Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 123-124 recibido 27-09-10 revisado 22-02-11 aceptado 24-02-11

An Approach to Word Sense Disambiguation based on Semantic Classes and Machine Learning*

Una Aproximación a la Desambiguación del Sentido de las Palabras basada en Clases Semánticas y Aprendizaje Automático

Rubén Izquierdo
Departamento Lenguajes y Sistemas Informáticos
Universidad de Alicante
Aptdo. Correos 99 E-03080 Alicante
ruben@dlsi.ua.es

Resumen: Tesis doctoral en Informática realizada por Rubén Izquierdo en la Universidad de Alicante (UA) bajo la dirección del Dr. Armando Suárez Cueto (UA) y del Dr. German Rigau Claramunt (EHU/UPV). El acto de defensa de la tesis tuvo lugar en Alicante el 17 de Septiembre de 2010 ante el tribunal formado por los doctores Manuel Palomar (UA), Paloma Moreda (UA), María Teresa Martín (UJA), Lluís Padró (UPC) e Irene Castellón (UB). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: Word Sense Disambiguation, Semantic Classes, Machine Learning, Semantic Classifiers

Abstract: Ph.D Thesis in Computer Science, specifically in the field of Computational Linguistics, written by Rubén Izquierdo at the University of Alicante (UA), under the supervision of Dr. Armando Suárez Cueto (UA) and Dr. German Rigau Claramunt (EHU/UPV). The author was examined on September 17^{th} 2010, by a panel formed by Dr. Manuel Palomar (UA), Dr. Paloma Moreda (UA), Dr. María Teresa Martín (UJA), Dr. Lluís Padró (UPC) and Dr. Irene Castellón (UB). The grade obtained was *Sobresaliente Cum Laude*.
Keywords: Desambiguación del Sentido de las Palabras, Clases Semánticas, Aprendizaje Automático, Clasificadores Semánticos

1. Introduction

Natural language is extremely complex and ambiguous. One of the main ambiguities that has received more attention is the lexical ambiguity.Words can have more than one meaning, and therefore texts containing these polysemous words are ambiguous. Word Sense Disambiguation (WSD) is an enabling task in Natural Language Processing (NLP) that tries to solve lexical ambiguity, or in other words, determine which is the proper sense of a polysemous word depending on its surrounding context.

From a Machine Learning perspective, WSD can be seen as a classification problem. From this point-of-view, a machine learning algorithm learns classification models from corpora annotated with word senses. The most used repository of senses has been WordNet. This kind of supervised approaches based on word senses have shown to reach a very high performance on SensEval/SemEval competitions. However, since SensEval–3 there is a general consensus about that traditional word sense–based approaches has reached a plateau very difficult to overcome. One of the possible reasons is that WordNet, as a repository of senses, provides too fine–grained senses, with very little differences between senses of the same word. This kind of differences are too subtle to be captured by a machine learning system. This problem is even more acute when the amount of training examples for each classifier is low for ensuring a robust generalization. Furthermore, the existing sense annotated resources are not very large. This phenomenon is usua-

* Este trabajo ha sido co–financiado por el Ministerio de Ciencia e Innovación (proyecto TIN2009-13391-C04-01), y la Conselleria de Educación de la Generalitat Valenciana (proyectos PROMETEO/2009/119 y ACOMP/2011/001).

ISSN 1135-5948

lly known as the *knowledge acquisition bottleneck*.

This research focuses on the use of semantic classes instead of traditional word senses. A semantic class is a concept that groups a set of sub–concepts that present lexical coherence and share certain properties. Examples of semantic classes are BUILDING, VEHICLE or FOOD. The use of semantic classes provides several advantages compared with the use of senses: reduction of polysemy, higher level of abstraction for WSD, increment of the amount of training data available for each classifier and also the possibility to disambiguate words not contained in training data. We have developed a semantic class based system for WSD, using Support Vector Machines (SVM) as a learning method, SemCor as a training corpus and two different sets of features. We also tests different sets of semantic classes: *Basic Level Concepts* (BLC), WordNet Domains, SUMO classes and SuperSenses. Semantic classes has been used for building both the classifiers and two types of semantic features.

2. Contributions of this thesis

The main contributions of this thesis are two: an automatic method to extract BLC from WordNet, and a semantic class based WSD system. Both the resources and the NLP tool are available for research[1].

2.1. Automatic extraction of BLC from WordNet

We have developed a method that automatically extracts a set of *Basic Level Concepts* from WordNet. The process follows a bottom–up approach through the hypernymy chains of WordNet. For each synset[2], the algorithm selects as its BLC the first local maximum (according to a certain criteria) in the hypernymy chain. To compute this local maximum different characteristics can be considered. For instance, the number of relations of each synset or the frequency of the synset (computed as the sum of frequencies of the words within the synset). This criteria is a parameter of the algorithm. Another parameter is the minimum number of concepts that each candidate must subsume to be an actual BLC. This is the second parameter of the method.

Combining different values for the two parameters, different sets of BLC can be automatically extracted from WordNet. This is very relevant since the method is flexible enough to derive different sets of BLC at a different abstraction level. Possibly, a particular NLP application requires a particular level of abstraction for representing its semantic information. Predefined set of classes (like WordNet Domains, SUMO or SuperSenses) have an static level of abstraction. Our approach provides a good alternative to these predefined semantic classes.

2.2. A semantic class based WSD system

Our WSD system uses a machine learning algorithm (SVM) and SemCor as training data, two sets of features, and several semantic class repositories: BLC, WordNet Domains, SUMO and SuperSenses. Each of the classifiers assigns the proper semantic class to each ambiguous word. We also use the semantic classes to build two types of semantic features to help the characterization of training and testing examples.

We have tested the WSD system using the evaluation corpora from SensEval–2, SenseEval–3 and SemEval–1. Moreover, we have participated in SemEval–1 and SemEval–2 with our semantic class based system reaching the 5^{th} best position in both cases. Finally, we compare our system with the participants at SensEval–2 and SensEval–3 at a sense and a semantic class levels. In both cases, the system outputs are transformed to work at the same abstraction level (sense or semantic class) for a fair evaluation.

This research empirically demonstrates that we can select a medium level of abstraction with semantic classes and reach a high performance (near to very abstract semantic classes) maintaining the discriminatory power of meanings. Moreover, the semantic class classifiers are more robust and more independent to domain shifting than sense classifiers. Finally we have shown that the amount of training data available for each classifier is increased when using the semantic class approach, alleviating slightly the lack of semantically annotated corpora.

[1] `http://www.dlsi.ua.es/~ruben`

[2] The method is only defined for nouns and verbs.