This watermark does not appear in the registered version - http://www.clicktoconvert.com

Índice

	Índice	**xi**
	Abstract	xv
1	**Introducción**	**1**
2	**Estado actual de la Desambiguación del Sentido de las Palabras**	**5**
2.1	**El conocimiento lingüístico que utiliza el algoritmo de Yarowsky**	**6**
2.1.1	**Conocimiento lingüístico para la DSP**	**6**
2.1.2	**La codificación del conocimiento lingüístico mediante rasgos**	**9**
2.2	**El algoritmo de Yarowsky como método de DSP basado en conocimiento**	**12**
2.2.1	**Métodos basados en conocimiento que utilizan el rasgo 9**	**12**
2.2.2	**Métodos basados en conocimiento que utilizan los rasgos 6 y 10**	**13**
2.2.3	**Métodos basados en conocimiento que utilizan el rasgo 7**	**13**
2.2.4	**Métodos heurísticos (rasgos 3, 4 y 8)**	**14**
2.3	**El algoritmo de Yarowsky como método de DSP no supervisado**	**16**
2.4	**El algoritmo de Yarowsky como método de DSP semisupervisado**	**18**
2.5	**El algoritmo de Yarowsky como método de DSP supervisado**	**20**
2.5.1	**Aprendizaje automático para la clasificación**	**20**
2.5.2	**Principales metodologías de DSP supervisado**	**21**
2.6	**DSP y dominio**	**24**
2.6.1	**Enfoques para la caracterización del dominio**	**25**
2.7	**La evaluación de los sistemas de DSP**	**34**
2.7.1	**Definición de tareas de evaluación**	**35**
2.7.2	**Uso de corpus para la evaluación**	**36**
2.7.3	**La puntuación en los sistemas de evaluación**	**36**
2.7.4	**La línea de base del sentido más frecuente**	**38**
2.7.5	**Acuerdo entre anotadores (inter-annotator agreement)**	**38**
2.7.6	**La evaluación en las competiciones Senseval**	**39**
2.7.7	**Causas de los desacuerdos en el ITA**	**43**
2.7.8	**La granularidad: los grupos de sentidos en WordNet**	**45**
2.7.9	**El certamen Senseval-3**	**47**

This watermark does not appear in the registered version - http://www.clicktoconvert.com

2.8 Los sistemas de DSP en aplicaciones de PLN 48
2.8.1 Argumentos de fe 49
2.8.2 Argumentos por analogía 50
2.8.3 Argumento de las aplicaciones concretas 51

3 El Cuello de Botella de la Adquisición de Conocimiento (CBAC) 67

3.1 El CBAC en los sistemas supervisados 68
3.1.1 Adquisición automática de ejemplos de entrenamiento 68
3.1.2 Autoarranque (bootstrapping) 75

4 La granularidad de los sentidos en PLN 77

4.1 La DSP en aplicaciones de PLN de los sentidos en PLN 80

4.2 El nivel de granularidad necesario 80

5 El algoritmo de Yarowsky 83

5.1 Introducción 83

5.2 La propiedad "un sentido por discurso" 84

5.3 La propiedad "un sentido por colocación" 84

5.4 El algoritmo de 1995 85

5.5 Opciones para elegir las semillas originales 90

5.6 Uso de la propiedad de *un sentido por discurso* 91

5.7 Evaluación y resultados 92

6 La dependencia de las colocaciones con el dominio 95

6.1 La hipótesis *un sentido por colocación* 95
6.1.1 Experimentos para probar la hipótesis 96
6.1.2 Resultados 99

This watermark does not appear in the registered version - http://www.clicktoconvert.com

6.2 Relación entre la hipótesis *un sentido por colocación* y las variaciones de dominio 101
6.2.1 El corpus DSO 102
6.2.2 Definición de los experimentos 103
6.2.3 Colocaciones 103
6.2.4 Adaptación de las listas de decisión a más de dos sentidos 105
6.2.5 Experimentos dentro de un mismo corpus 105
6.2.6 Experimentos entre corpus diferentes 107
6.2.7 Utilización de ejemplos de entrenamiento y test obtenidos de los mismos documentos 109
6.2.8 La variación de género y tópico 110

7 Sistema experimental 113

7.1 Sistema utilizado para el estudio de la distribución estadística de las fuentes de dominio en los corpus de texto 113
7.1.1 Corpus utilizados 113
7.1.2 Medida para comparar fuentes de dominio 115

7.2 Sistema utilizado para la evaluación del algoritmo de Yarowsky 118

7.3 Sistema utilizado para la implementación y evaluación del nuevo algoritmo de autoarranque 121

8 La dependencia con el dominio del algoritmo de Yarowsky 125

9 La distribución estadística del dominio 131

9.1 Una medida de la ambigüedad de muestras de texto 132

9.2 Muestras de los corpus bajo estudio 133
9.2.1 Los corpus BNC y WSJ 134
9.2.2 La muestras utilizadas en los experimentos 134

9.3 Diseño de los experimentos 135

9.4 La naturaleza diferente del BNC y del corpus WSJ 137
9.4.1 Ambigüedad cruzada dentro del BNC y dentro del WSJ 137
9.4.2 Ambigüedad compuesta: el efecto de la composición de muestras 141
9.4.3 El efecto contradictorio del tamaño del corpus 144
9.4.4 Los corpus utilizados en las competiciones Senseval 145

This watermark does not appear in the registered version - http://www.clicktoconvert.com

10 El autoarranque del algoritmo de Yarowsky en corpus reales 149

10.1 La vigencia de los algoritmos de aprendizaje semisupervisados 150
10.1.1 Aprendizaje semisupervisado y variación de dominio 152
10.1.2 Otras aproximaciones posibles al CBAC 153

10.2 Una metodología de autoarranque del algoritmo de Yarowsky en corpus de texto general 154
10.2.1 La aplicación de la propiedad *un sentido por discurso* 156
10.2.2 El algoritmo de autoarranque 156
10.2.3 Trascendencia del algoritmo de autoarranque 162

11 Conclusiones y líneas de investigación futuras 165

11.1 Conclusiones 165
11.2 Líneas de investigación futures 172

12 Bibliografía 175

This watermark does not appear in the registered version - http://www.clicktoconvert.com

Abstract

A new semi-supervised Word Sense Disambiguation (WSD) bootstrapping algorithm is developed which alleviates very strongly the knowledge acquisition bottleneck problem affecting current supervised WSD algorithms. WSD algorithms are shown to perform much worse in general purpose text corpora than in journalistic text corpora, due to the stereotyped and repetitive nature of this type of corpora. The new bootstrapping algorithm is shown to reach the performance of supervised algorithms in general text, i.e. non-journalistic text corpora, and potentially to outperform them due to the use of a binary decision methodology jointly with the *one-sense-per-discourse* (OSPD) property of natural language, and to the flexibility of bootstrapping algorithms, which is able to tackle the domain fluctuations present in general text corpora.

This watermark does not appear in the registered version - http://www.clicktoconvert.com