Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 345-346 recibido 03-05-2011 aceptado 21-05-2011

Procesamiento automático de metáforas con métodos no supervisados*

Automatic metaphors processing through unsupervised methods

B. Navarro-Colorado, D. Tomás, S. Vázquez, P. Moreda, R. Izquierdo, E. Saquete y F. Llopis
Grupo de Procesamiento del Lenguaje (GPLSI) - Universidad de Alicante
Carretera San Vicente del Raspeig s/n 03690
{borja, dtomas, svazquez, moreda, ruben, stela, llopis}@dlsi.ua.es

Resumen: Proyecto emergente centrado en la detección e interpretación de metáforas con métodos no supervisados. Se presenta la caracterización del problema metafórico en Procesamiento del Lenguaje Natural, los fundamentos teóricos del proyecto y los primeros resultados.
Palabras clave: metáforas, semántica, corpus, clustering

Abstract: The main objective of this project is the identification and interpretation of metaphors through unsupervised methods. We present the metaphor processing problem in Natural Language Processing, the project background and first results.
Keywords: metaphors, semantics, corpus, clustering

1. Caracterización del problema

La metáfora es un fenómeno lingüístico ubicuo y resbaladizo. Aparece en cualquier tipo de texto y es determinante para su correcta interpretación. Pero la interpretación automática de la metáfora es compleja ya que establece relaciones semánticas y conceptuales no convencionales de alto poder significativo. En palabras de Goatly (1997, p. 8): "Metaphor occurs when a unit of discourse is used to refer unconventionally to an object, process or concept, or colligates in an unconventional way".

Para el procesamiento automático del fenómeno metafórico se han especificado dos pasos: la detección de la expresión metafórica y su interpretación (Shutova, 2010). Éste es el más complejo, pues se debe determinar automáticamente el significado de la expresión metafórica asumiendo que ninguno de los sentidos literales es el correcto. Dada la complejidad del fenómeno, hoy día no hay una aproximación bien definida al problema. La mayoría de los sistemas de Procesamiento del Lenguaje Natural (PLN) asumen la "metáfora cognitiva" de Lakoff y Johnson (2004 (1980)), cuya interpretación se basa en determinar la relación conceptual subyacente a la metáfora.

2. Marco teórico y metodológico

En este proyecto queremos explorar nuevas perspectivas. Como marco teórico asumimos los estudios sobre la metáfora que se están desarrollando en Lingüística de Corpus (Stefanowitsch y Gries, 2006). Éstos consideran únicamente la metáfora como un fenómeno lingüístico, independientemente de que sea o no reflejo de una relación conceptual subyacente. Así, analizan la metáfora a partir de rasgos y relaciones lingüísticas.

Como marco metodológico seguimos aproximaciones no supervisadas (*clustering*) a la semántica computacional, basados sobre todo en el Modelo de Espacio Vectorial (Turney y Pantel, 2010). Este modelo asume que el significado de las palabras y expresiones lingüísticas depende de su uso en un contexto determinado, por lo que una palabra tendrá tantos sentidos como contextos de aparición diferentes (hipótesis distribucional). Los usos metafóricos, con ello, tendrán contextos de aparición específicos y diferentes a los usos literales. Esta agrupación de sentidos en función de los diferentes contextos es lo que pretendemos modelar con técnicas de *clustering*, tratando de determinar los contextos metafóricos.

3. Objetivos científicos

En tanto que proyecto emergente, nos proponemos sentar las bases para un proyecto

* Proyecto financiado por la Universidad de Alicante (GRE09-31) y por la Generalitat Valenciana (GV/2011/040), dentro del programa de ayudas a proyectos emergentes.

futuro de mayor envergadura, trabajando en los siguientes objetivos:

- Proponer un modelo formal de representación del significado metafórico basado en rasgos lingüísticos (palabras).
- Crear un corpus piloto anotado a mano con significados metafóricos, como base para un futuro *goldstandar* de evaluación de sistemas de detección e interpretación de metáforas.
- Analizar diferentes algoritmos no supervisados y diferentes configuraciones de rasgos contextuales para la detección e interpretación de usos metafóricos a partir de sus contextos de aparición.

4. Primeros resultados y líneas de desarrollo

En los seis primeros meses de proyecto, se ha trabajado en la creación del corpus y en el análisis de algunos algoritmos.

En primer lugar, se ha creado un pequeño corpus de evaluación en inglés y español. El corpus lo forman oraciones extraídas de los corpus del *Cross Language Evaluation Forum* (CLEF) que contienen usos literales y metafóricos dentro de un patrón sintáctico tipo (ESP) "N de N" o (ING) "N of N" con las palabras "desierto de" (512 oraciones en español, 2010 en inglés), "oasis de" (141 español, 173 inglés), "tormenta de" (2750 español, 3681 inglés) y "torrente de" (74 español, 80 inglés). Hasta ahora, de este pequeño corpus se han marcado a mano los usos literales y metafóricos de la palabra (ESP) "torrente" (46 casos metafóricos y 28 no metafóricos), (ING) "torrent" (58 casos metafóricos y 22 no metafóricos), y (ESP) "desierto" (25 casos metafóricos y 487 no metafóricos).

Para representar formalmente la información metafórica, se ha definido un modelo de anotación en XML que especifica (i) el grado de metaforicidad (Hanks, 2006) de la expresión ("high", "middle", "low" o "literal"), (ii) el sentido literal y (iii) el significado metafórico. Éste último se representa mediante un conjunto difuso de palabras denominado "Metaphor Related Words", formado por aquellas palabras que, a juicio del anotador, representan el contenido metafórico de la palabra en ese contexto.

Con este modelo se han anotado, por ahora, 80 instancias de "torrent": 58 metafóricas (53 "high" y 5 "low") y 22 no metafóricas. En los meses restantes, se continuará anotando el corpus y se ampliará con otros tipos de metáforas.

En segundo lugar, se han realizado los primeros experimentos con algoritmos no supervisados para analizar si los usos metafóricos y no metafóricos se agrupan en *clusters* definidos o no. Para evaluar los algoritmos se han utilizado los ejemplos marcados de (ESP) "torrente" e (ING) "torrent". El corpus se representa mediante una matriz *palabra* × *contexto*, donde cada contexto representa un uso concreto de la palabra (metafórico o literal). El baseline se ha situado considerando todas las palabras como metafóricas (dado que en este corpus la mayoría de palabras son metafóricas), resultando un 0.62 de precisión para español y 0.73 para inglés. El mejor resultado, por ahora, se ha obtenido con el algoritmo K-means, que ha obtenido una precisión de 0.77 en español y 0.84 para inglés. Esta mejora, a falta de análisis en profundidad, nos indica que las hipótesis del proyecto son ajustadas y los objetivos alcanzables.

En los meses restantes, se analizarán diferentes formas de modelar el contexto, otros algoritmos y demás información que pueda ser de utilidad. Se analizará especialmente la discrepancia entre la distancia semántica a nivel léxico y la distancia semántica a nivel contextual (co-ocurrencia) como medida definitoria de usos metafóricos.

Bibliografía

Goatly, A. 1997. *The Language of Metaphors*. Routledge, Londres - Nueva York.

Hanks, P. 2006. Metaphoricity is gradable. En A. Stefanowitsch y S. Gries, editores, *Corpus-Based Approaches to Metaphor and Metonymy*. Mounton de Gruyter, Berlin, páginas 17–35.

Lakoff, G. y M. Johnson. 2004 (1980). *Metáforas de la vida cotidiana*. Cátedra, Madrid.

Shutova, E. 2010. Models of Metaphor in NLP. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala (Suecia).

Stefanowitsch, A. y S. Gries, editores. 2006. *Corpus-Based Approaches to Metaphor and Metonymy*. Mounton de Gruyter, Berlin.

Turney, P. D. y P. Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.