Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 137-140 recibido 31-03-2017 revisado 12-05-2017 aceptado 29-05-2017

Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con WordNet 3.0

Design and development of the Galician SemCor corpus semantically tagged with WordNet 3.0

Miguel Anxo Solla Portela, Xavier Gómez Guinovart
Grupo TALG - Universidade de Vigo
Campus Universitario, 36310 Vigo
{miguelsolla, xgg}@uvigo.es

Resumen: En esta presentación describimos la metodología utilizada para la creación del Corpus SensoGal, un corpus paralelo inglés-gallego etiquetado semánticamente con WordNet 3.0 y basado en el SemCor de la lengua inglesa.
Palabras clave: SemCor, WordNet, corpus paralelos, anotación semántica

Abstract: In this presentation, we review the methodology used in the development of the SensoGal Corpus, an English-Galician parallel corpus semantically tagged with WordNet 3.0 and based on the English SemCor.
Keywords: SemCor, WordNet, parallel corpora, sense tagging

1 Introducción

En este artículo[1] se describe la metodología utilizada para la creación del Corpus SensoGal[2], un corpus paralelo inglés-gallego etiquetado semánticamente con WordNet 3.0 y basado en el corpus SemCor de la lengua inglesa. La construcción de este recurso se realiza en el marco del proyecto *TUNER*, enfocado al desarrollo de recursos multilingües (inglés, español, catalán, vasco y gallego) para el procesamiento de documentos en dominios específicos mediante tecnologías lingüísticas de base semántica. En relación con el gallego, los objetivos del proyecto incluyen el desarrollo del WordNet para la lengua asociado con el Multilingual Central Repository (MCR) (González Agirre, Laparra, y Rigau, 2012), y la construcción de un corpus etiquetado semánticamente del gallego alineado con el corpus SemCor del inglés (Landes, Leacock, y Tengi, 1998).

2 Alineamientos con SemCor

El corpus SemCor del inglés es un corpus textual anotado semánticamente a nivel léxico. Las palabras de este corpus están etiquetadas con una indicación del sentido concreto que poseen en su contexto de aparición. Las anotaciones indican los sentidos establecidos en la versión 1.6 del WordNet del inglés, un recurso léxico elaborado por el mismo equipo de la Universidad de Princeton que llevó a cabo la anotación del corpus SemCor (Miller et al., 1990).

El SemCor está formado por 360.000 palabras repartidas entre 352 textos tomados del Corpus Brown. Se trata del mayor corpus general de una lengua anotado semánticamente y de libre acceso, con 192.639 palabras con significado léxico (nombres, verbos, adjetivos y adverbios) anotadas con su sentido respecto a WordNet[3]. De estos 352 textos, tan solo 186 están completamente anotados con categoría gramatical, lema y sentido, mientras que en 166 solo están anotados semánticamente los verbos.

Existen diferentes proyectos de creación de corpus paralelos alineados con el SemCor del inglés, entre los que destaca el corpus MultiSemCor inglés-italiano, compuesto en su versión 1.1 por 116 textos en inglés

[1]Esta investigación se lleva a cabo en el marco del Proyecto de Investigación *TUNER* (TIN2015-65308-C5-1-R) financiado por el Ministerio de Economía y Competitividad del Gobierno de España y el Fondo Europeo para el Desarrollo Regional (MINECO/FEDER, UE).

[2]`http://sli.uvigo.gal/SensoGal/`

[3]Con respecto al SemCor, el corpus de glosas anotadas del WordNet del inglés, también elaborado por el equipo de la Universidad de Princeton, es mayor cuantitativamente, pero al ser un corpus de definiciones contiene texto de un registro metalingüístico de características muy específicas, por lo que debe ser considerado propiamente un corpus especializado.

ISSN 1135-5948

totalmente etiquetados del SemCor junto a sus correspondientes traducciones en italiano. Los textos italianos del MultiSemCor están alineados a nivel de frase con los del inglés, y anotados con categoría gramatical, lema y sentido. Se realizó un alineamiento automático a nivel de palabra y, a partir de este alineamiento, se proyectaron automáticamente sobre las palabras italianas los sentidos léxicos anotados en el inglés. De este modo, se logró proyectar un 77,14 % (92.420) del total de los tokens anotados semanticamente del inglés (119.802), quedando sin correspondencia en italiano el 22,86 % restante (27.382)[4]. Posteriormente, se ha incorporado también a MultiSemCor la traducción de doce textos del SemCor original al rumano.

Por otro lado, el SemCor paralelo del japonés, JSemCor[5], se ha elaborado a partir de los mismos textos usados en el MultiSemCor inglés-italiano. Tras el alinamiento a nivel de frase se llevó cabo la proyección manual de los sentidos léxicos anotados en inglés, etiquetando los tokens del japonés con respecto al WordNet 3.0 y dejando sin correspondencia un 39 % de los sentidos (Bond et al., 2012).

Nuestro objetivo, dentro del proyecto *TUNER*, es la construcción de un corpus paralelo SemCor del gallego, el corpus SensoGal, etiquetado semánticamente con referencia a Galnet[6] –el WordNet 3.0 del gallego que forma parte de la distribución del Multilingual Central Repository– y basado en la traducción al gallego de los 186 textos completamente anotados del SemCor del inglés original de Princeton, priorizando los textos ya disponibles en MultiSemCor. En los siguientes apartados, trataremos de explicar concisamente la metodología diseñada para la elaboración del corpus SensoGal.

3 Construcción del corpus

El proceso de creación del SensoGal se inicia con la adaptación automática al WordNet 3.0 de las etiquetas semánticas del SemCor. A continuación, se realiza la traducción manual al gallego de los textos y, simultáneamente, se introducen en el WordNet del gallego las nuevas variantes derivadas de la traducción. Tras la traducción, se proyectan en los textos

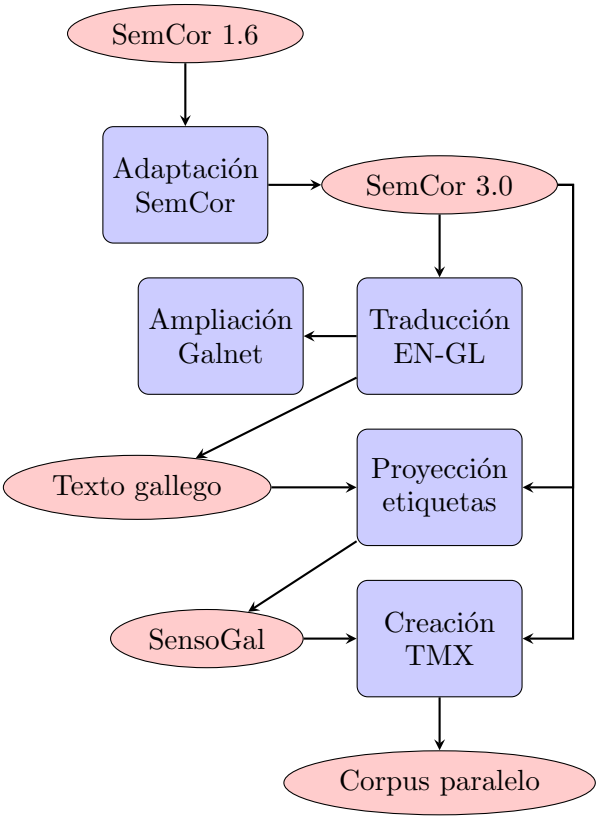

Figura 1: Proceso de elaboración del corpus

en gallego las etiquetas semánticas del inglés. Finalmente, se construye un corpus paralelo inglés-gallego en TMX con el resultado de la anotación semántica del gallego. Los detalles de este proceso, ilustrado de modo esquemático en la Figura 1, se presentan en los siguientes apartados de esta sección.

3.1 Adaptación del SemCor a WordNet 3.0

La construcción del SensoGal se abordó partiendo del SemCor 3.0 distribuido por Rada Mihalcea[7], que cuenta con la etiquetación de los sentidos en el formato de *Sense Keys* de WordNet 3.0. Sin embargo, observamos que algunos errores en la identificación de los sentidos en este corpus presentaban dificultades para la traducción humana y etiquetación semántica del texto gallego de destino. Por este motivo, se decidió emprender una nueva anotación del SemCor inglés a partir de su versión original 1.6[8], etiquetada con *Sense Keys* de WordNet 1.6, y proyectarla a *Inter-Lingual Index* (ILI) de WordNet 3.0 a través

[4]Datos disponibles en `http://multisemcor.fbk.eu/statistics.php`

[5]Disponible en `http://nlpwww.nict.go.jp/wn-ja/data/jsemcor/jsemcor-2012-01.tgz`

[6]`http://sli.uvigo.gal/galnet/`

[7]Disponible en `http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor`

[8]Disponible también en `http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor`

de un nuevo *mapping*. Este *mapping* solo tiene en cuenta los 34.960 *Sense Keys* empleados en el SemCor 1.6 y ha sido elaborado en tres etapas:

1. Identificación automática de la coincidencia en WordNet 1.6 y 3.0 del lema, categoría y glosa o, alternativamente, detección de una correspondencia unívoca (1.6/3.0) en el *Sense Key Index*[9]. De este modo obtenemos 26.269 alineamientos, lo que representa el 75,14 % del total.

2. Identificación del ILI de los lemas que, conforme a su categoría, solo tienen un sentido en WordNet 3.0 (separando, para su revisión, los alineamientos con lemas que en WordNet 1.6 no son monosémicos). Obtenemos así 7.438 alineamientos, lo que representa el 21,28 % del total.

3. Revisión humana de 1.254 de casos no resueltos en las dos fases anteriores (3,58 % del total), con el apoyo de los *mappings* elaborados por el Grupo TALP[10].

Tras estos procesos quedan sin asignar 263 *Sense Keys* (0,68 % del total), algunos de ellos irresolublemente, dada la supresión en WordNet 3.0 de *synsets* de marcado sentido gramatical.

Como resultado se ha obtenido una versión de SemCor (*SemCor-ILI*) que guarda mayor fidelidad con la anotación inicial de los sentidos. En la Tabla 1 se refleja la cantidad de *tokens* con anotación semántica, próxima a la totalidad de los incluidos en SemCor 1.6 (que cuenta con 709 *tokens* anotados con más de un sentido), en contraste con la cantidad de *Sense Keys* que figuran en SemCor 3.0 y que son compatibles con WordNet 3.0.

SemCor 1.6	234.136	100 %
SemCor 3.0	224.136	95,98 %
SemCor-ILI	233.148	99,58 %

Tabla 1: *Tokens* con anotación semántica

Cualitativamente, se han eliminado del *mapping* aquellos casos en los que no existe una coincidencia con el sentido en la versión 3.0 de WordNet. Se trata de un número reducido de casos en los que la anotación remitía a *synsets* de contenido predominantemente gramatical que se han suprimido con posterioridad, como verbos modales o locuciones prepositivas. Así se garantiza que todas las anotaciones en el corpus posean una correspondencia en WordNet.

En un número reducido de casos se ha mantenido la anotación pese a que el lema, por criterios ortográficos, ha desaparecido del *synset* de WordNet. La razón es que en estos casos se ha considerado pertinente mantener la anotación semántica para poderla heredar en el texto producido en la traducción al gallego.

3.2 Traducción y anotación del corpus SensoGal

La traducción humana del SemCor inglés al gallego se lleva acabo utilizando la versión del SemCor enlazada con Galnet y la interfaz de desarrollo del WordNet del gallego como recursos de referencia. El objetivo es que, en la medida de lo posible, el texto resultante mantenga una correspondencia con las secuencias anotadas en el original. Esta traducción controlada no implica necesariamente un alto grado de literalidad en el texto de destino, pero sí requiere cierta destreza estilística para mantener la misma categoría gramatical entre los lemas de las dos lenguas con equivalencia semántica. Durante el proceso traductivo se identifican los lemas utilizados en el texto gallego que todavía no están presentes en Galnet y se incluyen, a continuación, como variantes en el *synset* correspondiente.

Para aligerar la tarea de anotación semántica del texto gallego traducido y preservarla de errores humanos, se diseñó una aplicación que proyecta la etiquetación semántica desde el texto original al traducido. La aplicación deja sin resolver ciertas correspondencias que requieren una posterior intervención humana. El algoritmo procesa cada frase del texto analizando las etiquetas semánticas de la frase en inglés, una a una, y se comporta de forma diferente cuando detecta entidades –que en SemCor están identificadas con el sentido correspondiente a los lemas *persona*, *grupo* y *lugar*– a cuando identifica formas léxicas.

En el caso de las entidades, la anotación semántica solo se proyecta si existe coincidencia en la forma escrita entre inglés y gallego, con un algoritmo de substitución re-

[9] `https://github.com/ekaf/ski`

[10] `http://nlp.lsi.upc.edu/tools/download-map.php`

lativamente simple que, sin embargo, obtiene un índice de éxito en la transposición de aproximadamente el 90 % de los casos; en los casos en los que no lo consigue, indica con una marca que la anotación del gallego todavía está pendiente. Para el análisis de las demás etiquetas, el algoritmo utiliza diferentes resultados de la etiquetación de las frases gallegas proporcionados por el análisis de FreeLing [11] con sus dicionarios de sentidos y de términos pluriléxicos actualizados con las variantes procedentes de la traducción. Cada etiqueta semántica del original se proyecta al texto de destino según los siguientes pasos:

1. Si la etiqueta y su lema en Galnet coinciden con el análisis de una forma léxica de la frase en FreeLing, se proyecta una única vez la anotación sobre la forma léxica, comprobando que no haya sido ya etiquetada.

2. Si no se ha conseguido la proyección con el procedimiento anterior, se comprueba si la misma coincidencia que en la fase anterior se produce cuando la salida de FreeLing muestra todas las posibilidades de análisis morfológico de las formas léxicas de la frase. En caso afirmativo, se proyecta la anotación, previa comprobación de que la forma léxica no haya sido etiquetada con anterioridad.

3. Cuando la etiqueta no se ha podido cotejar con la salida de FreeLing, se hace una búsqueda directa en el archivo fuente con los lemas y etiquetas que utiliza FreeLing. De este modo, se identifican casos como nombres propios o lemas que tienen un sentido con una categoría gramatical en el diccionario principal de FreeLing diferente a la de WordNet.

En caso de que el algoritmo no logre identificar la forma léxica con estos procedimientos, se indica que la notación está pediente; sin embargo, el éxito de este procedimiento es prácticamente del 100 % de los casos.

4 Resultados y perspectivas

Se ha desarrollado una interfaz de consulta del corpus SemCor reetiquetado con ILIs de WordNet 3.0 y enlazado con Galnet a la que se puede acceder desde `http://sli.uvigo.gal/SemCor/`. Por otra parte, el SemCor reetiquetado y el *mapping* utilizado para su elaboración, se encuentran disponibles para descarga en `http://sli.uvigo.gal/download/`.

Hasta el momento, se han etiquetado semánticamente y alineado con el inglés 30 textos del SemCor, totalizando 2.734 unidades de traducción, con 61.236 palabras en inglés y 62.577 en gallego. El corpus paralelo resultante puede ser ya consultado a través de una interfaz web de consulta en `http://sli.uvigo.gal/SensoGal/`. Así mismo, las frases del corpus en gallego se emplean como ejemplos de uso de las variantes en la interfaz de consulta de Galnet.

Aunque hace falta aún mucho esfuerzo para finalizar esta tarea, el corpus SemCor del gallego representa sin duda un recurso de vital importancia para el desarrollo de las tecnologías lingüísticas en esta lengua. Su explotación adecuada debe permitir la construcción de herramientas de gran interés en el ámbito del procesamiento semántico, especialmente en tareas que requieran conocimiento plurilingüe, y la generación de aplicaciones más eficientes para el procesamiento del lenguaje.

Bibliografía

Bond, F., T. Baldwin, R. Fothergill, y K. Uchimoto. 2012. Japanese SemCor: A Sense-tagged Corpus of Japanese. En *Proceedings of the 6th Global WordNet Conference*, Matsue. GWN.

González Agirre, A., E. Laparra, y G. Rigau. 2012. Multilingual Central Repository version 3.0. En *Proceedings of the 6th Global WordNet Conference*, Matsue. GWN.

Landes, S., C. Leacock, y R.I. Tengi. 1998. Building Semantic Concordances. En C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, Cambridge. The MIT Press.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, y K. Miller. 1990. WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244.

[11] `http://nlp.cs.upc.edu/freeling/`