DETECCIÓN AUTOMÁTICA DE ERRORES EN EL CORPUS SENSEM

L.Alonso*, I. Castellón†, N. Tincheva†

*alemany@famaf.unc.edu.ar

†{icastellon; nevenatinkova}@ub.edu

GRIAL

†Dpto. Lingüística General

Universidad de Barcelona

*Sección de Ciencias de la Computación

Facultad de Matemáticas, Astronomía y Física

Universidad Nacional de Córdoba (Argentina)

Resumen

El proyecto SenSem tiene como objetivo la anotación sintáctico-semántica a nivel verbal de un amplio corpus del español. Una de las fases del proceso de anotación es la corrección de errores en el corpus. Para ello estamos desarrollando una serie de procedimientos que permiten la detección automática de errores para su posterior corrección. Estos procedimientos se encuentran actualmente en fase de estudio y diseño, fases previas a su implementación computacional. En este artículo presentamos un proyecto, el tipo de anotación que se realiza y la tipología de errores, así como algunos de los procedimientos de detección.

Palabras claves: *procesamiento del lenguaje natural, anotación de corpus, tipología de errores, detección automática de errores.*

Abstract

The goal of the SenSem project is the syntactico-semantic annotation at verbal level of a big corpus of Spanish. One of the phases in this process consists in correcting annotation errors in the corpus. To that aim, we are developing some procedures to detect errors automatically and correct them. These procedures are currently being studied and designed, before they are actually implemented computationally. In this paper we describe the project and the kind of annotation it involves, and we discuss a typology of errors, as well as some detection procedures.

Keywords: *Natural Language Processing, annotated corpora, typology of errors, automatic error detection*

1. Introducción

La anotación de corpus es una línea de trabajo que tiene un gran interés para la lingüística computacional, y que actualmente se está desarrollando para diversas lenguas (Palmer et al 2004, Palomar et al 2004). En este tipo de recursos uno de los problemas principales es la generación de errores por parte de los anotadores (Dickinson 2005, Civit et al 2003). Delimitar este tipo de errores permite conocer la situación del recurso y además posibilita su mejora en anotaciones posteriores, muchas veces mediante un conjunto de criterios elaborados a tal fin. En este artículo nos centramos en la detección y corrección de errores en el corpus anotado del proyecto SenSem.

El proyecto SenSem (Alonso et al 2005) tiene como objetivo construir una base de datos léxica describiendo el comportamiento sintáctico-semántico de los 250 verbos más frecuentes del español actual. La unidad de descripción es el sentido, de modo que un verbo puede tener varias descripciones – una por cada sentido observado. Cada sentido verbal se asocia a un conjunto de ejemplos de corpus real que posteriormente se analizan manualmente. El análisis se realiza a tres niveles diferentes: el verbo como unidad léxica, los constituyentes de la oración y la semántica oracional. En el proceso de anotación se pueden diferenciar los siguientes pasos: la identificación y definición del sentido verbal, el análisis de las estructuras sintácticas, la interpretación de los papeles semánticos y el análisis de la semántica oracional.

Con la inclusión de la semántica oracional pretendemos dotar a la descripción realizada de un nivel de información más que consideramos fundamental en el tratamiento de la interfaz sintáctico-semántica de los verbos. Asimismo, como una etapa más de nuestro proyecto, desarrollaremos un programa de detección automática de errores que nos permitirá corregir la anotación del corpus de oraciones desarrollado en SenSem para mejorar su fiabilidad – un desafío al que se enfrenta actualmente la lingüística computacional.

Para ello, primero se realizó un análisis exhaustivo de 5 de los verbos incluidos en el corpus, con un total de 500 frases, donde se estudiaron a fondo los tipos de errores que se producen y se planificó el tratamiento más adecuado para cada uno. En cada

nivel de anotación se presentan diferentes tipos de errores que se explicarán en el apartado 3. Finalmente presentaremos las conclusiones y las futuras líneas de investigación.

2. El corpus SenSem

El banco de datos de SenSem está compuesto por oraciones. En cada una de estas oraciones se analiza un verbo en forma personal y los constituyentes directamente relacionados con él. No se anotan otros predicados verbales que pudiera haber en la oración y aquellos constituyentes que estén más allá del alcance del verbo analizado. Veamos un ejemplo:

iniciar: ...El presidente, [que] [ayer] **inició** [una visita oficial a la capital francesa], hizo estas declaraciones…

hacer: ...[El presidente, que ayer inició una visita oficial a la capital francesa,] **hizo** [estas declaraciones]…

Si analizamos el verbo "iniciar", dejaremos sin analizar todos los constituyentes que están fuera de la oración de relativo, en cambio, si analizamos "hacer", el sujeto se tratará como un constituyente atómico, sin entrar a analizar su estructura interna. Los predicados se anotan a tres niveles:

- Semántica Oracional

Distinguimos tres tipos de semántica eventual: eventos, procesos y estados. Además, también anotamos los siguientes significados oracionales: anticausativa, antiagentiva, impersonal, reflexiva, recíproca o habitual. Este tipo de información es útil para especificar las estructuras argumentales de cada unidad verbal.

- Semántica Léxica Verbal

Cada predicado se asocia a uno de los sentidos del verbo al que pertenece. Para ello, se ha desarrollado un léxico verbal en el que se listan los sentidos posibles para cada verbo, su clase eventiva, su estructura de participantes, una lista de unidades léxicas sinónimas y antónimas y los synsets relacionados en EuroWordNet (Vossen et al 1998).

- Semántica de Constituyentes

Cada constituyente directamente bajo el alcance del predicado que se anota se asocia a:

- categoría morfosintáctica (p.ej.: *sintagma nominal, oración adverbial*),
- función sintáctica (p.ej.: *sujeto, objeto preposicional*),
- su relación con el verbo (p.ej.: *argumento* o *adjunto*),
- cada argumento es asociado a un rol semántico (p.ej: *iniciador, tema afectado, origen, tiempo*),
- se marcan los núcleos de los argumentos y su posible uso metafórico, ya que esta información resulta útil para adquirir las preferencias selectivas de cada sentido verbal.

Como los argumentos se definen como participantes de la escena representada por el predicado, su rol semántico forma parte de la semántica léxica del verbo. Por esta razón cada sentido verbal se asocia a una estructura temática prototípica que incluye los posibles argumentos verbales. Como en el caso de los sentidos, esta estructura temática es preliminar y es necesario modificarla si los ejemplos del corpus proporcionan una evidencia no contemplada inicialmente.

Además, en algunos casos, se ha incluido información relevante sobre unidades que pueden alterar alguna interpretación o bien que creemos interesantes para trabajos futuros, como los ítems de polaridad negativa.

3. Tratamiento de los errores

El objetivo de esta tarea es, en última instancia, corregir los posibles errores en la anotación del corpus SenSem. En primer lugar, consideramos que un porcentaje alto de estos errores pueden detectarse mediante procesos automáticos, y, de éstos, otro porcentaje nada desdeñable puede corregirse mediante procesos automáticos. En este momento, la investigación está en la etapa de detección automática de errores, por lo que nos centraremos en este aspecto. El segundo objetivo será objeto de investigaciones futuras.

La metodología que utilizamos para poder desarrollar los procedimientos de detección automáticos se basa en el análisis exhaustivo del corpus correspondiente a 5

verbos, un total de 500 frases. A partir de este análisis se desarrolló un programa de detección de errores que ha permitido establecer una tipología de errores que presentamos en la sección siguiente.

3.1. Tipología de errores

En la anotación manual se dan diferentes problemas, ya que el anotador humano analiza de una sola vez todos los niveles de análisis de una frase, incluso la selección de los fragmentos a analizar. Podemos identificar diferentes causas de error:

- Lapsus del anotador
- Categorías con definición poco específica en los criterios o inherentemente infraespecificadas
- Error de concepción gramatical

Uno de los errores más frecuente es el de los **constituyentes que no están bien delimitados** por los anotadores, en general esto responde a un lapsus del anotador, por ejemplo en los casos como en (a), en el que la segunda palabra del sintagma "*el uno*" no ha recibido etiqueta de papel semántico.

(a)

Tres de los objetos más brillantes de nuestra noche					se acercarán vertiginosamente
Ag-t-desp					*circunstancial*
SN					*Sadv*
Sujeto					
el	uno	al otro	para cortejar la luna.		
ag t-desp		*dest*	*circunstancial*		
Spron		*SP*	*SP-OInf*		
Sujeto					

Otro caso a tener en cuenta es la **asignación errónea de una categoría** a un constituyente, error que suele producirse por diferentes causas: lapsus como el de (b), donde un sintagma adverbial recibe la función sintáctica de Objeto Preposicional, o por una concepción errónea de la gramática por parte del anotador, como en (c), donde la oración adverbial es erróneamente etiquetada como relativa. Este tipo de error afecta a diferentes categorías, como función sintáctica, papel semántico y categoría sintáctica.

(b)

El cuerpo de bomberos hubiera tenido que actuar inmediatamente.
Agente *Ma*
SN *Sadv*
Sujeto *Obj-Prep-1*

(c)

La coordinadora no dispone de índices de siniestralidad de ciclistas,
T *Sadv-neg* *T*
SN *SP*
Sujeto *Obj Prep-2*

aunque sostiene que se trata de un fenómeno en auge.

Circunstancial
ORel

Otro tipo de error, que clasificamos aquí de diferente forma es el que consiste en utilizar una categoría basándose en **criterios tradicionales** en vez de los criterios desarrollados por el proyecto (Alonso et al 2005). Por ejemplo, las categorías circunstanciales tanto adverbiales como preposicionales, en ocasiones se confunden asignando a un SP temporal una categoría de tipo adverbial y a la inversa, en este caso, parece que el criterio funcional está más activo en el anotador que el categorial. Lo mismo pasa con las locuciones prepositivas o adverbiales que muchas veces presentan dudas al anotador, quizás por una falta de detalle en los criterios establecidos. También clasificamos dentro de este tipo los errores generados por categorías que en los propios criterios no están bien delimitadas conceptualmente, esto en ocasiones produce asistematicidad en la anotación ya que en muchas ocasiones las soluciones alternativas adoptadas también son correctas, un caso claro lo constituyen las oraciones reducidas. También encontramos el caso de construcciones complejas de relativo (preposición + relativo) que alternan entre una anotación como sintagma preposicional o como un pronombre de relativo (esta última asignada según los criterios) (d)

(d)
….hasta llegar al 2060, en el que la cifra bajará a los 10.000.
T-af *T-er*
SP *SN* *SP*
Circunstancial *Sujeto* *Obj-prep-1*

Por otro lado, consideramos errores algunos usos de categorías que aunque estén aceptadas tradicionalmente, en el proyecto no son contempladas. Es el caso de las oraciones completivas o de los pronombres relativos anotados como sintagmas nominales al tener la misma función que estos respecto al verbo.

La falta de anotación se da en algunos casos directamente establecidos por los criterios, como por ejemplo la función sintáctica de algunos adverbios. Sin embargo, encontramos otras faltas o carencias que vienen dadas por lapsus en la anotación. La falta de marca de los núcleos de los sintagmas anotados es muy frecuente, además de olvidos en algún constituyente, como se ve en (e), donde se ha olvidado la función del objeto directo y los tipos de constituyente.

(e)

Los mossos han abierto diligencias
Agente *T*
Sujeto

En el nivel de la anotación de semántica oracional encontramos frecuentes errores, diferentes según la naturaleza de la oración. Las causas de estos errores son difíciles de determinar. El error más frecuente es la anotación de oraciones de verbos estativos como antiagentivas o anticausativas (f).

(f)

Anticausativa
Estado
… permite que Vilanova i la Geltrú alcance en los próximos ocho años
T *SP*
SN *Circunstancial*
Sujeto
los 110.000 habitantes
T
SN
Obj-directo

Siguiendo la tipología presentada, el porcentaje total de errores en el corpus es de un 24,5% aproximadamente, y este error afecta a un porcentaje de 17% de frases en el corpus. En la tabla 1 podemos observar el porcentaje de cada tipo de error detectado:

Tipo de error	Porcentaje
Desambiguación de sentido verbal	**5,4**
Semántica oracional	**4**
Categorías	**51'4**
Papeles semánticos	**2,7**
Funciones sintácticas	**23**
Detección de núcleo	**9,5**
Segmentación errónea de constituyentes	**4,1**

3.2. Detección automática de errores

Como hemos mencionado, uno de los objetivos de esta investigación es conseguir realizar de forma automática la detección de errores en el corpus para su posterior corrección manual. En este sentido estamos actualmente desarrollando un sistema de detección que se basa en dos técnicas conocidas: la detección por co-ocurrencia de características y detección por anotación automática.

Hasta el momento nos hemos centrado en detectar la co-ocurrencia de características incompatibles, mediante heurísticas de búsqueda sobre la información de las etiquetas xml del corpus, como la que se muestra en el ejemplo (g), destacada en rojo. En (g) se buscan todas las ocurrencias de pronombres personales que están anotados funcionalmente como objetos preposicionales. Así, detectamos muchos errores por lapsus o por concepciones erróneas de la gramática.

(g)

<phr id='3' rs='Dest' cat='**PR-Pers**' fs='**Obj Prep-1**' Argumento='1'>
<w Id='16' forma='le' nucleo='1'>
</phr>

Sin embargo, una vez explotado este método nuestra idea es utilizar herramientas de anotación automática para encontrar otros posibles errores, que se escapan al alcance de la técnica anterior. En primer lugar, mediante la herramienta libre de análisis del castellano FreeLing (Carreras et al 2004) podemos captar los errores en la delimitación de los sintagmas, contrastando el análisis de constituyentes de FreeLing con el producido manualmente, de forma que si algún sintagma del corpus es más pequeño que el propuesto por FreeLing, este sintagma será candidato a contener error.

Conclusiones y líneas futuras

En este artículo hemos presentado la situación actual de nuestro trabajo en la detección de errores en la anotación manual del corpus SenSem, a nivel sintáctico-semántico. Hemos clasificado los errores de anotación en una tipología, que es la base para el desarrollo de herramientas de detección y corrección automática de los errores.

Referencias

Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2005). The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish. *Proceedings of the International Conference RANLP*, pp. 39-46. Borovets, Bulgaria.

Brants, T., W. Skut, H. Uszkoreit (1999). Syntactic annotation of a German newspaper corpus. In: *Anne Abeillé: ATALA sur le Corpus Annotés pour la Syntaxe Treebanks*, pp.69-76. Paris, France.

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

Civit, M., A. Ageno, B. Navarro, N. Bufí & A. Martí (2003). Qualitative and Quantitative Analysis of Annotators' Agreement in the Development of Cast3LB. *The Second Workshop on Treebanks and Linguistic Theories.* Växjö, Sweden.

Dickinson, M. (2005). *Error detection and correction in annotated corpora*, PhD Thesis, The Ohio State University.

Kingsbury, P., M. Palmer & M. Marcus (2002). Adding Semantic Annotation to the Penn TreeBank. *Proceedings of the Human Language Technology Conference.* San Diego.

Palmer, M. D.Gildea & P. Kingsbury (2004). The proposition bank: An annotated corpus of semantic roles. In: *Computational Linguistics Journal*, 31:1, 2005.

Palomar, M., M.Civit, A.Diaz, L.Moreno, E.Bisbal, M.Aranzabe, A.Ageno,M.A. Marti & B. Navarro (2004). 3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalan, euskera y castellano. *Proceedings of the XX Congreso Anual de SEPLN*. Barcelona, Spain.