Crespo Miguel, Mario (2020)
Automatic corpus-based translation of a Spanish FrameNet medical glossary

Sevilla
Editorial Universidad de Sevilla, colección Lingüística, Nº 65
ISBN 978-84-472-3005-1
728 págs.

La monografía *Automatic corpus-based translation of a Spanish FrameNet medical glossary* se presenta como un aporte de gran relevancia al análisis semántico computacional en español. Como es conocido, la lingüística computacional se ha centrado tradicionalmente en el análisis y desarrollo de recursos lingüísticos en inglés y, aunque en los últimos años la situación ha cambiado y se han desarrollado proyectos en otras muchas lenguas, queda aún mucho camino por delante. Esta realidad hace de este trabajo una herramienta aún más valiosa.

El estudio trabaja con dos de los recursos más importantes dedicados a la semántica léxica: FrameNet y WordNet. A grandes rasgos, lo que se propone es, por un lado, llevar a cabo la selección de los marcos de FrameNet que pertenecen al dominio de la medicina y, por otro, enlazar estos marcos de FrameNet con *synsets* –equivalentes– aportados por WordNet, para, finalmente, traducir la base de datos al español.

La monografía está dividida en seis capítulos, de los que el primero es la introducción; el segundo recoge el marco teórico; el tercero, cuarto y quinto conforman el desarrollo de la traducción del glosario y los experimentos con él realizados y el sexto se dedica a las conclusiones y a la presentación de las unidades léxicas traducidas al español. Le siguen las referencias bibliográficas y los anexos.

Como se indica en el capítulo introductorio, la motivación de este trabajo viene dada por la escasez de estudios que desde el procesamiento del lenguaje natural se han dedicado a la semántica. La razón de esta escasez en la lingüística computacional puede atribuirse al hecho de que el significado no puede observarse directamente. La semántica, a diferencia de la morfología o la sintaxis, es un nivel lingüístico más difícil de formalizar y, por tanto, de procesar automáticamente. Los objetivos, por su parte, se fundamentan a partir de la constatación de que la mayoría de los proyectos enfocados hacia la semántica computacional se han creado para el inglés, principalmente porque la mayoría de los enfoques modernos de la semántica léxica computacional surgieron en Estados Unidos. Esta situación está cambiando y algunos de estos proyectos se han ampliado posteriormente a otras lenguas; por ello, uno de los principales propósitos de este trabajo es investigar la posibilidad de extender estos recursos a otras lenguas como el español. Además, dado que FrameNet pretende expli-

car cómo las lenguas dan cuenta lingüísticamente de las situaciones cotidianas, el trabajo se centra en los marcos que mejor representan el dominio de la medicina, y presenta un método estadístico que, asistido por las asociaciones de palabras propuestas por WordNet, puede crear una selección y traducción de FrameNet médica para el español y también puede mejorar la cobertura de activación de la FrameNet inglesa. Los resultados, además, son comprobados manualmente para evaluar la fiabilidad del sistema.

La introducción recoge asimismo un resumen de la principal aportación del trabajo: desarrollar un método para emparejar los predicados de cada marco con los *synsets* de WordNet utilizando la información contextual proporcionada por un corpus representativo (corpus COCA). Este método se ha utilizado para desambiguar los predicados de FrameNet según WordNet. Una vez realizado el emparejamiento, toda la información que proporciona WordNet puede utilizarse tanto para traducir la unidad a otros idiomas como, en este caso, para ampliar la cobertura de FrameNet con nuevas unidades.

El capítulo segundo revisa diferentes recursos léxico-semánticos que están actualmente disponibles o en proceso de creación. La mayoría de estos recursos pretende mostrar la interfaz entre la información léxica y la sintáctica. De los recursos empleados en semántica léxica descritos en la obra, vamos a fijaros en los dos más importantes para este trabajo: WordNet y Framenet. WordNet (Miller *et al.*, 1993) es una base de datos léxica que organiza el vocabulario del inglés en grupos según conceptos y relaciones semánticas. Estos grupos, llamados *synsets*, pretenden reflejar la disponibilidad léxica en inglés para un determinado concepto. Los *synsets* son series de sinónimos cognitivos, cada uno de los cuales expresa un concepto distinto. WordNet contiene unas 150.000 palabras divididas en sustantivos, verbos, adjetivos y adverbios, y reunidas en grupos que representan conceptos. Se compone de más de 115.000 grupos de palabras o *synsets*, por lo que puede considerarse un tesauro u ontología lingüística más que un diccionario. Los *synsets* están etiquetados numéricamente, lo que permite ordenar los conceptos de una determinada lengua. Además, no se describen de forma aislada, sino que están interconectados según diferentes tipos de relaciones formales y semánticas. El resultado de estas relaciones conceptuales-semánticas y léxicas es una red de palabras y conceptos significativamente relacionados.

Por su parte, FrameNet (Baker *et al.*, 1998) es un recurso en línea para el inglés basado en la semántica de marcos y respaldado con pruebas de corpus (Ruppenhofer *et al.*, 2006). El objetivo de este proyecto es documentar la gama de posibilidades combinatorias semánticas y sintácticas de cada palabra en cada uno de sus sentidos. Siguiendo el concepto de "marco" definido por Fillmore (1977), un marco se fundamenta en que ciertas palabras evocan ciertas situacio-

nes en las que tienen lugar determinados participantes. La palabra que evoca un marco concreto se denomina palabra "objetivo" o "predicado". Al fijarnos en un dominio temático concreto como la medicina y tratar de describirlo en términos de FrameNet, se obtienen marcos como CURE, formado por palabras como cure.v, heal.v o palliative.a o MEDICAL CONDITIONS con unidades léxicas como arthritis.n, asphyxia.n o asthma.n. Diferentes palabras objetivo pueden evocar funciones semánticas similares si pertenecen al mismo marco y, a su vez, una palabra puede situarse en diferentes marcos si sus significados evocan situaciones o marcos diferentes. Las relaciones de los marcos pueden considerarse como una especie de ontología con las siguientes relaciones entre los diferentes marcos (tal y como se describe en las páginas 120-122):

1. Inheritance (is-a relation). Es la relación más común en las ontologías. Los submarcos son un subtipo de los marcos padre. Es el caso de vivo_o_muerto como subtipo de Estado. En FrameNet, esta relación se divide en Inherits_From and Is_Inherited_By.
2. Using. El marco hijo presupone el marco padre como background: medical_conditions usa observable_bodyparts. Esta relación se divide en Uses y Is_Used_By.
3. Subframes. El marco hijo es un sub-evento de un evento complejo representado por el padre: Criminal_process tiene como submarco Arrest. Esta relación puede dividirse en Subframe_of y Has_Subframes.
4. Perspective on. El marco hijo proporciona una perspectiva particular sobre un marco padre no perspectivizado: Hiring y Get_a_job, que perspectiviza el marco Employment_start. Esta propiedad tiene dos grupos: Perspective_on y Is_perspectivized_in.
5. Precedes. Estipula una relación temporal entre diferentes situaciones. Se compone de dos tipos: Precedes and Is_Preceded_by.
6. Causativity. Un marco es la causa de una situación determinada: Cause_change_of_position_on_a_scale y Change_position_on_a_scale.

Sobre estas bases teóricas, FrameNet ejemplifica las posibilidades sintácticas y semánticas de los *triggers* de un determinado marco. Junto con otros proyectos, FrameNet intenta mostrar cómo se combina la información sintáctica y semántica en una frase. Formalmente, las anotaciones de FrameNet son conjuntos de tres elementos que representan las realizaciones de elementos del marco para cada oración anotada, cada una de las cuales consiste en un nombre de elemento del marco (por ejemplo, comida), una función gramatical (por ejemplo, objeto) y un tipo de frase (por ejemplo, un sintagma nominal).

El capítulo 3 presenta el enfoque general del trabajo. El desarrollo de WordNet para otras lenguas ha hecho posible su uso para encontrar equivalencias léxicas, así que este recurso será el elemento clave en la traducción de FrameNet al español. La cuestión principal para ello es identificar a qué *synsets* apuntan los predicados o *triggers* de FrameNet. Si se pueden identificar, entonces la traducción es posible. Particularmente, en este capítulo se muestra que los marcos de FrameNet pueden clasificarse según ciertos dominios o temas y se presenta un enfoque basado en corpus que selecciona los marcos que resultan significativos tras un test de hipótesis. Para disponer de una lista fiable de ocurrencias de palabras que representen textos médicos, el trabajo emplea el corpus COCA (Davies, 2019). El Corpus of Contemporary American English (COCA) es uno de los mayores corpus de inglés de libre acceso. Contiene más de 560 millones de palabras de texto (20 millones de palabras cada año 1990-2007) y está dividido por igual entre textos hablados, de ficción, revistas populares, periódicos y textos académicos [20]. Las palabras han sido lematizadas y etiquetadas (POS-tagged). El corpus COCA contiene alrededor de 60.000 lemas.

Por tanto, el proceso que se lleva a cabo es el siguiente (véanse páginas 205-209):

> 1) Selección de un corpus representativo de inglés médico y general.
> 2) Selección automática de los marcos pertenecientes al ámbito médico mediante una prueba de hipótesis. Los resultados de la selección automática de marcos se comparan con un *benchmark* o selección manual de marcos, de manera que se asegura que los resultados son relevantes. Por último, se adopta el modo de selección de marcos que más se acerque a nuestro punto de referencia.
> 3) Traducción automática de los marcos al español. Para ello se utiliza WordNet y el procedimiento se basa también en corpus. Como dijimos anteriormente, WordNet organiza las unidades léxicas de una determinada lengua como una ontología en la que los términos que se refieren al mismo concepto (sinónimos y cuasi-sinónimos) comparten el mismo *synset*. Si somos capaces de unir los activadores de FrameNet con los *synsets* de WordNet, es posible traducir dichos elementos de FrameNet a otros idiomas, ya que el inglés y el español están unidos por *synsets*.

En el siguiente capítulo, el cuarto, se explican los diferentes experimentos realizados para la selección de marcos. Se presenta un método automático para la selección de marcos de FrameNet en función del tema de un determinado texto o corpus. Para comprobar que los resultados obtenidos son coherentes, se realiza una selección manual de marcos o *benchmark* para analizar el grado de coincidencia entre la selección automática y el resultado esperado. Los resultados se

analizan mediante F-score, una medida muy utilizada en este tipo de aplicaciones. A continuación, se aplican diferentes pruebas y técnicas estadísticas sobre los marcos, *triggers* o predicados para mejorar los resultados obtenidos, y se amplía cada grupo de marcos con la información léxica proporcionada por EuroWordNet. Por tanto, se observa que las relaciones léxicas de WordNet pueden ayudar en la selección de marcos. Por otro lado, se obtiene una puntuación F de 0,87 según el *benchmark*, lo que demuestra la aplicabilidad de este tipo de enfoque automático.

El capítulo 5 explora la mejor manera de traducir la selección de marcos al español mediante EuroWordNet. Para ello, primero se asocia cada *trigger* a todos los *synsets* en los que aparece en WordNet. Algunas palabras están asociadas a un solo *synset*, por lo que, en principio, podrían traducirse al español directamente, ya que EuroWordNet muestra cómo las distintas lenguas formalizan léxicamente la gama de conceptos de *synset*. A continuación, se analizan los problemas asociados. Para los activadores de FrameNet asociados a más de un sintagma, se han realizado varios experimentos. Para comprobar el rendimiento del sistema, se desambigua el punto de referencia empleado manualmente y se compara lo cerca que estaba la desambiguación automática de nuestra selección manual. Los resultados finales muestran una puntuación F del 85% en la asociación correcta de los activadores o predicados a los *synsets* de WordNet.

Por último, el capítulo sexto recoge los resultados y las conclusiones. Este capítulo está en su mayor parte constituido por la traducción al español de los *synsets* de marcos relacionados con la medicina. Para hacer esta lista, el autor indica que, tras la traducción automática llevada a cabo con EuroWordNet, comprobó manualmente que los *synsets* estuvieran bien traducidos: un 95,6% lo estaban. A continuación filtró dichas unidades para resaltar las que tienen una distribución de frecuencia estadísticamente significativa en el corpus de textos médicos creados para el español. Además, ofrece algunos ejemplos tomados de la base de datos bibliográfica médica Medline.

Tras las referencias, en los anexos se recoge la selección manual de *synsets* sobre tema médico y la ampliación de los *triggers* de FrameNet con los nuevos términos que ha aportado WordNet.

Estamos, pues, ante un trabajo sólido, sostenido mediante técnicas de corpus y avalado por pruebas estadísticas, de verdadero interés para el desarrollo de la semántica computacional en español, un campo que tiene una gran utilidad tanto para la traducción automática como para la terminología, la lexicografía y, en general, cualquier proceso automático en el que deba tenerse en cuenta el contenido léxico.

REFERENCIAS

BAKER, C., FILLMORE, C. J. & LOWE, J. B. (1998): "The Berkeley FrameNet project". Boitet, C. & Whitelock, P. (eds.), *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*. San Francisco, California: Morgan Kaufmann Publishers, pp. 86-90.

DAVIES, M. (2019): *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present.* Disponible en https://www.english-corpora.org/coca/.

FILLMORE, C. J. (1977): "Scenes and Frames Semantics". *Z*ampolli, A. (ed.), *Linguistic Structures Processing*. Amsterdam: North Holland, pp. 55-82.

MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D. & MILLER, K. (eds.) (1993): *Five Papers on WordNet, cls report 43. Tecnical report.* New Jersey: Cognitive Science Laboratory. Princeton University.

RUPPENHOFER, J., ELLSWORTH, M., PETRUCK, M. R. L., JOHNSON, C. & SCHEFFCZYK, J. (2006): *FrameNet II: Extended Theory and Practice*. URL: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf.

Marta Sánchez-Saus Laserna
Departamento de Filología
Instituto Universitario de Investigación en Lingüística Aplicada
Universidad de Cádiz
Avda. Dr. Gómez Ulla s/n
11003 Cádiz
marta.sanchezsaus@uca.es

Fecha de Recepción: 03/09/2021
Fecha de Publicación: 01/12/2021 DOI: http://dx.doi.org/10.25267/Pragmalinguistica.2021.i29.28