

# Resumen

Este trabajo de tesis se enmarca en los campos del Procesamiento del Lenguaje Natural, área de la ciencia y la tecnología encargada del procesamiento automático del lenguaje natural o lenguaje humano, y de la Recuperación de Información, cuyo objetivo es el de identificar, dada una colección de documentos, aquéllos que son relevantes a una necesidad de información del usuario.

Los sistemas convencionales de Recuperación de Información emplean técnicas estadísticas basadas en la distribución de los términos en el documento y en la colección para estimar la relevancia de un documento. Sin embargo, dado que un proceso de Recuperación de Información exige que el sistema comprenda en cierta medida el contenido del mismo, dicha tarea puede verse como perteneciente al ámbito del Procesamiento del Lenguaje Natural. Este razonamiento se ve apoyado por el hecho de que el mayor problema en la Recuperación de Información es la variación lingüística del idioma, consistente en que un mismo concepto se puede expresar de formas diferentes mediante modificaciones en la expresión.

El objetivo de esta tesis es el desarrollo de tecnología de base para el Procesamiento del Lenguaje Natural y el estudio de la viabilidad de su aplicación en sistemas de Recuperación de Información sobre documentos en español. Si bien existen estudios similares para otras lenguas, con un claro dominio del inglés, el español ha quedado relegado frecuentemente a un segundo plano. Además, la mayor complejidad lingüística del español frente al inglés en todos sus niveles no permite una extrapolación inmediata al español de los resultados obtenidos para el inglés, demandando la realización de experimentos específicos.

Por otra parte, hemos tenido que hacer frente a uno de los principales problemas en la investigación del procesamiento automático del español, la carencia de recursos lingüísticos libremente accesibles. La solución para minimizar este problema pasa por restringir la complejidad de las soluciones propuestas, centrándose en la utilización de información léxica, de obtención más sencilla. El hecho de acotar la complejidad de las aproximaciones planteadas permite que las técnicas desarrolladas sean fácilmente adaptables a otros idiomas de características y comportamiento similar, constituyendo de este modo una arquitectura general aplicable a otros idiomas mediante la introducción de las modificaciones oportunas, como sería el caso, por ejemplo, del gallego, el portugués o el catalán. Por otra parte, a la hora de minimizar el coste computacional de nuestras propuestas de cara a su aplicación en entornos prácticos, se ha hecho amplio uso de tecnología de estado finito.

En este contexto, en primer lugar, se ha desarrollado un preprocesador-segmentador avanzado de base lingüística para la *tokenización* y segmentación de textos en español, orientado principalmente a la etiquetación robusta del español, pero aplicable a otras tareas de análisis. Las modificaciones concretas para su aplicación en Recuperación de Información son también discutidas.

A nivel flexivo, se ha estudiado la utilización de técnicas de desambiguación-lematización para la normalización de términos simples, empleando como términos de indexación los lemas de las palabras con contenido del texto —nombres, adjetivos y verbos.

A nivel derivativo, se ha desarrollado una herramienta de generación automática de familias morfológicas —conjuntos de palabras ligadas derivativamente y que comparten la misma raíz— para su utilización también en tareas de normalización de términos simples. Esta propuesta, sin embargo, no es todavía por completo inmune a los problemas generados por la introducción de ruido por sobregeneración durante la creación de familias.

A nivel sintáctico se ha ensayado una aproximación basada en la utilización de dependencias sintácticas a modo de términos índice complejos más precisos y para tratar la variación sintáctica y morfosintáctica. Estas dependencias son generadas mediante análisis sintáctico superficial empleando dos analizadores de desarrollo propio: PATTERNS, basado en patrones, y CASCADE, basado en cascadas de traductores finitos. Se ha estudiado también el empleo de diferentes fuentes de información sintáctica, consultas o documentos —siendo ésta última más efectiva—, y la restricción a dependencias nominales para la reducción de costes.

Finalmente, también a nivel sintáctico, se ha evaluado una nueva aproximación pseudo-sintáctica que emplea un modelo sustentado sobre similaridades en base a distancias para la reordenación de resultados obtenidos mediante una aproximación clásica basada en la indexación de lemas. De las dos propuestas planteadas, la primera basada en la mera reordenación conforme el nuevo modelo, y la segunda basada en una nueva aproximación a la fusión de datos mediante intersección de conjuntos, ésta última ha sido la más fructífera.