



Técnicas de Procesamiento del Lenguaje Natural en la Recuperación de Información

Martin Tlacuatl Lozada¹, CHANTAL DANIELA AGUIRRE RAMIREZ¹ y CINTHIA SALAMANCA LÓPEZ¹

¹ Benemérita Universidad Autónoma de Puebla. martinmcfly42@gmail.com

El procesamiento del lenguaje natural es una rama de la ciencia que pertenece a la intersección de la lingüística aplicada y las ciencias de la computación, que estudia los métodos necesarios para que la computadora pueda ejecutar varias tareas relacionadas con el lenguaje humano, como el español, y requiere cierto grado de "entendimiento" de su contenido. Por otro lado, esta ciencia desarrolla las herramientas que ayudan al lingüista en su trabajo cotidiano e incluso pueden llevar a descubrimientos lingüísticos nuevos.

En la etapa pasada la lingüística computacional se ocupaba principalmente de codificar claramente el conocimiento lingüístico, y en este sentido fue, como lo indica su nombre, una rama de la lingüística. En la etapa contemporánea se ha convertido en una rama de la ciencia de la inteligencia artificial; a saber, el aprendizaje automático. Un reto muy interesante es el aprendizaje no supervisado (de los textos disponibles no marcados manualmente) del conocimiento lingüístico y de los datos (diccionarios y gramáticas) necesarios para desarrollar los sistemas prácticos.

Este artículo presenta las aplicaciones principales del procesamiento automático de lenguaje natural y discuten los métodos que se emplean para la resolución de problemas. Se describe el efecto de la integración de varias técnicas basadas en el procesamiento del lenguaje natural en sistemas de recuperación de información. Se estudiarán, en concreto, métodos de lematización entre otros. Una evaluación a gran escala con colecciones de documentos en español nos permitirá verificar que la combinación de estas técnicas con otras menos sofisticadas, tales como la eliminación de palabras gramaticales, contribuye a una mejora significativa de la calidad de los sistemas de recuperación.