Ejemplo de aplicación de un modelo basado en grafos para el procesamiento del lenguaje natural

Eva Sánchez Salido

1 de marzo de 2021

1. Introducción

Para la ejemplificación del uso de técnicas basadas en grafos para el procesamiento del lenguaje natural se ha elegido el artículo *Graph-based Word Sense Disambiguation of biomedical documents* de Aguirre, Soroa y Stevenson [1], publicado en 2010 en la revista *Bioinformatics*.

1.1. Desambiguación del sentido de las palabras

La desambiguación del sentido de las palabras (WSD: Word Sense Disambiguation) es la tarea de procesamiento de lenguaje natural que consiste en la identificación automática del significado de las palabras ambiguas en un cierto contexto. Para los humanos esta tarea es en general inconsciente, mientras que para automatizar este proceso es necesaria la implementación de algoritmos de decisión. Esto es, dada una palabra y sus posibles significados (por ejemplo sus acepciones en un diccionario), buscamos clasificar la aparición de una palabra en un contexto según una o varias de sus posibles interpretaciones. Las características que ofrece el contexto, como por ejemplo las palabras vecinas a la palabra objetivo, proporcionan evidencia para la clasificación.

1.2. Aplicación de la WSD y técnicas para su implementación

Las aplicaciones de la WSD van desde la Recuperación de Información, la traducción automática o la construcción de grafos de conocimiento. La investigación en sistemas de WSD ha avanzado firmemente hasta el punto de alcanzar niveles consistents de precisión en un amplio abanico de tipos de palabras y ambigüedades. Para ello se han empleado diferentes técnicas que van desde los modelos basados en diccionarios que usan el conocimiento codificado en recursos léxicos, a los modelos basados en aprendizaje

automático supervisado donde el clasificador se entrena con un corpus anotado con el sentido de las palabras manualmente, o los modelos no supervidados que agrupan las apariciones de las palabras, de manera que se induce el significado de las mismas. Entre todas estas técnicas, los modelos basados en aprendizaje supervisado han resultado ser los más exitosos durante la primera década de los 2000 [2]. En la última década se ha experimentado con modelos basados en aprendizaje profundo y redes neuronales, mostrando efectividad pero sin conseguir una mejora significativa respecto a modelos tradicionales de aprendizaje supervisado [3]. La mayoría de las soluciones a la WSD usan métodos tradicionales basados en aprendizaje supervisado y bases de conocimiento como WordNet o BabelNet, que siguen consiguiendo los resultados más competitivos [4].

1.3. Random Walks para la WSD

Los caminos aleatorios son una formalización matemática de la trayectoria que resulta de hacer sucesivos pasos aleatorios. Es bastante frecuente en el estudio de los caminos aleatorios que el espacio donde se requiere realizar el camino se modele con un grafo. La situación suele ser la siguiente: dado un grafo G, y comenzando en uno de sus vértices, seleccionamos de alguna manera al azar uno de sus vecinos y nos movemos a este. Después seleccionamos un vecino de este último vértice y nos movemos de nuevo, y así sucesivamente. La sucesión aleatoria de vértices obtenidos de esta forma es un camino aleatorio sobre el grafo G.

2. *Graph-based Word Sense Disambiguation of biomedical documents*

En [1] se presenta un enfoque basado en grafos para la WSD en el ámbito biomédico. La literatura científica en ámbito biomédico es tan extensa que para acceder a ella en modo efectivo son necesarias herramientas para la automatización, pero la recuperación de información se encuentra con el obstáculo de que el lenguaje natural puede ser ambiguo. Por ejemplo, en el dominio biomédico, la palabra *cold* en inglés es ambigua ya que puede referirse, al menos, al frío o al resfriado común. Al desambiguar el sentido de las palabras es posible mejorar las búsquedas en literatura asegurando que se obtienen tan solo los documentos que contienen los términos en su acepción más apropiada, lo cual es a su vez beneficioso para otras aplicaciones en ámbito biomédico como la indización automática o el descubrimiento de conocimiento.

2.1. Un enfoque no supervisado

A pesar de que los sistemas más populares de WSD se basan en aprendizaje supervisado, estos necesitan datos de entrenamiento etiquetados, a veces difíciles de conseguir y sobre todo costosos en cuanto al trabajo manual que requieren.

El método utilizado en [1] se basa en aprendizaje no supervisado, con lo cual no requiere datos etiquetados para el entrenamiento. En su lugar hace uso del conocimiento del UMLS (Unified Medical Language System) de la National Library of Medicine, que proporciona tanto terminología específica como estándares de clasificación y codificación y otros recursos para promocionar la creación de sistemas de información biomédica más eficientes e interoperables. A pesar de que la mayoría de métodos han explorado el problema de la WSD en dominios generales, en este caso nos encontramos con un enfoque específico en el ámbito biomédico. En particular el sistema hace uso del Metatesauro del UMLS, el cual se convierte en un grafo sobre el cual se aplica el algoritmo de PageRank personalizado para llevar a cabo la WSD.

2.2. PageRank

Uno de los algoritmos más famosos que emplean caminos aleatorios es PageRank. Este algoritmo es un método para clasificar los vértices en un grafo según su importancia estructural relativa que utiliza caminos aleatorios. Fue desarrollado originalmente para clasificar las páginas de la World Wide Web basándose en el número de páginas que enlazan a cada una de ellas.

La puntuación *PageRank* de un vértice es la probabilidad de encontrarnos en ese vértice al azar. En concreto, dado un grafo G con N vértices $(v_1, \dots v_N)$, llamamos $In(v_i)$ al conjunto de vértices que tienen un enlace que llega hasta v_i, y d_j el número de enlaces que salen de v_j. Se define el PageRank del vértice v_i como:

$$P(v_i) = c \sum_{v_j \in In(v_i)} \frac{1}{dj} P(v_j) + (1 - c)\frac{1}{N} \tag{1}$$

donde c es un factor de amortiguación que toma un valor entre 0 y 1. El PageRank se calcula entonces aplicando un algoritmo iterativo que ejecuta la ecuación 1 repetidamente hasta alcanzar la convergencia bajo un umbral especificado o hasta haber realizado un número de iteraciones concreto. La figura 1 muestra el ejemplo de un grafo y los valores de PageRank obtenidos, donde P se ha inicializado con distribución uniforme (es decir, con valor 0,25 en cada nodo).

En el caso de la WSD basada en grafos, es conveniente poder incluir información acerca de la importancia relativa de los vértices en el grafo. En otras palabras, dado un conjunto de vértices de interés, nos gustaría conocer qué vértices están relacionados con ellos. Por ejemplo, en el grafo del ejemplo podría interesarnos saber qué nodos en el grafo están relacionados con D, como muestra la figura 2. PageRank personalizado

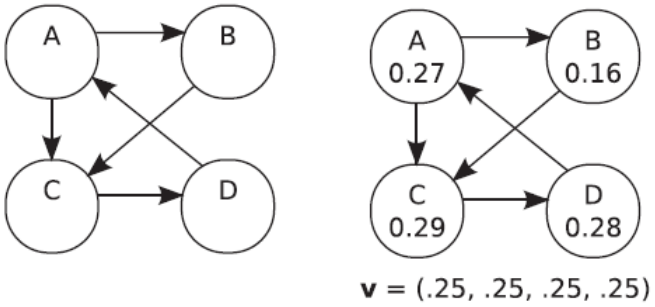

Figura 1: Aplicación del algoritmo PageRank.

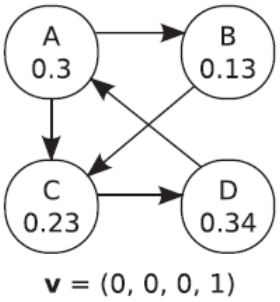

Figura 2: Personalización de PageRank.

computa la importancia estructural de los vértices en un grafo cuando unos vértices son más relevantes que otros para una tarea concreta. Para implementar PageRank personalizado reescribimos la ecuación 1 en forma compacta mediante el uso de matrices. Llamamos M a la matriz de transición de probabilidad de dimensión $N \times N$, donde $M_{ji} = \frac{1}{d_i}$ cuando existe un enlace desde v_i hasta v_j, y vale 0 en caso contrario. Llamamos v a un vector estocástico normalizado de dimensión $N \times 1$ cuyos elementos son todos $\frac{1}{N}$. Entonces el cálculo del vector PageRank (P) sobre el grafo G es equivalente a resolver la ecuación

$$P = cMP + (1 - c)v \tag{2}$$

2.3. PageRank Personalizado para la WSD

Para usar el algoritmo de PageRank personalizado para la WSD son necesarios dos recursos: una base de conocimiento (formado por conceptos y relaciones representadas por un grafo) y un diccionario (que relaciona palabras y expresiones de documentos con los posibles conceptos de la base de conocimiento).

En este caso se representa el Metatesauro del UMLS como un grafo en el que los conceptos son vértices y las relaciones entre ellos las aristas. En el ejemplo de la palabra *cold*, se inicializa v con valores iguales para todos los conceptos que aparecen en el contexto y cero en el resto para aplicar PageRank personalizado y seleccionar el concepto de *cold* con el PageRank más alto.

2.4. Evaluación y resultados

Al evaluarlo en el dataset NLM-WSD, el algoritmo supera los resultados que alcanzan otros sistemas que usan el Metatesauro del UMLS como base de conocimiento. Se alcanza una precisión del 68,1 %, que supera los resultados que se obtienen al aplicar PageRank de forma aleatoria. Existe el debate sobre cuál debe ser la precisión mínima de un sistema de WSD para que pueda mejorar el rendimiento de otras tareas. Algunos experimentos sugieren un mínimo del 90 % de desambiguaciones correcta para poder mejorar la tarea de Recuperación de Información. Sin embargo, otros experimentos muestran que un rendimiento más bajo, incluso del 65 %, puede mejorar igualmente los resultados, lo que avala la utilidad del sistema para la mejora de otras tareas de procesamiento del lenguaje.

3. Conclusiones finales

Afrontar el problema de la desambiguación del sentido de las palabras con el método propuesto tiene la ventaja de no necesitar datos etiquetados para el entrenamiento al ser un método no supervisado. Esto permite además poder aplicarlo a diversas situaciones. En concreto en el artículo elegido se ha aplicado este algoritmo basado en grafos para mejorar la recuperación de información en el dominio biomédico, donde la ambigüedad de los términos puede suponer una dificultad importante y sobre todo determinante.

En general, se ha demostrado la utilidad de los sitemas de desambiguación del sentido de las palabras para la mejora de diversas aplicaciones desde la recuperación de información multilingüe, la traducción automática y la extracción de información.

Además, la herramienta estudiada es de código abierto, programada en C++ y puede ser fácilmente integrada en software de terceros a modo de librería. De hecho, el paquete de código abierto multilingüe de procesamiento de textos *FreeLing* lo incorpora entre sus funcionalidades.

Referencias

[1] Eneko Agirre, Aitor Soroa y Mark Stevenson. «Graph-based word sense disambiguation of biomedical documents». En: *Bioinformatics* 26.22 (2010), págs. 2889-2896 (vid. págs. 1, 2, 3).

[2] P. Edmonds y E. Agirre. «Word sense disambiguation». En: *Scholarpedia* 3.7 (2008). revision #90370, pág. 4358. DOI: 10.4249/scholarpedia.4358 (vid. pág. 2).

[3] Luyao Huang y col. *GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge*. 2020. arXiv: 1908.07245 [cs.CL] (vid. pág. 2).

[4] Yinglin Wang, Ming Wang y Hamido Fujita. «Word sense disambiguation: A comprehensive knowledge exploitation framework». En: *Knowledge-Based Systems* 190 (2020), pág. 105030 (vid. pág. 2).