

1. Introducción

1.1. Antecedentes

Este trabajo surge dentro del marco del proyecto *Extracción terminológica en textos de especialidad*, llevado a cabo dentro del Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería de la UNAM, con el apoyo del Consejo Nacional para la Ciencia y la Tecnología (CONACyT), proyecto 82050.

Este proyecto, y en particular esta investigación, se ha enfocado en la extracción automática de contextos definitorios especializados. Particularmente, se ha creado un sistema de extracción automática llamado Ecode (Extractor de Contextos Definitorios) (Alarcón, 2009).

Con anterioridad, dentro del mismo GIL, ya se han realizado trabajos y tesis dentro de este mismo proyecto orientados a la extracción de contextos definitorios en textos especializados de manera automática. Con respecto a investigaciones teóricas sobre los contextos definitorios se han desarrollado trabajos tales como: *Análisis Lingüístico de definiciones en contextos definitorios* (Aguilar, 2009), *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática* (Hernández, 2009) y *La funcionalidad al interior de contextos definitorios con definiciones analíticas: el patrón sintáctico para + infinitivo* (Sánchez, 2009). Asimismo, se han publicado trabajos prácticos como: *Agrupamiento semántico de contextos definitorios* (Molina, 2009) o *Metodología de elaboración para un corpus informático de contextos definitorios* (Mijangos, 2011).

Los trabajos del Dr. Rodrigo Alarcón con relación al desarrollo del Ecode han sido: *Análisis lingüístico de contextos definitorios en textos de especialidad* (Alarcón, 2003) y *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios* (Alarcón, 2009).

1.2. Problemática

El problema que aborda esta tesis surge a raíz de los trabajos realizados con anterioridad, especialmente de los elaborados por Rodrigo Alarcón orientados a la extracción automática

de contextos definitorios. Como ya se ha dicho, este proceso se ha estado llevando a cabo por una herramienta denominada Ecode. Sin embargo, los trabajos de Rodrigo Alarcón aún requerían de refinamiento y necesitaban ser abordados desde una perspectiva desde el punto de vista ingenieril.

Es necesario que el Ecode extraiga de una manera más eficiente y rápida los contextos definitorios, y que la precisión y la exhaustividad de tal herramienta mejoren. Se busca que el Ecode sea una herramienta flexible y adaptable, es decir, que opere conforme a las necesidades que los usuarios (investigadores, alumnos, becarios, etc.) necesiten.

1.3. Objetivos de la tesis

Con relación a esta problemática y a lo dicho con anterioridad, me planteo los siguientes objetivos para la realización de esta tesis enfocados a las necesidades expuestas.

1.3.1 Objetivos generales

Desarrollar una herramienta integral que no requiera procesos intermedios para la obtención de resultados, integrando plenamente los procesos internos y los de entrada y salida que son requeridos.

1.3. 2 Objetivos particulares

- Desarrollar para Ecode un módulo de entrada que flexibilice la alimentación del texto requerido y lleve a cabo todo el pre-procesamiento requerido para que el sistema funcione.
- Optimizar el sistema proponiendo acciones concretas de ajuste al algoritmo y al código.
- Definir un módulo de salida que permita entregar resultados fáciles de interpretar y de integrar a otros sistemas.
- Generar una interfaz de usuario.
- Reescribir porciones del código redundantes o improductivas.

1.4. Estructura de la tesis

Con respecto a las necesidades y objetivos de esta investigación, se plantea comenzar la tesis por los temas más generales para abarcar al final los más particulares y mostrar las conclusiones a las que se llegó. La estructura de la tesis es la que se muestra a continuación:

1. Conceptos básicos: En este apartado planteo los conceptos que utilizaré a lo largo de la tesis. De manera general, se explica en qué consiste el procesamiento de lenguaje natural (PLN), la extracción terminológica, los contextos definitorios, el etiquetado POS, la lematización y los corpus lingüísticos. Por otra parte, también explicaré las herramientas utilizadas en el sistema final.
2. Ecode, una herramienta de extracción de CDs en textos de especialidad: Este capítulo aborda la temática acerca del sistema Ecode, su descripción, el algoritmo utilizado, las entradas y salidas del sistema, etc. También se analiza la funcionalidad del Ecode y del Describe, herramienta para la cual fue creado el Ecode.
3. Desarrollo y trabajo sobre Ecode: Este es el capítulo central para la tesis, pues aquí se describe el trabajo que se realizó sobre el primer Ecode, su optimización, la implementación web que se hizo de este y en general se describirá el trabajo que realicé conforme a las necesidades y objetivos planteados en el marco de la investigación de esta tesis.
4. Resultados y pruebas: En este apartado se muestran las pruebas que se realizaron con las optimizaciones realizadas al sistema y los resultados que se arrojaron. Se comparan los resultados estadísticos de precisión y exhaustividad conforme a los del Ecode original y se medita sobre los objetivos planteados en la introducción con base en los resultados obtenidos.
5. Conclusiones: Finalmente, se muestran las conclusiones sobre los logros alcanzados en esta tesis, su utilidad y el trabajo a futuro.