

<i>Índice General</i>	v
4.7 Resumen	89
5 Descripción y Evaluación del Sistema de Etiquetado Léxico	91
5.1 Descripción del Sistema de Etiquetado	91
5.1.1 Fase de Aprendizaje	92
5.1.2 Fase de Etiquetado	93
5.2 Descripción de los Corpora	94
5.2.1 Wall Street Journal (WSJ)	95
5.2.2 LexEsp	95
5.2.3 BDGEO	96
5.3 Evaluación del sistema de Etiquetado Léxico	99
5.3.1 Evaluación sobre el Corpus WSJ	100
5.3.2 Evaluación sobre el Corpus LexEsp	104
5.3.3 Evaluación sobre el Corpus BDGEO	108
5.4 Etiquetado Léxico usando Modelos Especializados	109
5.4.1 Resultados sobre el Corpus WSJ	109
5.4.2 Resultados sobre el Corpus LexESP	110
5.5 Comparación Experimental de las Prestaciones de Etiquetado	111
5.6 Resumen	113
6 Análisis Sintáctico Superficial	115
6.1 Introducción	115
6.2 Aproximación Unificada al Etiquetado y Análisis Superficial	116
6.3 Formulación Probabilística del Problema	119
6.4 Proceso de Decodificación: Etiquetado y Análisis Superficial	120
6.5 Evaluación del Sistema Integrado	123
6.6 Detección de NP sobre el WSJ	124

6.6.1	Integración de Modelos de Bigramas (BIG)	125
6.6.2	Integración de Modelos ECGI y BIG	129
6.7	Detección de Unidades Sintácticas sobre WSJ	130
6.7.1	Descripción de la Tarea	131
6.7.2	Características de las Unidades Sintácticas	131
6.7.3	Evaluación Experimental	133
6.7.4	Comparación con otras Aproximaciones	137
6.8	Detección de SN sobre LexEsp	140
6.9	Resumen	142
7	Entorno Gráfico para la Desambigüación de Textos	143
7.1	Funcionalidad de la Aplicación	144
7.1.1	Edición de Etiquetas	144
7.1.2	Edición de Gramáticas	145
7.1.3	Visualización y Corrección del Etiquetado Léxico y el Análisis Sintáctico	146
7.1.4	Evaluación de Prestaciones	149
7.2	Ventajas de la Herramienta Gráfica	150
8	Conclusiones y Trabajos Futuros	151
8.1	Conclusiones	151
8.2	Trabajos Futuros	153
8.2.1	Refinamiento de los Modelos	153
8.2.2	Aplicaciones del Sistema Desarrollado	154
A	Conjunto de Categorías Léxicas	157
A.1	Estructura Completa de las Categorías Léxicas PAROLE	158
A.2	Categorías Léxicas PAROLE	163

<i>Índice General</i>	vii
A.3 Categorías <i>Penn Treebank</i>	165
B Corpus BDGEO	171
B.1 Frases del Corpus BDGEO	171
B.2 Etiquetas Completas	173
C Palabras Especializadas en los Modelos Contextuales	175
C.1 Sobre el Corpus WSJ	175
C.2 Sobre el Corpus LexEsp	179
Bibliografía	181

Índice de Figuras

2.1	Proceso de etiquetado léxico.	10
2.2	Análisis global a) de la oración “Luis ve al hombre con el telescopio” .	24
2.3	Análisis global b) de la oración “Luis ve al hombre con el telescopio” .	24
2.4	Análisis parcial de la oración “Luis ve al hombre con el telescopio” . .	25
3.1	Descripción funcional de un etiquetador.	40
3.2	Representación de las secuencias de categorías léxicas posibles para la frase “Este río está seco”.	41
3.3	Representación de las secuencias posibles para la frase “Este río está seco” compatibles con el análisis morfológico.	42
3.4	Representación de las probabilidades de contexto y léxicas mediante un modelo de Markov.	45
3.5	Algoritmo de Viterbi	50
3.6	Proceso de especialización de una palabra w_i en la categoría C_i	62
4.1	Ejemplo de construcción de un modelo de categorías léxicas mediante el algoritmo ECGI.	69
4.2	Algoritmo ECGI.	71
4.3	Notación utilizada para el suavizado de un modelo ECGI	75
4.4	Evolución del número de estados del autómata y del valor $n1/N$ en función de la talla de entrenamiento sobre el corpus WSJ.	78

4.5	Comportamiento de la función de descuento en función de la frecuencia (BDFP) considerando distintos valores de $C(k)$	81
4.6	Distribución de $C(k)$ sobre el corpus WSJ con un conjunto de entrenamiento de 800,000 palabras.	82
4.7	Evaluación de los métodos de suavizado en función de la talla de entrenamiento considerando un modelo léxico equiprobable.	86
4.8	Evaluación de los métodos de suavizado considerando un modelo léxico equiprobable para un conjunto de entrenamiento de 700,000 palabras y uno de prueba de 100,000 palabras.	87
5.1	Descripción del sistema de etiquetado léxico.	92
5.2	Descripción del proceso de etiquetado léxico del corpus BDGEO.	97
5.3	Comparación de etiquetado léxico entre modelos ECGI con diferentes suavizados y los modelos BIG y LEX.	102
6.1	Esquema del sistema integrado de etiquetado léxico y análisis sintáctico superficial.	116
6.2	Proceso de construcción de un modelo de lenguaje integrado.	118
6.3	Trellis parcial de la programación dinámica para la frase W usando el modelo integrado de la figura 6.2 (c).	121
6.4	Modificación del algoritmo de Viterbi para contemplar transiciones ε	122
6.5	Evolución de la precisión de etiquetado usando modelos BIG, LEX y BIG-BIG.	126
6.6	Evolución de la precisión y la cobertura en la detección de NP usando modelos BIG-BIG.	127
6.7	Evolución del factor F_β en función del número de palabras especializadas en el modelo contextual.	136
7.1	Ventana de representación de una etiqueta léxica	144
7.2	Ventana del manipulador de etiquetas léxicas	145

7.3	Lista de etiquetas sintácticas y ventana de manipulación	145
7.4	Editor de gramáticas	146
7.5	Resultado del análisis con parentizado a izquierdas	147
7.6	Árbol sintáctico en “modo gráfico” salida de APOLN.	148
7.7	Árbol sintáctico en “modo gráfico” completado con un SV.	149