ONOMÁZEIN
Revista de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS

Revisión de métodos para la desambiguación léxica automática: aprendizaje automático y medidas de relación y similitud semánticas

A review of methods for word sense disambiguation: machine learning and measures of relatedness and semantic similarity

Fredy Núñez Torres
Pontificia Universidad Católica de Chile
Chile

María Beatriz Pérez Cabello de Alba
Universidad Nacional de Educación a Distancia
España

ONOMÁZEIN 64 (junio de 2024): 249-267
DOI: 10.7764/onomazein.64.14
ISSN: 0718-5758

Fredy Núñez Torres: Departamento de Ciencias del Lenguaje, Pontificia Universidad Católica de Chile, Chile.
| E-mail: frnunez@uc.cl
María Beatriz Pérez Cabello de Alba: Universidad Nacional de Educación a Distancia, España.
| E-mail: bperez-cabello@flog.uned.es

Fecha de recepción: julio de 2022
Fecha de aceptación: mayo de 2024

Resumen

Dentro de las posibles soluciones para la desambiguación léxica automática en tareas del procesamiento del lenguaje natural encontramos los métodos basados en algoritmos de aprendizaje automático, en medidas de relación semántica y en medidas de similitud semántica. Mientras los métodos de aprendizaje automático utilizan fuentes endógenas de conocimiento, las medidas de relación y similitud semánticas recurren a fuentes exógenas de conocimiento, como son las glosas de las definiciones de un recurso lexicográfico, o una ontología o tesauro, que provee relaciones léxicas de significado en el marco de una jerarquía conceptual. En este trabajo presentamos y analizamos los distintos tipos de métodos para la desambiguación léxica automática divididos en cuatro grupos: basados en algoritmos de aprendizaje automático, basados en medidas de relación semántica, basados en medidas de similitud semánticas y basados en medidas híbridas. Se propone que la ventaja de los métodos basados en medidas de relación y similitud radica en el hecho de que sus resultados no se derivan únicamente de la eficiencia estadística, sino que se tiene en cuenta el conocimiento lingüístico dentro de los parámetros que conforman cada medida utilizada.

Palabras clave: lingüística computacional; procesamiento del lenguaje natural; aprendizaje automático; desambiguación léxica automática; relación semántica; similitud semántica.

Abstract

Among the possible solutions for automatic lexical disambiguation in natural language processing tasks, we find methods based on machine learning algorithms, semantic relatedness, and semantic similarity measures. While machine learning methods use endogenous sources of knowledge, semantic relatedness and similarity measures resort to exogenous sources of knowledge, such as definitions from lexicographic resources or lexical meaning relations from ontologies or thesauri, which offer a conceptual hierarchy. In this work, we present and analyze the different types of methods for automatic lexical disambiguation divided into four groups: based on machine learning algorithms, based on semantic relatedness measures, based on semantic similarity measures, and based on hybrid measures. We postulate that the advantage of methods based on relationship and similarity measures lies in the fact that their results are derived from statistical efficiency and linguistic knowledge found in the parameters that make up each of the measures used.

Keywords: computational linguistics; natural language processing; machine learning; word sense disambiguation; semantic relatedness; semantic similarity.

1. El problema lingüístico de la ambigüedad léxica

La *ambigüedad léxica* se define como el fenómeno en el que una entrada léxica puede contener dos o más significados diferentes. En particular, esta investigación se centra en la resolución de la ambigüedad léxica para los casos de homonimia y polisemia, y el estudio de este fenómeno desde una perspectiva informática. Como ejemplo preliminar, la ambigüedad léxica ocurre debido a la existencia de dos significados denotativos para una misma etiqueta lingüística, como en "La gata está sobre el cobertizo", donde "gata" puede hacer referencia tanto al felino doméstico como a la herramienta levantapesos en el español de Chile.

En cuanto a la perspectiva tradicional para la descripción de la ambigüedad léxica, Ide y Véronis (1998) la caracterizan como la asociación entre un ítem léxico determinado en un texto o discurso y un sentido que pueda ser distinguible de otros significados potencialmente atribuibles a ese ítem léxico. Según lo anterior, algunos de los autores clásicos en el ámbito de la semántica y la lingüística teórica, como Lyons (1977) y Cruse (1986), distinguen dos tipos generales de ambigüedad léxica en tanto se le reconoce como un fenómeno no uniforme: homonimia y polisemia. Estas distinciones han sido actualizadas principalmente por Pustejovsky (1991, 1995), Pustejovsky y Boguraev (1996) y Bouillon y Busa (2001). En primer lugar, la homonimia corresponde a un fenómeno en el que un ítem léxico contiene accidentalmente dos o más significados. También se le ha llamado *ambigüedad contrastiva*. Un ejemplo para este caso es la unidad "velas" en las siguientes proposiciones:

(1)
a. El contramaestre mandó izar las *velas*.
b. Los monjes lograron encender las *velas*.

En el ejemplo (1a) "velas" se refiere a los 'instrumentos que capturan la fuerza del viento en una embarcación', mientras que en (1b) se refiere a los 'artefactos hechos de cera con una mecha que sirven para proveer de iluminación'. En segundo lugar, la polisemia[1] se refiere a los casos en los que un mismo lexema puede tener múltiples significados relacionados. Un ejemplo de lo anterior es la relación entre los distintos usos del verbo "abrir", como en las siguientes construcciones:

(2)
a. El niño *abrió* el debate.
b. El niño *abrió* la lata.

1 Llamada también polisemia complementaria en los trabajos de Weinreich (1964) y en las primeras versiones de los trabajos de Pustejovsky (1991, 1995).

En el primer caso (2a), la apertura supone el inicio de una acción comunicativa, mientras que en el segundo (2b) implica la apertura de un objeto físico. Si bien en ambos casos "el niño" funciona como movilizador de la acción, o agente, los sentidos de "abrir" para las construcciones sintácticas propuestas difieren.

Una vez descrito el problema de la ambigüedad léxica en este apartado, en (2) presentamos los métodos para resolver el problema de la desambiguación léxica automática (*word sense disambiguation*, en adelante WSD por sus siglas en inglés) basados en algoritmos de aprendizaje automático. Luego, se exponen dos grupos de métodos basados en medidas para WSD (Slimani, 2013; Ali, Alfayez y Alqhayz, 2018) en tareas del procesamiento del lenguaje natural (*natural language processing*, en adelante NLP por sus siglas en inglés), a saber, los métodos de relación semántica, que están descritos y analizados en (3), y los métodos de similitud semántica, analizados en (4). Finalmente, en (5) exponemos las conclusiones de este estudio donde resaltamos las ventajas de este tipo de métodos basados en medidas de relación y similitud semánticas para tareas de WSD.

2. Métodos basados en algoritmos de aprendizaje automático

Los métodos basados en algoritmos de aprendizaje automático para WSD utilizan fuentes exógenas (corpus previamente anotados o etiquetados) o endógenas (el mismo contexto oracional) para derivar reglas o modelos que realicen el proceso de desambiguación léxica automática. En términos generales, este tipo de métodos selecciona un conjunto de muestras en lenguaje natural a partir de un corpus, considerando las distintas clasificaciones de cada elemento. Luego, deben ser capaces de identificar regularidades asociadas a cada elemento, para así generalizar patrones de reglas que serán aplicadas para generalizar los elementos nuevos. A continuación, se exponen brevemente dos tipos: supervisado y no supervisado. Ambos están basados en el principio del aprendizaje automático, esto es, que la máquina pueda aprender automáticamente a partir de la observación de instancias o datos textuales, con el objetivo de predecir un determinado comportamiento de estos.

2.1. Aprendizaje automático supervisado

El aprendizaje supervisado se define como una técnica de aprendizaje automático en la que se dispone de un corpus de entrenamiento previamente etiquetado con el sentido correspondiente para cada instancia de una palabra objetivo (Núñez, 2019). Según las pruebas de Carpuat y Wu (2005), los métodos de aprendizaje automático supervisado más utilizados para tareas de WSD serían: bayesiano ingenuo, modelo de máxima entropía, *Boosting model* y *Kernel PCA-based model*. Estos métodos, a su vez, están basados en conocimiento contextual; es decir, utilizan el mismo corpus en análisis, típicamente controlado a partir del contexto oracional o ventana contextual, para derivar información estadística. Según

Núñez y Pérez Cabello de Alba (2022), la mayoría de los métodos supervisados tradicionales tienen, al menos, cuatro fases de aplicación en común:

a. Selección de un conjunto de datos textuales que muestre las diferentes clasificaciones para cada elemento (valores, atributos, características).
b. Identificación de los patrones asociados con cada elemento.
c. Generalización de patrones.
d. Aplicación de patrones para clasificar nuevos elementos no presentes en el conjunto de datos textuales inicial.

En cuanto a su aplicación, los diferentes métodos basados en aprendizaje automático supervisado, de acuerdo con Aung y otros (2011), Fulmari y Chaldak (2014), Gamallo y otros (2014) y Gosal (2015), pueden presentar el problema del defecto de conocimiento en comparación con los métodos basados en conocimiento contextual, pues el cotexto no necesariamente constituiría un conjunto de datos textuales suficiente para lograr que un sistema automático derive clasificaciones eficientes. A pesar de que los métodos basados en medidas de relación y similitud podrían presentar el problema inverso de exceso de conocimiento, esta sería una ventaja relevante por sobre los métodos de aprendizaje automático, como se expondrá más adelante.

2.2. Aprendizaje automático no supervisado

A diferencia de su contraparte, el aprendizaje automático no supervisado no depende de un recurso lingüístico informatizado que haya sido previamente anotado con las estructuras o rasgos que el algoritmo de clasificación pretende producir como valores de salida. En este caso, el algoritmo es provisto solamente con los datos que provienen del conocimiento contextual; es decir, de las instancias en análisis. A partir de esa información, debe analizar estructuras lingüísticas mediante la identificación de patrones textuales de distribución y de propiedades de agrupación de los rasgos que emergen desde los datos (Popescu y Hristea, 2010; Ustalov y otros, 2018).

El surgimiento de estos métodos de aprendizaje automático no supervisado, principalmente a partir de la década de 1990, se debe principalmente a que los corpus etiquetados manualmente, o previamente entrenados por un humano, agregan una dificultad significativa al procedimiento. Por lo tanto, este costo debe compararse con la precisión que proporcionaría la utilización de métodos supervisados. En la medida en que los algoritmos no supervisados no incurran en estos costos, ofrecen una ventaja importante solo si son capaces de mantener un nivel aceptable de rendimiento en las aplicaciones para las que están diseñados.

Según lo anterior, el cuello de botella de la adquisición del conocimiento en desambiguación léxica automática se ha mantenido como un problema relevante en el ámbito de la

representación del conocimiento y el desarrollo de métodos de aprendizaje automático. En efecto, la principal dificultad al comparar métodos de aprendizaje supervisado y no supervisado es que los algoritmos supervisados a menudo necesitan de una gran cantidad de instancias previamente etiquetadas con sentidos relevantes para llevar a cabo el proceso de desambiguación. Específicamente, en los métodos de aprendizaje automático no supervisados, el hecho de que no se utilice información etiquetada manualmente de manera previa también plantea una serie de desafíos. Estos se dan sobre todo en torno a la evaluación de sus resultados al implementar algoritmos basados en agrupaciones (*clustering*). Según Jurafsky y Martin (2009), las siguientes corresponden a las desventajas más relevantes para los métodos no supervisados:

a. Es altamente probable que no se conozcan los sentidos correctos de las instancias utilizadas en los enfoques supervisados.
b. Los grupos que derivan de la clasificación tienden a ser heterogéneos con respecto a los sentidos de las instancias contenidas en el corpus.
c. El número de grupos resultante es, en la mayoría de los casos, diferente del número de sentidos de las palabras objetivo que se desambiguan.

3. Métodos basados en medidas de relación semántica

El primer grupo para tareas de WSD es el basado en medidas de relación semántica, *word relatedness*, correspondiente a de tipo Lesk (Lesk, 1986, 1987) para el solapamiento de glosas de definiciones, y de tipo Lesk adaptado (Banerjee y Pedersen, 2002, 2003), que integra el contexto sintáctico en el proceso de solapamiento de información. Las palabras relacionadas son aquellas que comparten ciertos aspectos de su significado y que frecuentemente coocurren en el mismo contexto oracional. Por ejemplo, las unidades léxicas “médico” y “medicamento” son palabras relacionadas porque comparten algunos aspectos de su significado basados en un campo semántico, como el de la ‘salud’ en este caso.

3.1. Tipo Lesk (1986, 1987)

Este método, desarrollado inicialmente por Michael Lesk (1986), corresponde, en términos generales, a una medida basada en la información contenida en un diccionario legible por la máquina, a partir de la cual se calcula el solapamiento de palabras entre dos palabras objetivo[2] y las glosas de sus respectivas definiciones. El mayor mérito de este procedimien-

2 Se utilizará de aquí en adelante la expresión *palabra objetivo* para referirnos a aquella unidad léxica que se establece como referencia para la definición de la ventana de palabras durante el análisis de un texto de entrada.

to es su capacidad para encontrar la combinación de los sentidos de palabra que maximice la relación total entre los sentidos de las palabras objetivo, de modo que estas cumplan con el criterio de la relación semántica.

En términos lingüísticos, uno de los fundamentos para este método es la idea de que el significado de un concepto se expresa mediante una agrupación de palabras y, por tanto, la manera más eficiente para cuantificar esta relación semántica entre las unidades léxicas que componen una definición sería el solapamiento de palabras. Según lo anterior, se hace imprescindible la utilización de definiciones de diccionario, como una fuente de conocimiento externa que permitirá establecer los parámetros de coocurrencia para el solapamiento de glosas. En cuanto a su formalización, y según la explicación de Navigli (2009), dadas dos palabras objetivo (W_1, W_2), se calculará un puntaje para cada par de sentidos de palabra (S_1, S_2), donde $S_1 \in sentidos(W_1)$ y $S_2 \in sentidos(W_2)$. De esta forma, se establece la intersección entre las glosas de las definiciones de cada palabra objetivo:

$$puntaje_{LESK}(S_1, S_2) = glosa(S_1) \cap glosa(S_2)$$

De esta forma, la glosa de cada sentido de palabra consistiría en una bolsa de palabras correspondiente a la definición textual del sentido (S_i) para cada palabra (W_i).

El ejemplo más representativo para la explicación del procedimiento de solapamiento de glosas es el cálculo de la relación semántica entre los sentidos de las palabras "pine" y "cone", propuesto por Lesk (1986). Para esto, se utilizaron las definiciones provenientes del diccionario *Oxford Advanced Learner* (Hornby y otros, 1974), en el que se incluyen cuatro sentidos para "pine" y tres sentidos para "cone". Lo anterior se expone en las tablas 1 y 2:

TABLA 1

Sentidos y definiciones para "pine"

PINE	
SENTIDO	**GLOSA**
S_1 = *tree*	*glosa*(S_1) = seven kinds of evergreen tree with needle-shaped leaves
S_2 = *pine*	*glosa*(S_2) = pine
S_3 = *waste*	*glosa*(S_3) = waste away through sorrow or illness
S_4 = *something*	*glosa*(S_4) = pine for something, pine to do something

TABLA 2

Sentidos y definiciones para "cone"

CONE	
SENTIDO	**GLOSA**
S_1 = *body*	*glosa*(S_1) = solid body which narrows to a point
S_2 = *shape*	*glosa*(S_2) = something of this shape, whether solid or hollow
S_3 = *fruit*	*glosa*(S_3) = fruit of certain evergreen trees (fir, pine)

Posteriormente, se evaluó el solapamiento de cada una de las glosas correspondientes al conjunto de pares para los sentidos de "pine" y "cone" registrados en la fuente de conocimiento, para así establecer un puntaje que expresa el número de coocurrencias. Lo anterior se expresa de la siguiente forma:

TABLA 3

Puntajes para el solapamiento entre los conceptos "pine" y "cone"

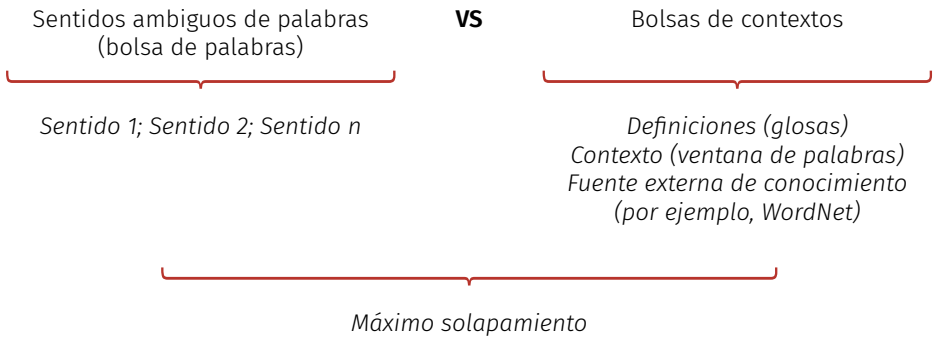
SOLAPAMIENTO	PUNTAJE
$cone_{glosa(S_1)} \cap pine_{glosa(S_1)}$	= 0
$cone_{glosa(S_2)} \cap pine_{glosa(S_1)}$	= 0
$cone_{glosa(S_3)} \cap pine_{glosa(S_1)}$	= 0
$cone_{glosa(S_4)} \cap pine_{glosa(S_1)}$	= 0
$cone_{glosa(S_1)} \cap pine_{glosa(S_2)}$	= 0
$cone_{glosa(S_2)} \cap pine_{glosa(S_2)}$	= 0
$cone_{glosa(S_3)} \cap pine_{glosa(S_2)}$	= 1
$cone_{glosa(S_4)} \cap pine_{glosa(S_2)}$	= 0
$cone_{glosa(S_1)} \cap pine_{glosa(S_3)}$	**= 2**
$cone_{glosa(S_2)} \cap pine_{glosa(S_3)}$	= 1
$cone_{glosa(S_3)} \cap pine_{glosa(S_3)}$	= 0
$cone_{glosa(S_4)} \cap pine_{glosa(S_3)}$	= 1

Como resultado, el máximo puntaje de solapamiento entre todas las posibles combinaciones de sentidos se obtiene para la intersección entre la glosa del sentido uno de "pine"

('seven kinds of evergreen tree with needle-shaped leaves') y la glosa del sentido tres de "cone" ('fruit of certain evergreen trees'), en el que se evidencian dos coocurrencias. En este caso, se solapan las palabras "evergreen" y "tree". Por lo tanto, la medida de Lesk seleccionará estos sentidos y los asignará en el proceso de desambiguación cada vez que las palabras "pine" y "cone" coincidan en la misma ventana contextual. Este tipo de procedimiento podría generalizarse en la siguiente figura:

FIGURA 1

Procedimiento general de métodos basados en solapamiento



Diversos estudios acerca del desarrollo y progresión de los métodos para la desambiguación automática de sentidos de palabra publicados a partir de la década de 1990 (Ide y Véronis, 1998; Navigli, 2009; Vihdu y Abirami, 2014) indican que el método Lesk es uno de los primeros algoritmos de WSD considerado como exitoso, y que logró de manera eficiente la consolidación del criterio de incorporar una fuente de conocimiento externa para abordar este tipo de procesamiento. Esto es particularmente relevante desde la perspectiva de la inclusión de un incipiente razonamiento lingüístico dentro de una propuesta íntegramente estocástica. A pesar de la aprobación de la que goza el método Lesk original, debido a la simpleza de su algoritmo y la efectividad que presenta en varios de los casos experimentales más canónicos, las críticas para este método son bastante consistentes en la mayor parte de la literatura especializada (Wilks y otros, 1989; Cowie y otros, 1992).

Se han establecido dos grandes deficiencias para este método. En primer lugar, se ve afectado directamente por el tamaño de la fuente de conocimiento (*i. e.*, el diccionario legible por la máquina) y por la exactitud de las palabras utilizadas en las glosas de las definiciones. Este problema implica que, dado que las definiciones son típicamente breves, la bolsa de palabras de cada una no permite establecer una relación entre los sentidos en análisis, porque la frecuencia de los casos de solapamiento es muy baja. En segundo lugar, se trata de un método que, en la medida que intenta encontrar la relación más significativa de sentidos de palabra, se ve sometido al problema de la explosión combinatoria. Esta limitación implica que, mientras más sentidos en análisis sean seleccionados, y junto con

ellos aumente el volumen de cada bolsa de palabras para sus correspondientes palabras objetivo, entonces se reduce la probabilidad de encontrar una relación óptima entre una palabra y su correspondiente sentido.

3.2. Tipo Lesk adaptado (Banerjee y Pedersen, 2002)

Esta propuesta de adaptación de la medida de Lesk (1986) surge como una solución para el problema de la limitación de las glosas de las definiciones de diccionario. Dado que las definiciones típicamente tienden a ser breves, la identificación de sentidos relacionados según la medida de solapamiento original se ve imposibilitada porque la información contenida en la fuente de conocimiento es insuficiente. Asimismo, la medida original es redundante, altamente iterativa e ineficiente dado el problema de la explosión combinatoria.

En los trabajos de Banerjee y Pedersen (2002, 2003) se propone una adaptación a la medida original de Lesk, mediante la incorporación del contexto sintáctico de la palabra objetivo. La crítica de estos autores a la medida original, además de las debilidades antes expuestas, es que el proceso de activación (*spreading activation*) del solapamiento es limitado, dado que no es un mecanismo suficiente para dar cuenta de relaciones más indirectas entre las palabras en análisis. Así, se proponen dos soluciones preliminares: (1) expandir las definiciones de diccionario para aumentar las posibilidades de encontrar coocurrencias y (2) considerar ventanas contextuales específicas, al mismo tiempo que se incorporan las glosas de los sentidos para aquellos conceptos que se encuentran relacionados léxica o semánticamente con la taxonomía de WordNet (Miller, 1985; Miller y otros, 1993; Fellbaum, 1998). Esta adaptación de la medida original considera un cotexto, correspondiente a *n tokens* de WordNet a la izquierda y a la derecha de la palabra objetivo; es decir, una ventana contextual de $2n + 1 = N$, donde $2n$ corresponde al número de palabras adyacentes. Si la palabra objetivo se encuentra al inicio o al final de la instancia a considerar en la ventana contextual, se adhieren palabras en la dirección opuesta que corresponda. Acerca del criterio para la definición de la ventana contextual, Choueka y Lusignan (1985) afirman que, desde una perspectiva cognitivista, el ser humano realiza decisiones de desambiguación basadas en intervalos de información estrechos que rodean a la palabra objetivo, usualmente una o dos palabras en cada dirección. Según esto, se establece un puntaje de combinación producto de la evaluación de todas las posibles combinatorias para la asignación de sentidos en la ventana contextual, mediante la comparación de glosas para los pares de palabras. Este puntaje corresponde a la siguiente ecuación:

$$\prod_{i=1}^{N} |W_i|$$

donde la multiplicatoria de todos los valores de W_i implica que i varía desde 1 hasta el valor total de la ventana contextual. Así, W_i corresponderá al número de sentidos posibles para

cada palabra. Esta técnica permite que el sentido desambiguado corresponda a aquella combinación que se activa con mayor fuerza, en relación con las palabras adyacentes en la ventana contextual.

4. Métodos basados en medidas de similitud semántica

A continuación, se exponen los métodos basados en medidas de similitud semántica, *word similarity*. Estas, a su vez, se dividen en dos subgrupos: de distancia entre rutas (Wu y Palmer, 1994; Leacock y Chodorow, 1998) y de contenido de información (Resnik, 1995; Jiang y Conrath, 1997; Lin, 1998).

En términos generales, las palabras similares establecen relaciones léxicas del tipo IS-A (hiponimia/hiperonimia) o HAS-PART (meronimia) en una jerarquía conceptual. Por ejemplo, las unidades léxicas "auto" y "bicicleta" son palabras similares porque, entre sus características, ambos significados son hipónimos de la categoría 'medios de transporte'. Así, los criterios para determinar la similitud están dados por las características o atributos (*features*) que presenta una determinada instancia o *wordform*. Se entenderán como características la o las palabras de contenido, cualquiera sea su clase (dejando fuera entonces a las palabras funcionales, o *stopwords*), que sean adyacentes a las unidades léxicas referidas en la definición de la palabra objetivo en un diccionario legible por la máquina, que a su vez se utilice como fuente de conocimiento lingüístico.

4.1. Medidas de distancia entre rutas

Las medidas de distancia entre rutas utilizan como método de desambiguación la clasificación direccional de relaciones taxonómicas. Utilizan como fuente de conocimiento la taxonomía de WordNet, que contiene relaciones taxonómicas del tipo *IS-A* (vertical) y *HAS-PART* (horizontal). Desde ahí es posible establecer la distancia de las rutas taxonómicas mediante etiquetas numéricas. La pregunta central de este tipo de medidas es qué tan cercanas son estas unidades léxicas en cuanto a su posición en una jerarquía conceptual. Así, dos conceptos serán similares, o tendrán sentidos similares, cuanto más cerca se encuentre el uno del otro en una jerarquía taxonómica. A continuación, se presentan tres medidas basadas en la distancia entre rutas.

4.1.1. Wu y Palmer (1994)

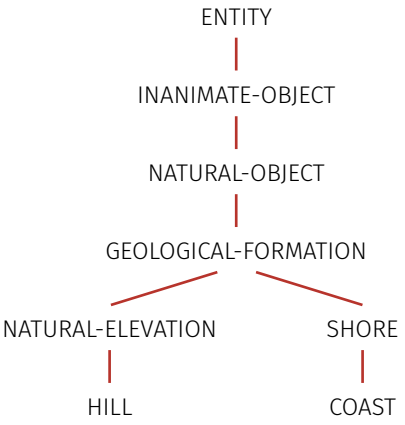
La propuesta de Wu y Palmer (1994) consiste en una medida de similitud semántica basada en las variables de distancias y profundidades en una taxonomía. De esta forma, se considera la distancia de la ruta entre dos conjuntos de sentidos, S1 y S2, y su hiperónimo común más cercano, S_3 (*lowest common subsumer*, de aquí en adelante LCS), así como la distancia entre el LCS y la raíz de la taxonomía en la que se encuentran los conjuntos de sentidos. Lo anterior se puede expresar de la siguiente manera:

$$Sim_{Wu\ y\ Palmer}(S_1, S_2) = \frac{2 \times\ Dist_{ruta}(S_3, raiz)}{Dist_{ruta}(S_1, S_3) + Dist_{ruta}(S_2, S_3) + 2 \times\ Dist_{ruta}(S_3, raiz)}$$

A modo de ejemplo, si se considera la siguiente jerarquía extraída de WordNet, es posible establecer la similitud entre dos conceptos como una función entre la longitud de la ruta que los pone en relación *IS-A* y la posición de esos conceptos en la taxonomía:

FIGURA 2

Taxonomía de los conceptos HILL y COAST, tomado de Jurafsky y Martin (2009)



Luego, la medida de similitud corresponderá al número de enlaces *N* presentes en la distancia de la ruta desde un concepto a otro:

$$Sim_{Wu\ y\ Palmer}(S_1, S_2) = \frac{2 \times N_3}{N_{1+}N_2 + 2 \times N_3}$$

donde N_1 y N_2 corresponden al número de enlaces *IS-A* desde el concepto S_1 hasta el concepto S_2, en relación con su LCS, considerando N_3 como el número de enlaces *IS-A* desde el LCS hasta el nodo raíz de la taxonomía. De esta forma, según los valores de la taxonomía de ENTITY para HILL y COAST:

$$Sim_{Wu\ y\ Palmer}(hill, coast) = \frac{2 \times 3}{2 + 2 + 2 \times 3} = 0{,}6$$

En este caso, la medida de similitud para las unidades “hill” y “coast” requiere establecer GEOLOGICAL-FORMATION como la superclase específica común. Luego N_1 y N_2 contabilizan dos enlaces hasta GEOLOGICAL-FORMATION, respectivamente, y N_3 contabiliza tres enlaces desde GEOLOGICAL-FORMATION hasta el nodo raíz ENTITY.

4.1.2. Leacock y Chodorow (1998)

Se trata de una medida basada en la longitud de la distancia entre rutas; es decir, la similitud entre dos conceptos correspondería a una función entre la longitud de la ruta que relaciona esos conceptos en una jerarquía conceptual de tipo *IS-A*, y su posición respecto a otros hiperónimos en la taxonomía. En este caso, para las definiciones de sustantivos se utiliza la jerarquía conceptual de WordNet. La medida se formaliza de la siguiente manera (Leacock y Chodorow, 1998):

$$Sim_{Leacock\ y\ Chodorow}(S_1, S_2) = -\log\left(\frac{dist_{node}(S_1, S_2)}{2 \times depth}\right)$$

Para comprender esta medida es necesario considerar, con especial relevancia, dos variables. Primero, el valor de la profundidad de la taxonomía (*depth*) entre los conceptos en análisis, que corresponde a la longitud del camino más corto entre cada *synset* y el nodo raíz de la taxonomía. Segundo, el valor de la distancia entre los nodos, que corresponderá en este caso al LCS entre ambos conceptos. Por ejemplo, si se considera la distancia entre rutas para los conceptos SHORE y HILL, según la figura dos, la profundidad máxima es igual a 5, mientras que la distancia entre sus nodos tiene un valor de 3. Luego, el puntaje de similitud corresponderá a:

$$Sim_{Leacock\ y\ Chodorow}(hill, coast) = -\log\left(\frac{3}{2 \times 5}\right) = -\log(0{,}3) = 1{,}20$$

La ventaja de esta medida es que pone en relación la profundidad junto con el concepto que contiene los atributos comunes a los conceptos en análisis. Sin embargo, presenta la dificultad de que el puntaje final se ve afectado significativamente por la presencia o ausencia del hiperónimo común más cercano.

4.2. Medidas de contenido de información

Las medidas basadas en el contenido de información (*information content*, de aquí en adelante IC) constituyen una propuesta que surge a partir de la noción de LCS, donde *P*(*c*) es la probabilidad de que una palabra elegida aleatoriamente sea una instancia del concepto *c*. Así, la frecuencia de la instancia de un concepto *c* en la taxonomía, o *freq*(*c*), se puede calcular a partir de la sumatoria de aparición de las palabras que son hipónimos del nodo al que pertenece *c*:

$$freq(c) = \sum_{w \in w(c)} count\ (w)$$

Luego, la probabilidad de que una palabra elegida aleatoriamente desde un corpus sea una instancia del concepto *c* corresponderá a la frecuencia de *c* dividida por el número total de palabras en el corpus:

$$P(c) = \frac{freq(c)}{N}$$

De esta forma, el IC puede ser representado matemáticamente como el logaritmo negativo de esa probabilidad, puesto que, cuando *P*(*c*) aumenta, el valor de *IC*(*c*) disminuye:

$$IC(c) = -\log P(c)$$

Finalmente, un puntaje de IC alto se puede interpretar como el hecho de que un concepto representa un significado conceptual altamente específico cuando ocurre en un texto, mientras que un puntaje de IC bajo corresponde a un significado conceptual general. Así, el indicador de IC correspondería a la medición de la especificidad de un concepto en cuanto a su significado.

4.2.1. Resnik (1995)

La propuesta de Resnik (1995) es la precursora del concepto de IC. Según esto, se establece que el valor asociado a cada concepto en una jerarquía se basa íntegramente en la evidencia disponible en corpus. A partir de esta aproximación, los conceptos se relacionan de tal manera que un IC alto está ligado a conceptos mayormente especificados. Por otra parte, un IC bajo se relaciona con conceptos menos especificados, o generales. Por ejemplo, el concepto HERRAMIENTA tendrá un IC bajo, mientras que un tipo específico de herramienta, como MARTILLO, tendrá un IC alto, que seguirá aumentando en la medida que existan más subtipos de martillos. La formalización de la medida de Resnik es la siguiente:

$$Sim_{resnik}(S_1, S_2) = IC\big(LCS(C_1, C_2)\big)$$

En términos generales, la medida de Resnik (1995) considera el IC como una cuantificación de la información que dos conceptos tienen en común. Luego, se incorpora el número de ocurrencias que un concepto tiene en un corpus en relación con el número de sentidos disponibles para cada concepto. Además, se integra el valor del LCS entre ambos conceptos, para determinar la información solapada según su IC. Así, según esta medida, la frecuencia en la que un concepto aparezca en el corpus incluirá la frecuencia de todos sus conceptos subordinados. Entonces, el conteo de los conceptos específicos aporta al resultado de los conceptos genéricos. Esto tiene un impacto relevante en el valor de su probabilidad asociada: a mayor probabilidad de una frecuencia de aparición alta, tendrán un valor bajo de IC y, por tanto, se tratará de conceptos generales, o cercanos a metaconceptos.

4.2.2. Jiang y Conrath (1997)

En la propuesta de Jiang y Conrath (1997), primero se utiliza el algoritmo de Resnik (1995) para calcular el IC, y luego se incorpora el cálculo de la longitud de distancia entre rutas:

$$Sim_{Jiang\ y\ Conrath}(S_1, S_2) = IC(S_1) + IC(S_2) - 2(IC(LCS(S_1, S_2))$$

Se trata de una medida híbrida en cuanto a los criterios que selecciona para calcular la similitud del *synset*, puesto que incluye la diferencia entre el contenido de información (*i. e.*, $IC(S_1) + IC(S_2)$) y el hiperónimo común más cercano (*i. e.*, $2(IC(LCS(S_1, S_2))))$.

4.2.3. Lin (1998)

Si bien Lin (1998) concuerda con las definiciones de Resnik (1995) en cuanto a las medidas de similitud, propone que la similitud entre dos conceptos no solamente depende de aquello que tienen en común, sino también de sus diferencias. Por tanto, la medida de Lin propone que, mientras más diferencias existan entre un S_1 y S_2, entonces menos similares serán. Luego se establecen dos conceptualizaciones para dar cuenta de los tipos de relaciones que se pueden establecer entre dos *synsets*:

a. *Commonality*: mientras más en común tengan A y B, más similares serán.
b. *Difference*: mientras más diferencias entre A y B existan, menos similares serán.

Según Torres-Ramos (2012), la medida de Jiang y Conrath (1997) es bastante parecida a la propuesta de Lin (1998), en cuanto a la definición del concepto de similitud que subyace a ellas. No obstante, aunque ambas se basan en el cálculo de la cantidad de información necesaria para escribir la información común entre ambos conceptos, fueron postuladas y publicadas de manera independiente[3]. Luego, la formalización de la medida de Lin (1998) es la siguiente:

$$Sim_{Lin}(S_1, S_2) = \frac{2 \times \left(IC\left(LCS(S_1, S_2)\right)\right)}{IC(S_1) + IC(S_2)}$$

Específicamente, esta medida está diseñada para representar la similitud como una función entre el IC dado el LCS, dividido por la suma del IC de ambos.

3 Si bien el trabajo de Lin (1998) no hace referencia a la propuesta de Jiang y Conrath (1997), considera como punto de partida fundamentalmente los trabajos ya citados de Resnik (1995) y de Wu y Palmer (1994).

5. Conclusión

En este trabajo se han presentado diferentes métodos de WSD. En primer lugar, los métodos basados en algoritmos de aprendizaje automáticos o conocimiento contextual, divididos en aprendizaje supervisado y no supervisado. Posteriormente, se han revisado las medidas de relación semántica, correspondientes a los métodos Lesk (Lesk, 1986, 1987) y Lesk adaptado (Banerjee y Pedersen, 2002, 2003). Luego, las medidas de similitud semántica, que se pueden dividir en medidas de distancia entre rutas, correspondientes a las propuestas de Wu y Palmer (1994) y Resnik (1995), y medidas basadas en contenido de información, que consideran los aportes de Resnik (1995), Jiang y Conrath (1997) y Lin (1998). Todas estas medias requieren como insumo la incorporación de fuentes exógenas de conocimiento, ya sean las glosas de las definiciones a través de un recurso lexicográfico, o bien una ontología o tesauro que provea relaciones léxicas de significado, en el marco de una jerarquía conceptual, necesarias para establecer los parámetros de distancia entre rutas o contenido de información. En términos generales, las medidas de similitud se basan en relaciones jerárquicas del tipo IS-A (hiponimia/hiperonimia) o HAS-PART (meronimia), mientras que las medidas de relación utilizan como fuente de conocimiento relaciones más amplias, basadas en información lexicográfica. Finalmente, la mayoría de las aplicaciones de PLN que incorporan soluciones para tareas de WSD optan por este tipo de medidas. Una de las ventajas de estos métodos es que no derivan sus resultados únicamente a partir de la eficiencia estadística, sino que son capaces de explicar el efecto de sus parámetros a partir del conocimiento lingüístico del que depende su rendimiento.

6. Bibliografía citada

Ali, Ashraf, Fayez Alfayez y Hani Alquhayz, 2018: "Semantic similarity measures between words: a brief survey", *Sci. Int. (Lahore)* 30, 907-914.

Aung, Nyein Thwet, Khin Mar Soe y Ni Lar Thein, 2011: "A Word Sense Disambiguation System Using Naïve Bayesian Algorithm for Myanmar Language", *International Journal of Scientific & Engineering Research* 2 (9), 1-7.

Banerjee, Satanjeev, y Ted Pedersen, 2002: "An adapted Lesk algorithm for word sense disambiguation using WordNet", comunicación presentada en The Third International Conference on Intelligent Text Processing and Computational Linguistics.

Banerjee, Satanjeev, y Ted Pedersen, 2003: "Extended gloss overlaps as a measure of semantic relatedness", comunicación presentada en The 18th International Joint Conference on Artificial Intelligence (IJCAI).

Bouillon, Pierrete, y Federica Busa, 2001: "Qualia and the Structuring of Verb Meaning" en *The Language of Word Meaning*, Cambridge: Cambridge University Press, 149-167.

CARPUAT, Marine, y Dekai WU, 2005: "Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation", comunicación presentada en The Second International Joint Conference on Natural Language Processing (IJCNLP).

CHOUEKA, Yaakov, y Serge LUSIGNAN, 1985: "Disambiguation by short contexts", *Computer and the Humanities* 19, 147-157.

COWIE, Jim, Joe GUTHRIE y Louise GUTHRIE, 1992: "Lexical disambiguation using simulated annealing", comunicación presentada en The 14th International Conference on Computational Linguistics (COLING-92).

CRUSE, David Allan, 1986: *Lexical semantics*, Cambridge: Cambridge University Press.

FELLBAUM, Christiane (ed.), 1998: *WordNet: An Electronic Lexical Database*, Cambridge: MIT Press.

IDE, Nancy, y Jean VÉRONIS, 1998: "Introduction to the special issue on word sense disambiguation: the state of the art", *Computational Linguistics* 24 (1), 1-40.

FULMARI, Abhishek, y Manoj CHANDAK, 2014: "An Approach for Word Sense Disambiguation using modified Naïve Bayes Classifier", *International Journal of Innovative Research in Computer and Communication Engineering Organization* 2 (4), 3867-3870.

GAMALLO, Pablo, Susana SOTELO y José Ramom PICHEL, 2014: "Comparing ranking-based and naive bayes approaches to language detection on tweets", *Workshop TweetLID: Twitter Language Identification Workshop at SEPLN 2014*, Girona, Spain.

GOSAL, Gurinder, 2015: "A Naïve Bayes Approach for Word Sense Disambiguation", *International Journal of Advanced Research in Computer Science and Software Engineering* 5 (7), 336-340.

HORNBY, Albert Sidney, Anthony Paul COWIE y Windsor LEWIS, 1974: *Oxford Advanced Learner's Dictionary of Current English*, Londres: Oxford University Press.

JIANG, Jay, y David W. CONRATH, 1997: "Semantic similarity based on corpus statistics and lexical taxonomy", comunicación presentada en The International Conference on Research in Computational Linguistics (ROCLING X).

JURAFSKY, Daniel, y James MARTIN, 2009: *Speech and language processing: an introduction to natural language processing, speech recognition, and computational linguistics*, New Jersey: Prentice Hall.

LEACOCK, Claudia, y Martin CHODOROW, 1998: "Combining local context and WordNet similarity for word sense identification" en Christiane FELLBAUM (ed.): *WordNet: An electronic lexical database*, Cambridge: MIT Press, 265-283.

Lesk, Michael, 1986: "Automatic sense disambiguation: How to tell a pinecone from an ice cream cone", comunicación presentada en The ACM SIGDOC Conference.

Lesk, Michael, 1987: "Can Machine-Readable Dictionaries Replace a Thesaurus for Searches in Online Catalogs?", comunicación presentada en The 3rd Annual Conference of the UW Centre for the New OED: The Uses of Large Text Databases.

Lin, Dekang, 1998: "An information-theoretic definition of similarity", comunicación presentada en The 15th International Conference on Machine Learning.

Lyons, John, 1977: *Semantics*, Cambridge: Cambridge University Press.

Miller, George, 1985: "Wordnet: A Dictionary Browser", *Proceedings of the First Conference of the UW Centre for the New Oxford Dictionary*, University of Waterloo.

Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross y Katherine Miller, 1993: "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography* 3 (4).

Navigli, Roberto, 2009: "Word sense disambiguation: a survey", *ACM Computing Surveys (CSUR)* 41 (2), 1-69.

Núñez, Fredy, 2019: "An experimental review of a supervised method for word sense disambiguation using DAMIEN (Data Mining Encountered)" en Brian Nolan y Elke Diedrichsen (eds.): *Linguistic Perspectives on the Construction of Meaning and Knowledge*, Cambridge: Cambridge Scholars Publishing, 372-386.

Núñez, Fredy, y Beatriz Pérez Cabello de Alba, 2022: "Desarrollo de un sistema de aprendizaje automático supervisado para la desambiguación léxica automática utilizando DAMIEN (Data Mining Encountered)", *RaeL Revista Electronica de Linguistica Aplicada* 21 (1), 150-178.

Popescu, Marius, y Florentina Hristea, 2010: "State of the art versus classical clustering for unsupervised word sense disambiguation", *Artificial Intelligence Review* 35, 241-264.

Pustejovsky, James, 1991: "The Generative Lexicon", *Computational Linguistics* 17, 409-41.

Pustejovsky, James, 1995: *The Generative Lexicon*, Cambridge: The MIT Press.

Pustejovsky, James, y Brad Boguraev (eds.), 1996: *Lexical Semantics: The Problem of Polysemy*, Oxford: Oxford University Press.

Resnik, Phillip, 1995: "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", comunicación presentada en The 14th International Joint Conference on Artificial Intelligence.

Slimani, Thabet, 2013: "Description and Evaluation of Semantic Similarity Measures Approaches", *International Journal of Computer Applications* 80, 25-33.

Torres-Ramos, Sulema, 2012: "Estudio sobre métodos tipo Lesk usados para la desambiguación de sentidos de palabras", *Research in Computer Science* 47, 139-148.

Ustalov, Dmitry, Denis Teslenko, Alexander Panchenko, Mikhail Chersnoskutov, Chris Biemann y Simone Ponzetto, 2018: "An Unsupervised Word Sense Disambiguation System for Under-Resourced Languages", *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC).*

Vidhu Bhala, Vidyasagard, y Murugappan Abirami, 2014: "Trends in word sense disambiguation", *Artificial Intelligence Review* 42 (2), 159-171.

Weinreich, Uriel, 1964: "Webster's Third: A Critique of its Semantics", *International Journal of American Linguistic* 30, 405-409.

Wilks, Yorick, Dan Fass, Cheng-Ming Guo, James McDonald, Tomy Plate y Brian Slator, 1989: "A tractable machine dictionary as a resource for computational semantics" en Branimir Borugaev y E. J. Briscoe (eds.): *Computational lexicography for natural language processing*, Harlow: Longman, 193-228.

Wu, Zhibiao, y Martha Palmer, 1994: "Verbs semantics and lexical selection", comunicación presentada en The 32nd Annual Meeting on Association for Computational Linguistics.