Desambiguación del sentido de las palabras

Jordy Joaquín Cuan Robledo [1], Eduardo Cantorán Flores [1],
Rafael Pérez Aguirre [1]
[1] Facultad de Ciencias de la Computación, Benemérita Universidad
Autónoma de Puebla, México

1. Introducción

En la última década con el uso de la tecnología se ha generado una cantidad inmensa de información sin embargo todo esto no tiene sentido si no se interpretan todos estos datos. Para que una computadora pueda procesar toda esta información de manera eficaz se hace uso de la WSD que es de gran relevancia en el procesamiento del lenguaje natural.

2. Ambigüedad en el lenguaje

La ambigüedad en el lenguaje es aquella palabra que no tiene un solo significado, este puede ser interpretado de distintas formas lo cual genera cierta confusión.
Puede surgir por aspectos léxicos, sintácticos, por referencia, entre otros.

2.1. Ambigüedad semántica de la palabra

Esta surge cuando el contexto cambia por completo el significado de la palabra. Podemos mencionar como ejemplo la palabra gato la cual puede tener tres interpretaciones diferentes ya que puede significar gato como animal, gato como una herramienta hidráulica o gato como un juego. Este tipo de palabras se denominan como polisemia.

3. Desambiguar el Sentido de las Palabras

WSD es el proceso de elegir el significado más adecuado de una palabra acorde a su contexto. Dicho contexto hace referencia a las demás palabras que conforman a la oración.
Algunas aplicaciones relevantes son la traducción automática, la recuperación de información y la búsqueda de respuestas.

3.1. Problemas

Hay tres tipos de problemas que presenta utilizar WSD que son: determinar qué repositorio utilizar o cuál tendría mayor relevancia para los resultados que se deseen obtener, modelar una técnica de desambiguación y evaluar el porcentaje de efectividad de dicho procedimiento.

4. Enfoques de WSD

4.1. Métodos basados en conocimiento

Estos métodos se basan en utilizar recursos externos tales como diccionarios o tesauros. La idea de esta técnica es proporcionar una lista la cual contiene los significados de las palabras. Un ejemplo de estos diccionarios puede ser el WordNet.
Algunos otros hacen uso de las traducciones de otro lenguaje los cuales utilizan un corpus en otro idioma buscando todas las posibles traducciones de una palabra, se cuenta contabiliza el número de traducciones y se elige aquella que sea mayor.
El principal problema de este tipo de métodos es no contar con grandes repositorios y la mayoría solo existen en pocos idiomas.

4.2. Métodos supervisados

Este tipo de métodos se basan en utilizar un clasificador que dictamine el significado que mejor se adapte a la oración.
Hacen uso de un corpus etiquetado por ejemplo SemCor y Senseval. Principalmente se dividen en dos secciones: el entrenamiento y las pruebas.
Los algoritmos que más se utilizan son Bayes y SVM. Utilizan las frecuencias de las palabras, el total de palabras, sus lemas y posiciones.
Para la fase del entrenamiento se hace uso del corpus etiquetado para así encontrar las relaciones entre las palabras. Son mínimamente supervisados.
Después, en la fase de pruebas se analiza el texto contra el modelo generado por el entrenamiento y se calcula el porcentaje de efectividad.
Un problema puede ser que los corpus que se utilizan tiene un algo nivel de polisemia.

4.3. Métodos no supervisados

Estos métodos discriminan el sentido de las palabras, es decir, identifican patrones sin la necesidad de un dataset.
Se basan en que las palabras con significados iguales normalmente tienen contextos iguales, es decir, que reúnen las palabras respecto a la similitud del contexto. Como no se tiene un sentido de las palabras simplemente se basa en la discriminación.
La gran ventaja de estos métodos es que no dependen de un corpus. Sin embargo por desventaja tenemos que algunos grupos no contienen el verdadero significado y es complicado medir la similitud de pequeños contextos.

5. Evaluación

Para medir la efectividad y rendimiento de los métodos de desambiguación de las palabras se utilizan algunas métricas que conllevan la precisión, Recall y F-Measure.