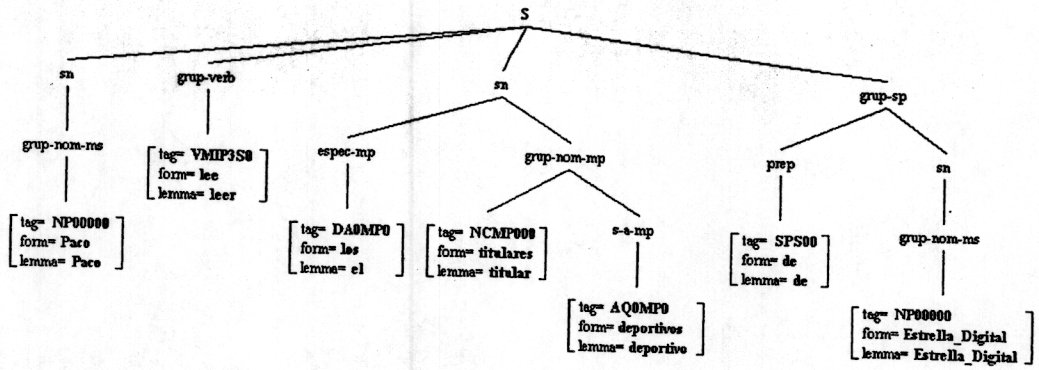

Figura 1: Árbol sintáctico de la frase "Carlos lee los titulares de la Estrella Digital"

Cuadro 2: Correspondencia entre funciones sintácticas anotadas en el corpus 3LB-Cast y los argumentos de la formas lógicas verbales

Función sintáctica	Argumento
SUJ	2
CD	3
CD.NF	3
CD.Q	3
CI	4
CI.DATI	4
CI.NF	4
ATR	5
CPRED	5
CPRED.SUJ	5
CPRED.CD	5
CAG	5
CREG	5

tales como aprender con características alfanuméricas o clasificación multi-clase, lo que hace que sea adecuado para tareas relacionadas con el PLN como la que aquí se plantea.

5.2. Un caso de uso

En este apartado se ilustra con un sencillo ejemplo el funcionamiento global del modelo mixto propuesto, utilizando los recursos descritos en la sección anterior. La frase de partida es:

> Carlos lee los titulares deportivos de la Estrella Digital

Y se debe llegar a su forma lógica:

$$\text{Carlos\_N}(x_1) \quad \text{leer\_V}(e_1,x_1,x_2,\_,x_4)$$
$$\text{titular\_N}(x_2) \quad \text{deportivo\_J}(x_2)$$
$$\text{de\_P}(x_2, x_3) \quad \text{Estrella\_Digital\_N}(x_3)$$

El cuarto argumento de la forma lógica del verbo queda vacío al carecer esta frase de objeto indirecto. Además, "Estrella_Digital" es marcado como un nombre gracias a que el paquete lingüístico FreeLing detecta a esta expresión como una entidad con nombre.

- El primer paso requiere la obtención del árbol sintáctico de la frase (fig. 1).
- Al lema de cada palabra de la frase, junto con la etiqueta de la parte de la oración, se le asigna un predicado.
- Se aplican reglas sintácticas *horizontalmente*. De esta forma se detecta que "deportivo" está adjetivando a "titular" y en consecuencia debe compartir argumento (cuadro 3).
- Las formas lógicas obtenidas se propagan hacia atrás en el árbol sintáctico, intentando aplicarse reglas sintácticas *verticalmente* (cuadro 4).
- Finalmente, es necesario extraer las características de cada constituyente para inferir con qué argumento verbal se corresponde, utilizando aprendizaje automático tal como ha sido descrito en sección 4.

5.3. Experimentos realizados

La evaluación del módulo de aprendizaje en la predicción de los argumentos de la forma lógica del verbo se ha realizado extrayendo 15000 constituyentes, procedentes de 3LB-Cast, para la fase de entrenamiento, y 1000 para la evaluación. Nótese que este corpus incluye un completo análisis sintáctico manual de cada frase. Es por ello que aquí no ha sido necesario utilizar el paquete FreeLing ni

Cuadro 3: Ejemplo de regla horizontal

regla	aplicable a	resultado
SN→NJ	noticia_N(x_2) deportiva_J(_)	noticia_N(x_2) deportiva_J(x_2)
SP→P SN	de_P(_,-) Estrella_Digital_N(x_3)	de_P(_,x_3) Estrella_Digital_N(x_3)

Cuadro 4: Ejemplo de regla vertical

regla	aplicable a	resultado
S→SN V SN SP	Paco_N(x_1) leer_V($e_{1,-,-,-,-}$) noticia_N(x_2) deportiva_J(x_2) de_P(_,x_3) Estrella_Digital_N(x_3)	Paco_N(x_1) leer_V($e_{1,-,-,-,-}$) noticia_N(x_2) deportiva_J(x_2) de_P(x_2,x_3) Estrella_Digital_N(x_3)

limitar el tipo de frases utilizadas en la experimentación.

Por otra parte, y con la finalidad de comprobar la viabilidad del modelo mixto propuesto se han seleccionado 15 frases con su correspondiente forma lógica, procedentes de la tarea de identificación de formas lógicas de SENSEVAL-3. Estas frases han sido traducidas manualmente al español, así como su forma lógica, que han sido adaptadas a este idioma. Debido a las limitaciones del analizador sintáctico utilizado, las 15 frases presentan un único sintagma verbal. Finalmente, para dar cobertura a estas 15 frases se han codificado 37 reglas.

5.4. Resultados obtenidos

En el cuadro 5 se muestra cómo han quedado clasificados los mil constituyentes extraídos del corpus 3LB-Cast. La precisión media alcanzada es del 84 %. Un resultado muy destacado es la precisión exacta alcanzada en la detección del sujeto (categoría 2) acertando 238 de 238 posibles. A pesar de la gran flexibilidad sintáctica del castellano, la detección del sujeto se ve facilitada por el hecho de que el sujeto usualmente aparece a la izquierda del verbo, y con un nombre como *head word*. Estas dos características son tenidas en cuenta durante el entrenamiento.

Cuadro 5: Tabla de confusión sobre los 1000 argumentos evaluados (frases procedentes del corpus 3LB-Cast)

	5 (CP)	3 (CD)	2 (S)	4 (CI)
5 (CP)	439	82	0	1
3 (CD)	20	164	0	13
2 (S)	0	0	238	0
4 (CI)	10	6	0	27
Precisión media: 0,84				

El objeto directo (categoría 3) y los diversos complementos predicativos (categoría 5) obtienen también un buen resultado situado entre el 85-90 %. Sin embargo, la detección del objeto indirecto (categoría 4) es la que peor resultado obtiene, con una precisión próxima al 65 %. La mayoría de los errores asociados al objeto directo e indirecto se deben a constituyentes erróneamente marcados como complementos predicativos. Esto puede deberse a un sobre-entrenamiento de la categoría 5, que engloba aquellos complementos de verbo que no son ni el objeto directo ni el indirecto. Como puede apreciarse en la tabla de confusión, la gran mayoría, casi un el 70 %, de los complementos del verbo resultan pertenecer a esta categoría 5.

Posiblemente, para mejorar estos resultados se requiera alimentar el sistema con información semántica. Una ventaja del uso de algoritmos basados en aprendizaje frente aquellos basados en reglas es la posibilidad de integrar nuevas fuentes de conocimiento, sin las cuales es imposible que el sistema averigüe con qué argumento se corresponde cierto constituyente. Así por ejemplo la frase

Pedro recomienda a el niño que lea.

debería marcar "a Pedro" como objeto indirecto. Sin embargo en la frase

Ana quiere a el niño

aparece el mismo constituyente en una estructura sintáctica similar, y sin embargo actúa como objeto directo. Resolver tales situaciones requiere un conocimiento adicional sobre la naturaleza del verbo, los roles semánticos que puede recibir éste y el tipo de argumento. Por ejemplo, de un recurso como FrameNet (Baker et al., 1998) podría aprenderse que "lo querido"son personas, y de WordNet se desprendería que un niño es una persona.

Con la finalidad de comprobar el modelo

Cuadro 6: Algunas frases y la forma lógica obtenida

Juan vuela desde Tokio hasta Nueva York	Juan_[P](x1) volar_[V](e1 x1) desde_[P](e1 x2) Tokio_[N](x2) hasta_[P](x2 x3) Nueva_York_[N](x3)
John es golpeado por una pelota	John_[N](x1) golpear_[V](e1 x2 x1) por_[P](e1 x2) pelota_[N](x2)
En vez de alubias comeré pizza	En_vez_de_[P](x2 x1) alubias_[N](x1) comer_[V](e1 _ x2) pizza_[N](x2)
El baloncesto y el tenis son grandes deportes	baloncesto_[N](x1) y_[C](x3 x1 x2) tenis_[N](x2) ser_[V](e1 x3 x4) grande_[A] (x4) deporte_[N](x4)
El profesor permitió un periodo de descanso	profesor_[N](x1) permitir_[V](e1 x1 x3) periodo_[N](x3) de_[P](x3 x2) descanso_[N](x2)

Cuadro 7: Tabla de confusión sobre los 43 argumentos evaluados (15 frases traducidas del *SENSEVAL-3 LFI workshop*)

	5 (CP)	3 (CI)	2 (S)	4 (CD)
5 (CP)	10	2	0	0
3 (CI)	6	4	0	0
2 (S)	2	0	8	0
4 (CD)	0	2	0	5
Precisión media: 0,74				

completo, integrando tanto la parte basada en reglas como la basada en aprendizaje, se ha realizado un pequeño experimento en el que se ha obtenido la forma lógica de 15 frases procedentes del workshop *Senseval-3 Logic Forms workshop*. Debido a las restricciones del analizador sintáctico, las frases seleccionadas cuentan todas con un único verbo. Estas frases han sido traducidas manualmente al español. De igual manera la forma lógica que se deriva de ellas ha sido también adaptada del inglés al español. Para obtener la forma lógica de los constituyentes no verbales se han necesitado implementar 37 reglas como las mostradas en los cuadros 3 y 4. Tras la aplicación de estas reglas, han quedado 44 constituyentes repartidos entre 15 frases, que deben ser clasificados según resulten ser un sujeto, objeto directo, indirecto u otros. En los cuadros 7 se muestra la precisión alcanzada por el algoritmo basado en aprendizaje sobre esos argumentos. En el cuadro 6 se muestran algunos ejemplos de frases junto con la forma lógica obtenida. Los resultados siguen en la línea, a la baja, de los alcanzados previamente sobre el corpus 3LB. Un posible motivo es que la evaluación sobre 3LB-Cast está realizada con un corpus etiquetado manualmente, lo cual elimina en gran medida errores debidos a un análisis sintáctico erróneo.

6. *Conclusiones y trabajo futuro*

Se ha presentado un nuevo modelo para la derivación de formas lógicas basado en reglas y en aprendizaje automático. Además, este modelo ha sido aplicado al español, idioma especialmente complicado para esta tarea, dada la flexibilidad de su sintaxis. Justamente por ello, un sistema basado únicamente en reglas no puede rendir tan bien como en otros idiomas como el inglés. Sin embargo, la introducción de aprendizaje automático permite integrar conocimiento que va más allá de la producción sintáctica. En este trabajo, el aprendizaje automático se ha utilizado al modo que (Gildea y Jurafsky, 2002) proponen para la clasificación de roles semánticos. Como trabajo futuro más inmediato, quedan por integrar otras fuentes de conocimiento de carácter marcadamente semántico, como puede ser el uso de un probable futuro 3LB-Cast etiquetado con roles semánticos al modo en que lo está el corpus PropBank. Igualmente, el siguiente paso será la evaluación del modelo propuesto sobre otros idiomas y con un conjunto de frases más amplio.

Bibliografía

Baker, Collin F., Fillmore, Charles J., Lowe, y John B. 1998. The Berkeley FrameNet project. En *Proceedings of the COLING-ACL*, Montreal, Canada.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. Freeling: An open-source suite of language analyzers. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Carreras, Xavier, Lluís Màrquez, y Grzegorz Chrupala. 2004. Hierarchical Recognition

of Propositional Arguments with Perceptrons. En *Proceding of CoNLL 2004 Shared Task: Semantic Role Labeling*, Boston, USA, May.

Collins, Michael John. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. tesis. Supervisor-Mitchell P. Marcus.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, y Antal van den Bosch. 2004. Timbl: Tilburg memory based learner, version 5.1, reference guide. Technical Report Series 04-02, ILK Research Group.

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

Moldovan, Dan I. y Vasile Rus. 2001. Logic Form Transformation of WordNet and its Applicability to Question Answering. En *Proceedings of the ACL 2001 Conference*, Toulouse, France, July.

Palomar, M., M. Civit, A. Díaz, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, M.A. Martí, y B. Navarro. 2004. 3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y español. *XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, páginas 81–88, Julio.

Punyakanok, V., D. Roth, W. Yih, y D. Zimak. 2004. Semantic Role Labeling via Integer Linear Programming Inference. En *Proc. the International Conference on Computational Linguistics (COLING)*.

Rus, Vasile. 2002. *Logic Form For WordNet Glosses and Application to Question Answering*. Ph.D. tesis, Computer Science Department, School of Engineering, Southern Methodist University, Dallas, Texas.

Rus, Vasile. 2004. A first evaluation of logic form identification systems. En Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, páginas 37–40, Barcelona, Spain, July. Association for Computational Linguistics.