INTRODUCCIÓN

En la actualidad la cantidad de información a la que se puede tener acceso es enorme, ésta puede provenir de libros, revistas, recursos electrónicos, entre otros. Estos medios de información ciertas veces incluyen términos nuevos o desconocidos que deben ser buscados para comprender de manera óptima la información. Sin embargo, a pesar de la importancia que tienen en la vida académica y cotidiana los términos, su obtención es una tarea complicada y costosa, tanto en recursos como en tiempo, que tradicionalmente se lleva a cabo de manera manual con la ayuda de expertos en terminología como del área de donde se quieran obtener los términos.

Por lo anterior, es que desde la década de 1990, y más recientemente desde el año 2000, se han desarrollado numerosos sistemas extractores de términos que emplean diversos conocimientos, ya sean lingüísticos, estadísticos o híbridos, para la obtención de los términos de un conjunto de documentos de manera automática. Esto para emplear los términos en la creación de diccionarios y glosarios, verbigracia, pero igualmente en el enriquecimiento de sistemas de traducción automática, sistemas de generación de texto, bases de conocimiento, etcétera.

Pero aún con el desarrollo de diversos sistemas de extracción terminológica, pocos de ellos emplean recursos léxicos, como los son Wikipedia o EuroWordNet, para llevar a cabo una validación de los candidatos a término que se obtienen de los extractores terminológicos. Siendo que estos tienen como ventaja el poder obtener listas de unidades terminológicas mucho más fiables y acordes al área de análisis; así como la inclusión de información extra que sólo se encuentra en los recursos léxicos, como la relación entre diversos términos, sinónimos, abreviaturas, etcétera.

Objetivo

El objetivo general de esta tesis es desarrollar un sistema extractor de términos que extraiga la mayor cantidad de candidatos a término y que a su vez los resultados que se obtengan de éste se validen empleando la enciclopedia gratuita Wikipedia.

Asimismo, como un objetivo particular de esta tesis es la obtención de listas de términos validadas, es decir, un conjunto de unidades terminológicas propias de un corpus determinado o de una selección de él.

Estructura de la tesis

Esta tesis está conformada por cinco capítulos; los dos primeros capítulos son los antecedentes necesarios para la comprensión de la tesis. De manera más específica, el primer capítulo es sobre el procesamiento de lenguaje natural, donde se hablará sobre lo que es, algunos recursos y herramientas que emplea y una de sus aplicaciones, la recuperación de información; en cambio, en el segundo capítulo se abordará lo que es la terminología, no sólo como disciplina sino también como conjunto de términos, asimismo se darán a conocer algunos de los extractores terminológicos desarrollados en los últimos años, los métodos de evaluación de estos y algunos de los recursos léxicos más representativos que existen.

En el tercer capítulo se dará a conocer la metodología empleada en el desarrollo de esta tesis, es decir, el corpus que se empleó, el método de extracción terminológica seleccionado, el método de validación que se usó para validar los resultados del extractor de términos y la arquitectura del sistema desarrollado para la obtención de términos y su validación empleando Wikipedia.

A lo largo del cuarto capítulo se describirán los resultados obtenidos tanto del extractor terminológico como los obtenidos de la validación en donde se empleó la enciclopedia Wikipedia. Asimismo, se hablará sobre el método de evaluación de los resultados empleado y cómo es que se llevó a cabo éste. De igual manera, se darán a conocer algunas observaciones que se vieron durante el proceso de validación de términos y de evaluación del sistema desarrollado en la tesis.

Finalmente, en el quinto capítulo se mostrarán las conclusiones a las que se llegaron al finalizar este trabajo de tesis, así como el trabajo a futuro que se puede llevar a cabo.