

1. Introducción 9

1.1. ANTECEDENTES.....	9
1.2. PROBLEMÁTICA.....	9
1.3. OBJETIVOS DE LA TESIS	10
1.3.1 <i>Objetivos generales</i>	10
1.3.2 <i>Objetivos particulares</i>	10
1.4. ESTRUCTURA DE LA TESIS	11
2. CONCEPTOS BÁSICOS	12
2.1. PROCESAMIENTO DE LENGUAJE NATURAL	12
2.1.1 <i>Extracción terminológica</i>	13
2.1.2 <i>Contextos definitorios</i>	14
2.1.3 <i>Etiquetado de partes de la oración</i>	17
2.1.4 <i>Lematización</i>	23
2.1.5 <i>Corpus</i>	24
2.2 HERRAMIENTAS UTILIZADAS Y EL LENGUAJE PERL Y SUS MÓDULOS	25
2.2.1 <i>Herramientas externas</i>	25
2.2.2 <i>Perl y el procesamiento de texto</i>	26
3. ECODE.....	29
3.1. DESCRIPCIÓN Y ANÁLISIS DE ECODE.....	29
3.1.1 <i>Entrada al sistema</i>	31
3.1.2 <i>Algoritmo General</i>	34
3.1.3 <i>Filtrado e Identificación de término y definición</i>	37
3.1.4 <i>Salida del Sistema</i>	53
3.2. EMPLEO DEL ECODE.....	54
3.2.1 <i>Página del Ecode.</i>	54
3.2.2 <i>Describe</i>	55
3.3. RECAPITULACIÓN DE PROBLEMAS Y LÍNEAS DE TRABAJO	57
4. DESARROLLO Y TRABAJO SOBRE EL SISTEMA ECODE.....	59
4.1 OPTIMIZACIÓN DEL ECODE ORIGINAL	60
4.1.1 <i>Operaciones de Entrada/Salida</i>	61
4.1.2 <i>Simplificación de código</i>	62
4.1.3 <i>Impacto de la optimización del Ecode original y trabajo futuro</i>	73
4.2 EL MÓDULO DE ENTRADA DE DATOS.....	73
4.2.1 <i>La fuente de los datos</i>	74
4.2.2 <i>El problema de las codificaciones</i>	75
4.2.3 <i>Separador de oraciones</i>	76
4.2.4 <i>El etiquetador POS</i>	77
4.3 EL MÓDULO DE SALIDA	80
4.3.1 <i>Filtrado y agrupamiento de términos semejantes y asociación de CDs por sus términos</i>	81
4.3.2 <i>El hash de CDs</i>	82
4.4. IMPLEMENTACIÓN DE ECODE.....	84
4.4.1 <i>Interfaz de usuario web de Ecode (WebCode)</i>	85
4.4.2 <i>Interfaz de consola</i>	85
4.4.3 <i>Integración a Describe</i>	87
5. RESULTADOS Y PRUEBAS	89
5.1 SIMPLIFICACIÓN DE CÓDIGO	89
5.1.1 <i>Reducción de operaciones de entrada y salida</i>	90

5.1.2 <i>Compresión del código</i>	92
5.2 TIEMPO DE EJECUCIÓN	93
5.2.1 <i>Tiempo de ejecución de etiquetadores POS</i>	93
5.2.2 <i>Tiempo de ejecución del núcleo de Ecode</i>	94
5.3 PRECISIÓN Y EXHAUSTIVIDAD DE LOS CONTEXTOS EXTRAÍDOS CONTRA EL ECODE ORIGINAL	95
5.4 RECAPITULACIÓN DEL TRABAJO	96
6. CONCLUSIONES	98
6.1 RECAPITULACIÓN DEL TRABAJO	98
6.2 CONCLUSIONES.....	98
6.2.1 <i>Conclusiones sobre la optimización</i>	98
6.2.2 <i>Sobre la expansión</i>	99
6.2.3 <i>Sobre la adaptación</i>	99
6.3 TRABAJO FUTURO	99
6.3.1 <i>Trabajo futuro sobre los módulos de entrada y salida</i>	100
6.3.2 <i>Trabajo futuro sobre el núcleo del sistema</i>	101
7. BIBLIOGRAFÍA	102

Índice de figuras

- 2.1 Estructura de un CD
- 2.2 Un árbol de decisión de TreeTagger
- 2.3 Árbol de sufijos de longitud tres
- 3.1 Panorama general de la arquitectura de Ecode
- 3.2 Extracción de candidatos a CD
- 3.3 Filtro de candidatos no relevantes
- 3.4 Identificación de términos y definiciones
- 3.5 Árbol de decisión para la identificación de términos y definiciones
- 3.6 Evaluación de CDs
- 3.7 Esquema del sistema Describe
- 3.8 Flujo del Sistema Describe
- 4.1 Panorama final del Ecode
- 4.2 El módulo de entrada de Ecode
- 4.3 El módulo de salida de Ecode
- 4.4 Ejemplo de agrupamiento de términos semejantes
- 4.5 Interacción de Ecode con la base de datos de Describe
- 5.1 Formulas de precisión y exhaustividad

Índice de tablas

- 2.1 Conjunto de etiquetas del CEMC
- 2.2 Ejemplo de regla en el algoritmo de Porter
- 2.3. Número de palabras por lengua y ámbito
- 3.1 Módulos del Ecode
- 3.2 Gramáticas de Ecode
- 3.3 Parámetros de la gramática de patrones verbales
- 3.4 Reglas de filtrado de contextos no relevantes
- 3.5 Elementos de la identificación de términos y definiciones
- 3.6 Inferencias del árbol de decisión
- 3.7 Reglas de evaluación de término y definición
- 3.8 Reglas globales de evaluación de CDs
- 3.9 Ejemplo de términos y de definiciones evaluados
- 3.10 Resultados de GEN por Tipo de definición y PVD en la página de Ecode
- 4.1 Archivos agregados a Ecode
- 4.2 Ocurrencias de operaciones de Entrada/Salida en los módulos de Ecode.
- 4.3 Herramientas para la extracción de texto
- 4.4 Excepciones y reglas de separación de oraciones
- 4.5 Tabla de relación entre etiquetas de etiquetadores
- 4.6 Reglas de filtrado y agrupamiento de términos semejantes
- 4.7 Estructura del Hash de CDs
- 4.8 Argumentos y uso de la interfaz de consola de Ecode
- 5.1 Ocurrencias de operaciones de Entrada/Salida en los módulos de Ecode después de su

optimización

5.2 Número de líneas de cada archivo del Ecode original y el final

5.3 Comparación de tiempos de ejecución de etiquetadores POS

5.4 Comparación de tiempos de ejecución del Ecode original contra el producto de la optimización

5.5 Resultados globales por CDs

5.6 Resultados de precisión y exhaustividad