Resolución de la Ambigüedad Semántica de las palabras mediante Modelos de Probabilidad de Máxima Entropía

Armando Suárez Cueto
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Alicante
Aptdo. Correos 99 E-03080 Alicante
armando@dlsi.ua.es

Resumen: Tesis doctoral en Informática realizada por Armando Suárez Cueto bajo la dirección de los doctores Manuel Palomar Sanz, de la Universidad de Alicante, y German Rigau Claramunt, de la Universidad del País Vasco. El acto de defensa de la tesis tuvo lugar el 28 de junio de 2004 ante el tribunal formado por los doctores Lluís Padró Cirera (Univ. Politécnica de Cataluña) , Andrés Montoyo Guijarro (Univ. de Alicante), Eneko Agirre Bengoa (Univ. del País Vasco), Alfonso Ureña López (Univ. de Jaén) y Bernardo Magnini (Istituto Trentino di Cultura) . La calificación obtenida fue Sobresaliente Cum Laude por unanimidad.
Palabras clave: desambiguación léxica, máxima entropía, basado en corpus, bootstrapping, co-training

Abstract: PhD Thesis in Computer Science written by Armando Suárez Cueto under the supervision of Dr. Manuel Palomar Sanz, (Univ. of Alicante), and German Rigau Claramunt (Univ. of Basque Country). The author was examined in June 28^{th}, 2004 by the commitee formed by Dr. Lluís Padró Cirera (Politechnic University of Cataluña) , Andrés Montoyo Guijarro (Univ. of Alicante), Eneko Agirre Bengoa(Univ. of Basque Country), Alfonso Ureña López (Univ. of Jaén) y Bernardo Magnini (Istituto Trentino di Cultura). The grade obtained was *Sobresaliente Cum Laude*
Keywords: Word Sense Disambiguation, Maximum Entropy, corpus-based, bootstrapping, co-training

1. Introducción

La *resolución de la ambigüedad semántica de las palabras* (de ahora en adelante WSD, de *Word Sense Disambiguation*) es una tarea que se encuadra dentro de un conjunto más amplio de técnicas llamado *procesamiento del lenguaje natural* (PLN) que, básicamente, trata los fenómenos lingüísticos de toda índole de forma mecanizada mediante ordenadores. Concretamente, WSD trata de la asignación automática de sentidos a las palabras de un texto.

WSD es lo suficientemente compleja como para requerir la concurrencia de múltiples aproximaciones, métodos, heurísticas, etc. Dentro de esta complejidad debemos mencionar la fuerte dependencia de la fuente de los textos destinados al aprendizaje y la desambiguación, además de las características propias de cada sentido o concepto que puede obligar al refinamiento del proceso.

Parece comúnmente aceptada la clasificación de las aproximaciones a WSD en dos categorías muy generales: **métodos basados en el conocimiento** (*knowledge-based methods*) y **métodos basados en corpus** (*corpus-based methods*). Los primeros hacen uso del conocimiento adquirido en forma de diccionarios, tesauros, lexicones, ontologías, etc. Podemos decir que este conocimiento es preexistente al proceso de desambiguación y, en la mayoría de los casos, adquirido de forma manual. Los segundos extraen el conocimiento de grandes cantidades de ejemplos (de un corpus) mediante métodos estadísticos y aprendizaje automático. Cuando esos ejemplos están anotados previamente con la etiqueta correcta (el sentido, en nuestro caso), se dice que son métodos de **aprendizaje supervisado**, y **no supervisado** cuando no existe tal anotación. Dado el gran número de métodos y soluciones propuestos actualmente, la clasificación suele simplificarse y se habla de métodos supervisados o no, esto es, únicamente si necesitan de un corpus anotado o no.

De entre los métodos supervisados, nosotros vamos a proponer los **modelos de probabilidad condicional de máxima entropía** (MME), usados también en muchas de las tareas de PLN. ME se encuadra en los **métodos estocásticos de aprendizaje automático**, grupo en el que también se citan *naive* Bayes, *expectation maximization*, *log-linear models* y modelos ocultos de Markov. También se le puede denominar como un **método de aprendizaje inductivo basado en vectores de atributos**.

A partir de la construcción de un sistema de WSD supervisado basado en los MME, el objetivo es mejorar su rendimiento. Así, esta Tesis Doctoral tiene dos objetivos fundamentales:

Desarrollo y selección de atributos. Básicamente, cuál es la información necesaria y adecuada para aprender. Cada conjunto de ejemplos tiene su propias características, cada palabra también, y hasta cada sentido. Detectar estas características y aprovecharlas para el aprendizaje requiere, en muchos casos, un análisis de los corpus de aprendizaje previo a la construcción de los clasificadores.

Construcción de un sistema de alta precisión. Todavía no existe el sistema aceptablemente fiable para WSD. Nuestro sistema basado en ME, aún cuando es competitivo, puede ser mejorado. Uno de los problemas de WSD es su escasa efectividad para ser usado en otras tareas como búsqueda de respuestas. Una forma de hacerlo, aunque no sea la ideal, es primar la precisión en detrimento de la cobertura, esto es, clasificar sólo aquellas palabras para las que tenemos una alta confianza en que la etiqueta a asignar es la correcta.

Proponemos un algoritmo iterativo, una adaptación de un método incremental en el que se usa como núcleo nuestro propio sistema de WSD, con el objetivo de asegurar esa alta precisión en la clasificación.

2. Contribuciones de esta Tesis Doctoral

Tras abordar los objetivos antes mencionados, estos son, en nuestra opinión, los méritos destacables de todo este trabajo de investigación.

Los resultados del trabajo que originó la redacción de este documento se derivan de la implementación de un sistema de WSD supervisado basado en los modelos de máxima entropía. Así, se dispone de un software de propósito general que en la actualidad está siendo utilizado, aparte de en WSD, en la construcción de sistemas de reconocimiento de entidades y de análisis sintáctico parcial.

Una vez que ya disponíamos de nuestro propio sistema, la elección de la información que debíamos suministrarle para realizar un correcto aprendizaje y clasificación fue nuestra siguiente tarea. El estudio de dichas fuentes de información reveló la disparidad de criterios de selección que se puede llegar a dar, dependiendo del conjunto de ejemplos de aprendizaje y clasificación de la palabra a desambiguar.

También ofrecemos otra alternativa a la definición de atributos mediante una compactación de los mismos, reduciendo drásticamente la cantidad de proceso necesario en el aprendizaje y clasificación, al tiempo que se observa una degradación mínima de los resultados.

Se prueba empíricamente que la selección de atributos puede ser beneficiosa tanto para palabras como para categorías gramaticales (nombres, verbos, adjetivos y adverbios). No obstante, nuevamente la dependencia de los datos de entrenamiento se muestra como un obstáculo difícil de salvar.

La combinación por votación simple de nuestro sistema con un método no supervisado obtiene resultados tan buenos como otros más sofisticados. Esto nos lleva a pensar que, posiblemente, el problema de la resolución de la ambigüedad semántica de las palabras no radica tanto en los métodos utilizados como en los datos que manejamos, los ejemplos de entrenamiento y los textos donde se aplican, finalmente, para su clasificación.

El algoritmo Reentrenamiento, alternativo a *co-training*, un conocido método de *bootstrapping*, que hemos desarrollado y que presentamos en esta Tesis Doctoral, aplicado a WSD es capaz de mantener altos niveles de precisión en la clasificación. La novedad de este algoritmo radica en la división del problema de desambiguar una palabra de n sentidos en n subproblemas binarios a los que se aplican varios filtros para asegurar la certeza en la clasificación. La comparación con el algoritmo original de coentrenamiento es ventajosa en cuanto la degradación de la precisión en el proceso iterativo puede detectarse y eliminarse, aún a costa de no clasificar todo el conjunto no etiquetado.