Modelos de semántica distribucional

M. Antònia Martí Antonín
CLiC-Centre de Llenguatge i Computació
Universitat de Barcelona
amarti@ub.edu

Resumen

Los modelos de semántica distribucional (MSD) construyen representaciones semánticas de manera dinámica en forma de espacios vectoriales multidimensionales a través del análisis estadístico de los contextos en que cada palabra aparece. Las palabras son los vectores y los contextos son las coordenadas de los mismos. El vector de cada palabra está constituido por los contextos en que ésta aparece y por el número de veces que ocurre en cada uno de ellos. Se trata de una aproximación cuantitativa del significado ya que representamos la información lingüística en términos de representaciones geométricas, los vectores. Gracias a esta representación cuantitativa del significado, en el espacio vectorial podemos comparar vectores (palabras) y obtener el grado de similitud que hay entre ellos de manera objetiva. Un aspecto importante de esta metodología, y donde los lingüistas pueden realizar una aportación significativa, es la determinación del tipo de contexto que se toma en consideración y si se aplica o no un procesamiento previo al corpus.

En el marco de la Lingüística Computacional esta aproximación al significado aparece como una alternativa a las limitaciones que presentan las aproximaciones tradicionales al significado de base simbólica. Frente a los modelos simbólicos, los modelos de semántica distribucional y su representación mediante modelos de espacios vectoriales, tienen unas propiedades que los hacen especialmente atractivos para el tratamiento computacional del significado. En primer lugar, el contenido semántico de una palabra se basa en su distribución y no en rasgos inherentes (rasgos semánticos o componentes del significado). Como resultado, las representaciones léxicas son cuantitativas y graduales, no simbólicas ni categoriales; se trata de representaciones relacionales, no referenciales; las relaciones semánticas entre las palabras representadas de este modo se pueden cuantificar y son graduales. Se trata de una aproximación radicalmente empírica. El modelo de aprendizaje del significado de las palabras es inductivo y es fácilmente escalable. Finalmente, el método es independiente de la lengua.

1. Introducción

A partir de los años 90 del siglo XX asistimos a un rápido y progresivo proceso de digitalización de la información debido a dos factores fundamentales: el uso de ordenadores personales y la aparición de internet como plataforma de comunicación y de intercambio y difusión de la información. Actualmente el soporte digital ha sustituido en su práctica totalidad el soporte analógico en todo tipo de actividad relacionada con comunicación y el intercambio de la información escrita. El hecho de disponer de grandes volúmenes de textos en formato digital ha incidido muy directamente en los métodos de investigación en Lingüística y en el desarrollo de técnicas y aplicaciones en el área de la Lingüística Computacional. Los métodos simbólicos basados en reglas, característicos de los años 70 y 80, han sido sustituidos por métodos empíricos basados en el aprendizaje automático a partir de muestras representativas de la lengua (y de los fenómenos que se quiere tratar) o por métodos híbridos que combinan ambos tipos de conocimiento.

Además, en el marco de la Lingüística Computacional, a las aplicaciones tradicionales como son la traducción automática y la extracción y recuperación de información, hay que añadir ahora nuevas aplicaciones como son la detección de la polaridad en documentos de opinión (Polanyi y Zaenen, 2006; Morante y Sporleder, 2012; Wiegand et al 2010), la detección de emociones y sentimientos (Turney, 2002; Pang et al. 2002; Wiebe et al. 2006), la detección de ironía y sarcasmo (Kovaz et al 2013; Whalen et al.

2014) y la distinción entre conocimiento factual y no factual (Vicze et al. 2008; Vicze, 2010), entre otras. El objetivo del procesamiento del lenguaje no se circunscribe ya al tratamiento de las lenguas en tanto que código formal, en un ámbito de aplicación delimitado y siguiendo un uso normativo de la lengua (estándar), sino que se tratan en el marco más amplio de la comunicación que incluye todo tipo de registros y una gran variedad de interacciones comunicativas: foros de discusión, redes sociales, conversaciones online, etc.

Ante la dificultad de procesar textos digitales resultantes del uso real de la lengua utilizando modelos simbólicos, siguiendo el modelo de las tecnologías del habla, a finales de los años 90, se abandonan las aproximaciones simbólicas basadas en reglas y se empiezan a desarrollar modelos basados en el aprendizaje automático a partir de corpus anotados, de base fundamentalmente estadística.

La anotación de corpus abarca actualmente todos los niveles del lenguaje (morfología sintaxis, semántica) e incluye también la anotación de aspectos emotivos (sentimientos), valorativos (polaridad) y referenciales (correferencia y anáfora), entre otros. Los textos que se anotan proceden de entornos de comunicación muy diversos, normalmente de aplicaciones de internet, y no siguen las normas de la lengua escrita. Es por ello, en parte, que los métodos basados en reglas se muestran totalmente ineficientes en el tratamiento de este tipo de textos.

Los lingüistas computacionales, que se habían centrado en la representación del conocimiento lingüístico y en el desarrollo de gramáticas y léxicos para el procesamiento del lenguaje, cambian su foco de atención y centran su actividad en la anotación de corpus. Como resultado, cobra relieve la Lingüística de Corpus cuyos objetivos son el desarrollo de métodos y técnicas para la recopilación, anotación y evaluación de los mismos. En la medida en que un corpus esté anotado de manera correcta y consistente, garantiza la calidad de la herramienta de análisis que de él se derive aplicando técnicas de aprendizaje automático.

Se ha comprobado que las herramientas de procesamiento del lenguaje desarrolladas partir de muestras de uso real de la lengua (estrategia *bottom-up*) se adecuan mejor al objetivo de análisis que los sistemas simbólicos basados en reglas (aproximación *top-down*).

2. Semántica distribucional y modelos de espacios vectoriales

La existencia de corpus de gran tamaño, de cientos de millones de palabras, ha permitido el desarrollo de nuevas técnicas de análisis que, sorprendentemente, tienen su fundamento lingüístico en propuestas aparecidas en los años 50 del pasado siglo.

Z.-H. Harris, en su artículo '*Distributional Analysis*' (Harris, 1954) sostiene que mediante la aplicación del método distribucional se pueden establecer las entidades básicas de la lengua: si dos unidades lingüísticas w1 y w2 tienden a tener las mismas propiedades distribucionales, podemos inferir que w1 y w2 pertenecen a la misma clase. En la segunda parte de su artículo aplica el análisis distribucional al significado. Considera que el significado de una palabra está determinado por el contexto en qué se usa y que importantes aspectos del significado de una palabra son una función o pueden estar representados por el conjunto de contextos en los que aparece. Las diferencias de significado están en correlación con las diferencias de distribución. Se trata de una

aproximación extensional y relacional al significado que sortea el problema de la representación formal del mismo. El significado de una palabra se puede representar mediante la suma de contextos en qué aparece.

En la misma época en que Harris propone el análisis distribucional, diversos autores desde ámbitos de conocimiento diferentes llegan a conclusiones muy parecidas. J. R. Firth (1957), profesor de lengua inglesa en la Universidad del Punjab y posteriormente profesor de Lingüística General en la Escuela de Estudios Orientales y Africanos de Londres, llega a la conclusión de que el significado de las palabras depende de sus contextos de uso: 'You shall know a word by the company it keeps'. Para Firth el lenguaje no se tenía que estudiar como un sistema mental aislado, sino como respuesta a determinadas situaciones comunicativas. La lingüística funcional y la lingüística cognitiva derivan, entre otras, de esta línea de pensamiento.

Wittgenstein, cuya obra inspiró dos de los principales movimientos filosóficos del siglo XX, el positivismo lógico y la filosofía del lenguaje ordinario, en su obra *Investigaciones Filosóficas* (1953) se cuestiona el significado referencial de las palabras y propone que el significado de una palabra depende de su uso en un determinado contexto.

La visión que estos tres autores tienen del lenguaje y, más en concreto, del significado, está en la base de los modelos de semántica distribucional, que actualmente constituyen uno de los focos de interés en Lingüística Computacional. Ahora bien, la aplicación del análisis distribucional para la representación del significado requiere disponer de información estadística sobre amplias muestras de uso y no se ha dispuesto de este tipo de información hasta hace muy poco. La tecnología digital ha hecho posible que estas propuestas teóricas aparecidas en los años 50 del pasado siglo, dispongan ahora de modelos matemáticos de base estadística para su representación y procesamiento.

Los modelos de semántica distribucional (MSD) construyen representaciones semánticas de manera dinámica en forma de espacios vectoriales multidimensionales a través del análisis estadístico de los contextos en que cada palabra aparece. Las palabras son los vectores y los contextos son las coordenadas de los mismos. El vector de cada palabra está constituido por los contextos en que ésta aparece y por el número de veces que ocurre en cada uno de ellos. Si para cada palabra del corpus obtenemos un vector, podemos agruparlos en una tabla o matriz que tendrá tantas filas como palabras tenga el corpus. El número de columnas variará según el tipo de contexto que se tome en consideración. Cada fila es un vector que tiene codificada numéricamente la coaparición de la palabra que da lugar a la fila con todos los contextos posibles, que son las columnas.

Se trata de una aproximación cuantitativa del significado ya que representamos la información lingüística en términos de representaciones geométricas, los vectores. Gracias a esta representación cuantitativa del significado, en el espacio vectorial podemos comparar vectores (palabras) y obtener el grado de similitud que hay entre ellos de manera objetiva. Para calcular el grado de similitud se aplican diferentes medidas, siendo el coseno la más utilizada. Así, dadas dos palabras representadas mediante vectores en un espacio vectorial, para cualquier ángulo existente entre dos vectores, el coseno dará un valor entre cero y 1, pero siempre inferior a 1; en este caso las palabras no son semánticamente idénticas. Si el ángulo entre los dos vectores es cero, es decir, si ambos vectores apuntan al mismo lugar, el coseno dará como valor 1. Gracias a esta tipo de representación, el significado de las palabras se puede expresar numéricamente con

valores entre el 0 y el 1: se trata, por tanto, de una representación no categorial, ya que la similitud se expresa de modo gradual.

Un aspecto importante de esta metodología, y donde los lingüistas pueden realizar una aportación significativa, es la determinación del tipo de contexto que se toma en consideración y si se aplica o no un procesamiento previo al corpus. Turney y Pantel (2010) proponen tres tipos básicos de contexto: palabra-documento, palabra-palabra y palabra-patrón.

Las matrices palabra-documento se han utilizado principalmente en recuperación de información para detectar qué documentos son relevantes respecto de una pregunta del usuario (Salton et al. 1975). Las palabras de la pregunta se representan en las filas de la matriz y cada documento se representa en su correspondiente columna. Los documentos que contienen más palabras de la pregunta se proponen como resultado de la búsqueda. Esta aproximación se utiliza también en tareas como clasificación de documentos por su temática o en la detección de temas dentro de un mismo documento.

Las matrices palabra-contexto son las más adecuadas para medir la similitud entre palabras. Como ya se ha indicado, la medida de similitud más utilizada es el coseno del ángulo entre los diferentes vectores en una matriz palabra-contexto. Existe una gran variedad de aplicaciones basadas en la similitud entre palabras como son la creación de clústers de palabras semejantes, la creación de tesauros y la desambiguación semántica, entre otras.

Las matrices palabra-patrón son adecuadas para medir la similitud semántica de pares de palabras respecto de un determinado patrón. Los vectores son pares de palabras (carpintero: madera, pintor: pintura; río: agua, personas: avenida) y las columnas, los contextos o patrón (‘usar’, ‘trabajar con’; ‘discurrir’). Así obtenemos que ‘carpintero: madera’ y ‘pintor: pintura’ mantienen una relación de analogía ya que comparten el patrón ‘usar’ y que ‘río: agua’ y ‘personas: avenida’ son análogas en la medida que comparten la relación ‘discurrir’ Este tipo de matrices se utilizan para tareas como la similitud de relaciones o la similitud de patrones (similitud entre los elementos de las columnas): los patrones ‘usar’ y ‘trabajar con’ son similares en la medida que diferentes pares de palabras los comparten.

Las diferentes formulaciones de estas matrices coinciden en el hecho de que el contexto de una palabra aporta información sobre su significado, en que el grado de similitud semántica entre dos palabras depende de si comparten o no un número relevante de contextos y, finalmente, en que se pueden captar aspectos fundamentales del significado a partir de la simple coocurrencia estadística.

En el marco de la Lingüística Computacional esta aproximación al significado aparece como una alternativa a las limitaciones que presentan las aproximaciones tradicionales al significado de base simbólica. En concreto, son una alternativa a la falta de criterios objetivos para determinar la mejor opción de entre las diferentes propuestas teóricas existentes sobre el metalenguaje de representación del significado léxico y oracional: Framenet (Fillmore et al. 2012), WordNet (Miller, 1991b), VerbNet (Feeley et al, 2012), y PropBank (Palmer, 2005) entre otros. También son una alternativa a la falta de consistencia interna entre los diferentes recursos existentes (p.e.: la vinculación de recursos como WordNet, VerbNet y PropBank); y, finalmente, son una alternativa a la falta de cobertura de los léxicos con información semántica. Cuando hay que procesar

grandes cantidades de textos, de temática muy diversa y resultado de situaciones comunicativas espontáneas, estos recursos se muestran claramente insuficientes en su cobertura.

Frente a los modelos simbólicos, los modelos de semántica distribucional y su representación mediante modelos de espacios vectoriales, tienen unas propiedades que los hacen especialmente atractivos para el tratamiento computacional del significado:

a. El contenido semántico de una palabra se basa en su distribución y no en rasgos inherentes (rasgos semánticos o componentes del significado). Como resultado, las representaciones léxicas son cuantitativas y graduales, no simbólicas ni categoriales.
b. Se trata de representaciones relacionales, no referenciales, en las que el significado de las palabras se expresa en términos de vectores obtenidos a partir de los contextos de aparición de las palabras. Este modo de representar el significado es fácilmente tratable computacionalmente.
c. Las relaciones semánticas entre las palabras representadas de este modo se pueden cuantificar y son graduales. Se puede medir y cuantificar el grado de similitud entre ellas de manera objetiva.
d. Se pueden construir tensores con más de una relación (Baroni y Lenci 2010) que permiten representar adecuadamente el contenido semántico de un corpus. Se derivan modelos de lenguaje más compactos sobre los que se puede operar con mucha más eficiencia.
e. Se trata de una aproximación radicalmente empírica. El modelo de aprendizaje del significado de las palabras es inductivo.
f. El modelo es fácilmente escalable, no plantea problemas de cobertura.
g. El método es independiente de la lengua.

Como cabe esperar, la aproximación cuantitativa al significado presenta también limitaciones importantes, entre las que cabe destacar la dificultad para tratar el significado más allá de unidades mínimas como son las palabras, y la representación de la polisemia. ¿Cómo llevar a cabo la composición de vectores? ¿Cómo representar el significado de unidades lingüísticas complejas, más allá de las palabras? ¿Cómo detectar para una palabra polisémica qué contextos son propios de un sentido u otro?

Estas cuestiones centran el interés de los investigadores en semántica distribucional. Propuestas metodológicas como el 'Deep Learning' abordan el problema de la representación del significado de las oraciones y de la polisemia léxica. Esta metodología se basa en sucesivas capas de redes neuronales, donde cada red aprende nuevo conocimiento de las anteriores y los transmite a la siguiente de modo secuencial. El problema de esta aproximación al significado es que se trata de un modelo de 'caja negra', en el que el investigador pierde el control de lo realmente hace el sistema. Solo puede comprobar si los resultados, en evaluaciones intrínsecas y extrínsecas, son buenos.

Cabe preguntarse también el impacto que estos modelos de base estadística pueden tener en nuestra concepción del lenguaje humano. El cerebro es un buen almacén de memoria, y todo apunta a que nuestro aprendizaje se basa en gran medida en almacenar información en forma de patrones que organizamos a diferentes niveles de abstracción. Interpretamos los nuevos inputs a partir de los patrones que hemos aprendido y almacenado en memoria. En este modo de enfocar el aprendizaje/adquisición del lenguaje coinciden tanto investigadores en neurociencia (Hawkins y Blakeslee, 2004) como en lingüística cognitiva (Bybee 2014) y en psicología del comportamiento (Tomasello, 2000). La

investigación actual en Lingüística Computacional encaja también con estos enfoques sobre el lenguaje. Algunos autores, como Landauer y Dumais (1997) consideran que estos modelos tienen un fundamento cognitivo.

Referencias bibliográficas

Baroni, M. (2013). ‘‘Composition in distributional semantics’’. *Language and Linguistics Compass*, 7(10):511-522, October.

Baroni, M. y Alessandro L. (2010). ‘‘Distributional memory: A general framework for corpus-based semantics’’. *Computational Linguistics*, 36(4):673--721, December.

Feeley, W., Claire Bonial and Martha Palmer, Evaluating the coverage of VerbNet, In the Proceedings of ISA-8: Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation, Pisa, Italy, October, 2012.

Fillmore, Charles-J, Russell Lee-Goldman, y Russell Rhodes (2012). ‘‘The framenet constructicon’’. *A Sign-based Construction Grammar*. CSLI, Stanford, CA.

Firth, J.R. (1957). *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Harris, Z. (1954). ‘‘Distributional structure’’. *Word*, 10(23):146--162.

Hawkins, St. y S. Blakeslee (2004). *On intelligence*. H. Holt and Co. Publishers.

Kovaz, D., R.J. Kreuz, y M.A. Riordan (2013). ‘‘Distinguishing sarcasm from literal language: evidence from books and blogging’’. *Discourse Processing*. Num. 50. Pp. 598-615.

Landauer, T. K y S. Dumais (1997). ‘‘A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge’’. En *Psychological Review*. Vol. 104, núm. 2, pp. 211. American Psychological Association Publish.

Miller, G. (1991). ‘‘Contextual correlates of semantic similarity’’. *Language and Cognitive Processes*, 6(11), 1-28.

Mitchell, J. y M. Lapata (2010). ‘‘Composition in distributional models of semantics’’. *Cognitive Science*, 34(8):1388--1439.

Morante, R. y C. Sporleder. (2012). ‘‘Modality and negation: An introduction to the special issue’’. *Computational linguistics*, 38(2), 223-260.

Palmer, M., P. Kingsbury y D. Gildea (2005). ‘‘The Proposition Bank: An Annotated Corpus of Semantic Roles’’. *Computational Linguistics*, 21 (1).

Pang, B., L. Lee, y S. Vaithyanathan, S. (2002). ‘‘Thumbs up?: sentiment classification using machine learning techniques’’. Proceedings of the *ACL-02 conference on Empirical methods in natural language processing*-Volume 10, páginas 79-86. Association for Computational Linguistics.

Polanyi L., Zaenen, A. (2006). ‘‘Contextual Valence Shifters. Computing affect and attitude in text: Theory and applications’’, 20, páginas 1-10. *The Information Retrieval Series*.

Salton, G., A. Wong y C.S. Yang (1975). ‘‘A Vector Space Model for Automatic Indexing’’. En, *Information Retrieval and Language Processing*. Vol. 18, num. 11.

Taboada, M., C. Anthony y K. Voll. (2006). ‘‘Methods for creating semantic orientation dictionaries’’. Proceedings of the *5th Conference on Language Resources and Evaluation* (LREC’06), páginas 427-432.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). ‘‘Lexicon-based methods for sentiment analysis’’. *Computational linguistics*, 37(2), 267-307.

Tomasello, Michael (2000). ‘‘First steps toward a usage-based theory of language acquisition’’. En *Cognitive Linguistics*, 11(1-2):61--82.

Turney (2002). ‘‘Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification reviews’’. En *Proceedings of the 40th Annual Meeting of the ACL*. Pp. 417-424.

Turney, Peter D. (2008). ‘‘The latent relation mapping engine: Algorithm and experiments. A Artificial Intelligence Res’’, (JAIR)}, 33:615-655.

Turney, Peter D. y Pantel, P. (2010). ‘‘From frequency to meaning: Vector space models of semantics. A Artificial Intelligence Res’’, 37(1):141--188, January.

Vincze, V., Szarvas G., Farkas R., Móra G. y Csirik J. (2008). ‘‘The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes’’. BMC Bioinformatics, 9:1-9.

Vincze, V. (2010). ‘‘Speculation and negation annotation in natural language texts: what the case of bioscope might (not) reveal’’. *Proceedings of the workshop on negation and speculation in natural language processing*, páginas 51-59, Uppsala, Association for Computational Linguistics.

Whalen, J., P.M. Pexman, J.G. Alistair y S. Nowson (2014). ‘‘Verbal irony use in personal blogs’’. *Behaviour and Information Technology*. Vol. 32, num 6, pp.560-569.Taylor and Francis.

Wiebe, J. , T. Wilson y C. Cardie (2006). ‘‘Annoating expressions of opinions and emotions in language’’. En *Language Resources and Evaluation* num. 39, pp. 165-2010.

Wiegand, M., A. Balahur, B. Roth, D. Klakow, y A. Montoyo (2010). ‘‘A survey on the role of negation in sentiment analysis’’. In *Proceedings of the workshop on negation and speculation in natural language processing*, páginas 60-68. Association for Computational Linguistics.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). ‘‘Recognizing contextual polarity in phrase-level sentiment analysis’’. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347-354). Association for Computational Linguistics.

Wittgentein, L. (1958). *Philosophical investigations*. Blackwell. Trad. Catalá de J.M. Terricabras, *Investigacions Filosòfiques*, E. Laia 1983.