Anotación de roles semánticos en el corpus 3LB

Borja Navarro, Paloma Moreda, Belén Fernández,
Raquel Marcos y Manuel Palomar

Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información.
Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.
Apartado de correos, 99 E-03080. Alicante, Spain.
{navarro, moreda, bfernan, rmarcos, mpalomar}@dlsi.ua.es

Resumen In this paper, the proposal and the method of annotation with semantic roles of 3LB corpus are presented. The semantic roles have been specified bearing in mind the application of the corpus to the development of Question Answering Systems. A semiautomatic method is followed with 3LB-SeRAT tool.

En este trabajo se presenta la propuesta y método de anotación con roles semánticos del corpus 3LB. Los roles semánticos se han especificado teniendo en cuenta el uso del corpus para el desarrollo de sistemas de Búsqueda de Respuestas. Se sigue un proceso de anotación semiautomático con la herramienta 3LB-SeRAT.

Palabras clave: Anotación de corpus, roles semánticos, búsqueda de respuestas, recursos lingüísticos y herramientas.

1. Introducción

En aplicaciones reales de Procesamiento de Lenguaje Natural (PLN) resulta imprescindible la obtención de gramáticas computacionales a partir de amplios corpus anotados con información lingüística. El proyecto 3LB nació con el objetivo de desarrollar tres corpus anotados con información lingüística: uno para el euskera (corpus Eus3LB), otro para el catalán (corpus Cat3LB) y otro para el español (Cast3LB)[1]. Los tres corpus han sido anotados a nivel sintáctico, y actualmente están siendo anotados a nivel semántico con el sentido desambiguado de nombres, verbos y adjetivos; y a nivel pragmático-textual con las relaciones anafóricas [9]. En este trabajo vamos a dar un paso más en la anotación del corpus presentando la anotación con roles semánticos.

Dada una oración como (0):

(0) Los empleados dieron una cerrada ovación al jefe.

[1] Proyecto financiado por el Gobierno Español FIT-150-500-2002-244, FIT-150-500-2003-411 y TIC-2003-07158-C04-01. En este proyecto participan las siguientes universidades: Universidad del País Vasco, Universidad de Barcelona, Universidad Politécnica de Cataluña, Universidad Politécnica de Valencia y Universidad de Alicante.

Corpus	Cast3LB	Cat3LB	Eus3LB
Constituyentes	100.000 palabras	106.000 palabras	-
Funciones	100.000 palabras	53.000 palabras	-
Dependencias	-	-	56.000 palabras

1

Cuadro 1. Datos anotación corpus 3LB a nivel sintáctico

Los empleados tiene el rol "agente", *el jefe* el rol "beneficiario" o "receptor" y *una cerrada ovación* el rol "tema".

Los Sistemas de Búsqueda de Respuestas, por sus características, requieren información lingüística para afrontar con garantías la tarea de localización de la respuesta correcta. Entre la información lingüística requerida, los roles semánticos juegan un papel fundamental. Con esta información se podría responder a preguntas como "quién", "cuándo", "dónde" o "qué". Por ejemplo, las siguientes preguntas podrían ser contestadas con la oración anterior (0): el rol agente responde a la cuestión (1), y el rol tema a la pregunta (2).

(1) ¿Quién dio al jefe una cerrada ovación?

(2) ¿Qué dieron los empleados al jefe?

Este trabajo se estructura del siguiente modo: tras presentar el estado actual del proyecto, expondremos los principios que rigen la anotación de roles (secciones 3 y 4), así como una propuesta de roles semánticos (sección 5). Después presentaremos la herramienta de anotación 3LB-SeRAT y el método de anotación semiautomático (sección 6). El trabajo finalizará con algunas conclusiones y trabajos futuros (sección 7).

2. 3LB corpus

Actualmente, el corpus Cast3LB y el corpus Cat3LB están formado por 100.000 palabras aproximadamente, y el corpus Eus3LB por 300.000 palabras.

A nivel sintáctico, en los corpus Cast3LB y Cat3LB se han anotado los constituyentes sintácticos (oraciones, sintagmas, etc.) y las relaciones funcionales básicas (sujeto, objeto directo, etc.)[1]. En el corpus Eus3LB se ha seguido una anotación de dependencias sintácticas. El cuadro 1 muestra el estado actual de la anotación de cada corpus.

A nivel semántico se ha anotado el sentido de cada verbo, nombre y adjetivo. Para la representación del sentido se ha utilizado el número de sentido del WordNet de cada lengua. Esta representación del sentido es la misma para las tres lenguas, dado que es el número del Interlingua Index de EuroWordNet [10]. El cuadro 2 muestra el estado actual de la anotación.

A nivel pragmático-textual se están anotando las principales anáforas (sujetos elípticos, pronombres personales y clíticos, etc.) y su antecedente.

Corpus	Cast3LB	Cat3LB	Eus3LB
palabras	33.000	5.000	5.000

1

Cuadro 2. Datos anotación corpus 3LB a nivel semántico

3. Principios generales para la especificación de los roles semánticos

A diferencia del nivel sintáctico, donde hay más o menos acuerdo entre la comunidad científica sobre los constituyentes sintácticos y su definición, con los roles semánticos no hay acuerdo alguno sobre qué roles semánticos existen ni las características de cada uno. Por ello, a la hora de anotar un corpus con roles semánticos se debe especificar, primero, qué roles se van a anotar y, después, definir cada uno de ellos. Los principios que seguimos a la hora de establecer estos roles son los siguientes:

1. **Principio de aplicabilidad:** El objetivo de la anotación del corpus con roles semánticos no es demostrar ni justificar ninguna teoría concreta sobre el tema, sino desarrollar un recurso útil para tareas de PLN. Por ello, no pretendemos definir unos roles semánticos universales, sino establecer un conjunto de roles semánticos consensuados y justificados tanto desde un punto de vista teórico como aplicado a partir de los ejemplos del corpus, de los cuales se pueda obtener una anotación consistente. Como otras propuestas de anotación de roles semánticos [8], se intenta desarrollar una anotación teóricamente neutra, que no sigue ninguna teoría en concreto.
   La anotación de roles semánticos del corpus 3LB tiene una aplicación clara a Recuperación de Información y a Búsqueda de Respuestas. Como se expondrá más adelante, los roles semánticos responden a posibles entidades semánticas por las que se puede preguntar en una consulta a partir del verbo.
2. **Principio de generalidad:** Otros proyectos de anotación de roles semánticos anotan, en algunos casos, roles específicos de un solo verbo [8] [3]. En nuestra propuesta, la lista de roles definidos son roles generales, aplicables a diferentes verbos que compartan rasgos semánticos similares (es decir, a toda una clase verbal).
3. **Principio de conexión con otras propuestas de anotación:** Etiquetar el corpus con una lista de roles semánticos propios no serviría de nada si los roles propuestos no están relacionados con los roles de otros corpus con anotación similar. Así, la lista de roles propuesta para el corpus 3LB está basada en los roles generales (*theta roles*) de PropBank [8] y VerbNet [4] (y teniendo en cuenta los utilizados en FrameNet [3]). Estos roles han sido desarrollados para el inglés, y se basan en la clasificación de verbos del inglés desarrollada por B. Levin [5]. Esta propuesta ha sido adaptada a casos concretos del español, manteniendo la relación de cada rol definido para el español con su correspondiente rol en inglés. De esta manera, el corpus anotado resultante quedará relacionado con los corpus similares anotados para el inglés.

4. **Principio de jerarquía:** Al igual que en otros ámbitos de la semántica, como son las relaciones léxicas, y teniendo en cuenta trabajos sobre el tema [2], consideramos que es posible establecer una jerarquía de roles semánticos. Con ello, el conjunto de roles con el que se etiqueta el corpus es más consistente: no es una simple lista de roles que puede asumir un argumento verbal, sino que, según el contexto, puede ser semánticamente más generales o más específicos.
   El nivel más general es aquél que no tiene ninguna información semántica: sólo se indica la presencia de un argumento. En un primer nivel de concreción semántica están los roles de carácter universal, como "Tiempo", "Lugar" o "Modo", junto al conjunto de roles relacionados con el Agente y el conjunto de roles relacionados con el Paciente. En un tercer nivel se sitúan los roles específicos de cada uno de estos: "Causa", "Agente", "Paciente", "Tema", etc. En algunos casos, como se expondrá luego, por debajo de esta nivel aún se especifican subroles.

4. Principios metodológicos para la anotación del corpus

1. Sólo se etiquetan constituyentes explícitos. No se tratará, por tanto, ningún argumento elíptico. La única excepción es el sujeto elíptico, que ya ha sido marcado en la fase de anotación sintáctica.
2. Sólo se anotan constituyentes que en fases anteriores hayan sido anotados con función sintáctica. Con ello, se tienen marcados de antemano todos los constituyentes que son susceptibles de recibir rol semántico, así como su límite.
3. Los roles semánticos anotados se relacionan con el sentido del verbo, no con el verbo (en tanto que palabra). Dado que el sentido del verbo ya está marcado no es necesario desambiguarlo. Los roles quedan relacionados directamente con el verbo y su sentido, no sólo con la palabra.
4. Se sigue un proceso de anotación léxica o transversal. No se anota el corpus oración a oración, sino por sentidos verbales: los roles asociados a todas las apariciones de un sentido verbal son anotados a un mismo tiempo y por el mismo anotador. Con ello se evitan problemas de inconsistencia en la anotación, en la que un mismo verbo pueda estar anotado con criterios diferentes por dos anotadores diferentes.

5. Propuesta de roles semánticos para el corpus 3LB

De acuerdo con los principios y especificaciones indicadas anteriormente, la propuesta de roles semánticos para la anotación del corpus 3LB se especifica en los siguientes roles.

En primer lugar, los roles que suelen actuar como argumentos:

- **Agente-Causa:** Argumento que denota la entidad que desde un punto de vista general produce la acción o evento (o es la principal entidad del estado) expresado en el verbo. En general, este rol responde a la pregunta

"¿Quién?" Si tiene el rasgo [+animado] se considera Agente, y si tiene el rasgo [-animado] se considera Causa. Relacionados con estos roles está también el rol **"Instrumento"**.

- **Tema-Paciente:** Argumento que denota la entidad directamente afectada por el verbo. Suele responder a la pregunta "¿Qué?". Si tiene el rasgo [+animado] se considera Paciente, y si tiene el rasgo [-animado] se considera Tema. Dentro de este grupo se incluye también el rol **"Tópico"**, que hace referencia a lo expresado o pensado en verbos de dicción y pensamiento.
- **Beneficiario-Receptor:** Argumento que denota la entidad que resulta beneficiada o afectada indirectamente por el verbo. Responde a preguntas tipo "¿a/para qué/quién?"

Estos roles se pueden relacionar, a grandes rasgos, con las funciones sintácticas de sujeto, objeto directo y objeto indirecto de las oraciones transitivas, respectivamente.

Otro grupo de roles son aquellos que suelen aparecer como adjuntos (si bien hay determinados verbos que los exigen como argumentos).

- **Tiempo:** Sólo se anota si aparece un sintagma que especifique de manera explícita el tiempo en el que la acción/estado del verbo se desarrolla. Responde a la pregunta "¿cuándo?"
- **Lugar:** Pueden hacer referencia tanto a lugares físicos como a lugares abstractos. Responde a la pregunta "¿dónde?". Este rol se puede especificar en tres sub-roles: origen (lugar "desde donde"), meta (lugar "a donde") y trayectoria (lugar "por donde").
- **Modo:** Es complemento similar a los anteriores que indica el modo o manera en que se lleva a cabo la acción, evento o estado del verbo. Responde a la pregunta "¿cómo?".

Ésta es una lista inicial de roles basados en los fundamentos teóricos anteriores. Estos roles son consecuencia de las necesidades expresadas por sistemas de Búsqueda de Respuestas, y podrían variar durante el proceso de anotación.

6. 3LB-SeRAT. Herramienta de apoyo a la anotación

Durante el transcurso del proyecto se han desarrollado herramientas de ayuda a la anotación de los corpus tanto sintáctica como semánticamente y a nivel de discurso. Un esquema de la arquitectura resultante puede verse en la figura 1.

Puesto que a nivel sintáctico se ha anotado con constituyentes el español y catalán, y con dependencias el euskera, se han desarrollado herramientas diferentes para cada uno. Para el caso de la anotación sintáctica de constituyentes y funciones se ha adaptado el editor de árboles TreeTrans AGTK versión 9.2. Esta adaptación ha consistido en el formato de entrada/salida que permite la entrada en formato PennTreebank (TBF o parentizado) y XML. Para el caso del euskera se ha desarrollado la herramienta computacional 3LBAbarHitz. La utilización de 3LBAbarHitz facilita y agiliza la anotación sintáctica manual del corpus, además

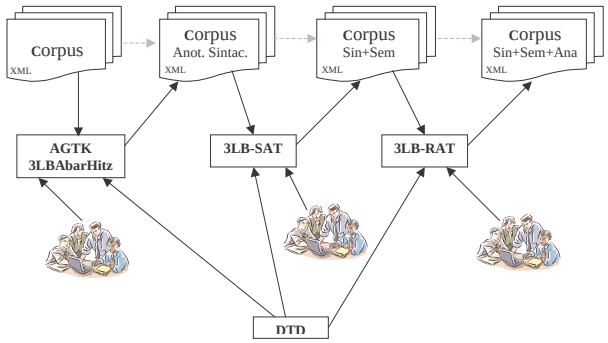

Figura 1. Arquitectura herramientas anotación 3LB

de evitar errores de anotación al controlar el número y tipo de campos descritos en cada etiqueta de dependencia. En ambos casos la oración es visualizada en forma de árbol, donde el anotador puede realizar distintos cambios (supresión de nodos, corrección, etc).

Para el caso de la anotación semántica se ha desarrollado la herramienta 3LB-SAT (3LB-Semantic Annotation Tool). Sus principales características son: (i) está orientada a palabra, (ii) permite introducir el corpus en diferentes formatos (TBF y XML) y (iii) utiliza EuroWordnet para consultar el sentido de las palabras.

A nivel de discurso se ha desarrollado la herramienta 3LB-RAT (Reference Annotation Tool) con el fin de agilizar la anotación y supervisión de anáforas y correferencias. Esta herramienta proporciona al anotador dos modos de trabajo: manual y semiautomático.

Continuando con esta línea, en este trabajo se propone el desarrollo de una nueva herramienta de apoyo a la anotación de roles semánticos: 3LB-SeRAT (Semantic Role Annotation Tool). Al igual que las herramientas anteriores, 3LB-SeRAT proporcionará a las anotadores dos modos de trabajo: manual y semiautomático. En el modo manual, la herramienta simplemente marca en el texto los diferentes verbos y sus constituyentes sintácticos y muestra la lista de roles genéricos considerada en esta anotación. De esta manera, el anotador puede seleccionar el rol adecuado para cada constituyente e indicar, además, el grado de seguridad en la selección mediante el campo del estado (cierto, incierto, en espera).

En el modo semiautomático, la herramienta desambigua el rol de cada uno de los constituyentes sintácticos. El anotador supervisa estos resultados aceptando como válida la propuesta del etiquetador o modificando dicha propuesta. Si la propuesta es aceptada, el anotador simplemente establece el grado de seguridad mediante el campo del estado. Si la propuesta no se acepta, el anotador puede modificar manualmente el rol asignado eligiendo otro de la lista de posibles roles e indicando además el grado de seguridad de la selección.

El módulo de desambiguación de roles de 3LB-SeRAT [6] utiliza estrategias de aprendizaje automático basadas en modelos de probabilidad condicional de Máxima Entropía (ME). Su implementación se llevará a cabo utilizando un método de aprendizaje supervisado que consista en la construcción de clasificadores rol-argumento haciendo uso de un corpus etiquetado sintáctica y semánticamente. Un clasificador obtenido por medio de una técnica de ME consta de un conjunto de parámetros o coeficientes los cuales son estimados utilizando un procedimiento de optimización. Cada coeficiente se asocia a una característica observada en los datos de entrenamiento. El principal propósito es obtener la distribución de probabilidad que maximiza la entropía, es decir, se asume máxima ignorancia y únicamente se consideran los datos de entrenamiento. Por tanto, y tal como se expone en [7], el proceso necesita una fase de aprendizaje/entrenamiento previa, además de la definición de un conjunto de características que extraigan la información a aprender. Para la fase de aprendizaje se utilizará la parte del corpus que haya sido etiquetada de forma manual con roles semánticos. El conjunto de características utilizado hace uso de la información sintáctica, semántica y morfológica disponible, tal y como muestra la siguiente lista:

- palabras con carga semántica (verbos, nombres, adjetivos y adverbios) que forman parte del constituyente (PS);
- categoría gramatical de las palabras con carga semántica que forman parte del constituyente (CG);
- sentido de nombres, verbos y adjetivos que forman parte del constituyente (SP);
- función sintáctica del constituyente (FS);
- posición del constituyente respecto al verbo (anterior o posterior) (PC);
- cadena de sintagmas que forman el constituyente (SC);
- distancia del constituyente al verbo en cuanto a número de palabras (DP);
- distancia del constituyente al verbo en cuanto al número de sintagmas (DS);
- distancia del constituyente al verbo en cuanto al número de constituyentes (DC).

Si consideramos la oración de (3)

(3) Pasó frente a la casa donde vive algún gato,

junto al argumento relativo al sintagma preposicional, la información extraída sería la mostrada en la figura 2.

7. Conclusiones y trabajos futuros

En este trabajo hemos expuesto los principios generales y los principales roles con los que anotar el corpus 3LB. La especificación y anotación de roles tiene una clara aplicación en tareas relevantes de PLN como Búsqueda de Respuestas, y teniendo en cuenta esta aplicación final se ha desarrollado la propuesta. Además

PS	casas, vive. gato
CG	ncfp, vmip, ncms
SP	n02728393, v01499405, n01457160
FS	sp-CC
PC	posterior
SC	prep-sn
DP	0
DS	0
DC	0

Figura 2. Conjunto de características para el argumento *frente a la casa donde vive algún gato*.

se ha presentado la herramienta de anotación semiautomática 3LB-SeRAT y la metodología de anotación semiatomática.

Entre los trabajos futuros, destacan determinados problemas lingüísticos que deben ser solucionados, como los argumentos con dobles funciones y predicación secundaria, o los constituyentes con roles específicos que no han sido considerados, entre otros. Por otro lado, en el proceso de anotación semiautomático de roles semánticos se quiere hacer uso de la información referencial ya anotada.

Referencias

1. M. Civit, Mª.A. Martí, B. Navarro, N. Bufí, B. Fernández y R. Marcos. 2003. Issues in the Syntactic Annotation of Cast3LB. En *4th International Workshop on Linguistically Interpreted Corpora (LINC03), EACL03* Budapest.
2. D. Dowty. 1991. Thematic Proto-roles and Argument Selection. En *Language*, 67.
3. D. Gildea y D. Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245-288.
4. K. Kipper, H. Trang Dang y M. Palmer. 2000. Class-Based Construction of a Verb Lexicon. En *Seventeenth National Conference on Artificial Intelligence (AAAI2000),* Austin, Texas.
5. B. Levin. 1993. *English Verb Clases and Alternations. A Preliminary Investigation.* Chicago, UCP.
6. P. Moreda, M. Palomar y A. Suárez. 2004. Assignment of Semantic Roles based on Word Sense Disambiguation. En *Proceedings of the IX Ibero-American Conference on Artificial Intelligence (IBERAMIA),* Puebla, México.
7. P. Moreda, M. Fernández, M. Palomar y A. Suárez. 2004. Identifying Semantic Roles Using Maximum Entropy Models. En *Proceedings of the Seventh International Conference on TEXT, SPEECH and DIALOGUE (TSD),* Brno, Czech Republic.
8. M. Palmer, D. Gildea y P. Kingsbury. 2004. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics.* Submitted.
9. M. Palomar, M. Civit, A. Diaz, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, Mª.A. Martí y B. Navarro. 2004. 3LB: Contrucción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano. En *Procesamiento del Lenguaje Natural*, 33.
10. P. Vossen. 1998. A Multilingual Database with Lexical Networks. Kluwer Academic Publisher.