Supervised Word Sense Disambiguation: Facing Current Challenges

David Martínez Iraola
Grupo IXA - Universidad del País Vasco
Donostia, 20018
davidm@si.ehu.es

Resumen: Tesis presentada en diciembre de 2004 por la Universidad del País Vasco bajo la supervisión del Dr. Eneko Agirre Bengoa.
Palabras clave: desambiguación léxica supervisada, tipos de atributos, adquisición automática de ejemplos, portabilidad

Abstract: Dissertation presented in December of 2004 at the University of the Basque Country under the supervision of Dr. Eneko Agirre Bengoa.
Keywords: supervised word sense disambiguation, feature types, automatic acquisition of examples, portability

1. Goals of the thesis

The main goal of this thesis is to study supervised Word Sense Disambiguation (WSD), and propose ways to overcome its limitations to be applied for NLP. Thus, we started describing the main challenges facing WSD systems. These factors limit the performance of these systems to around 70 % accuracy in the literature[1].

- *The definition of the problem is wrong.* Some authors claim that defining the meaning of a word as a discrete list of senses is hopeless, as it does not model correctly its behavior.
- *Sense inventory and granularity.* The task depends on the applied sense inventory, which has to be chosen adequately in order to build a flexible and comparable system.
- *ML algorithms are not adequately applied to the problem.* Methods coming from the Machine Learning (ML) community have been widely applied to the WSD problem. However, the comparative results show that even the most sophisticated methods have not been able to make a qualitative jump in performance.
- *The feature sets used to model the language are too limited.* Traditionally simple feature sets consisting in bigrams, trigrams, and "bags of words" have been used to model the contexts of the target words. But in order to be robust, the ML methods should rely in as much information from the texts as possible.
- *The sparse data problem.* In NLP most of the events occur rarely, even when large quantities of data are available. This problem is specially noticeable in WSD, where hand-tagged data is difficult to obtain.
- *Need of extra training data.* Existing hand-tagged corpora is not enough for current state-of-the-art systems. Hand-tagged data is difficult and costly to obtain, and methods to obtain data automatically have not reached the same quality of hand-tagged data so far.
- *Portability.* The porting of the WSD systems to be tested on a different corpora than the one used for training also presents difficulties. Previous work (Escudero, Màrquez, y Rigau, 2000) has shown that there is a loss of performance when training on one corpora and testing on another.

[1] We refer to all-words systems with fine-grained sense distinctions.

2. Main contributions

We explored two main hypotheses in this dissertation:

1. The use of richer features (syntactic, semantic, or domain features) can provide relevant information of the contexts, and it should improve significantly baseline methods that are trained on classic features.
2. The automatic acquisition of examples by means of WordNet relatives can alleviate the knowledge acquisition bottleneck, and improve over other unsupervised (or minimally supervised) approaches.

All in all, we think that our main contributions on these initial hypotheses are the following:

2.1. Syntactic features

We explored the contribution of an extensive set of syntactic features to WSD. The study included two different ML methods (Decision Lists (DL) and AdaBoost (AB)), and a precision/coverage trade-off system using these feature types. The results show that basic and syntactic features contain complementary information, and that they are useful for WSD. This is specially noticeable for the AB algorithm in the standard setting, and for DLs when applying the precision/coverage trade-off.

2.2. Semantic features

We applied two approaches to study the contribution of semantic features using WordNet and the Semcor corpus. On the one hand, we constructed new feature types based on the synsets surrounding the target word, the hypernyms of these synsets, and also their semantic files. On the other hand, we learned different models of selectional preferences for verbs, using the relations extracted from the Semcor corpus by Minipar. Our main conclusions were that the "bag-of-synsets" approach does not benefit much from the WordNet hierarchy. Instead, selectional preference acquisition offers promising results.

2.3. Automatic acquisition of examples

We evaluated up to which point we can automatically acquire examples for word senses and train WSD systems on them. The method we applied is based on the monosemous relatives of the target words (Leacock, Chodorow, y Miller, 1998), and we studied some parameters that affect the quality of the acquired corpus, such as the distribution of the number of training instances per sense (bias). We built three systems with different supervision requirements

We showed that the fully supervised system combining our web corpus with the examples in Semcor improves over the same system trained on Semcor alone (specially for nouns with few examples in Semcor). Regarding minimally supervised and fully unsupervised systems, we demonstrated that they perform well better than the other systems of the same category presented in the Senseval-2 lexical-sample competition.

2.4. Genre/topic shift

We measured the strength of the "one sense per collocation" hypothesis (Yarowsky, 1993) using different corpora for training and testing. Our goal was to measure the importance of introducing examples from different sources in WSD performance. We focused on the domain/genre factor, and performed our experiments in the DSO corpus, which comprises sentences extracted from two different corpora: the balanced BC, and the WSJ corpus containing press articles.

Our experiments show that the one sense per collocation hypothesis is weaker for fine-grained word sense distinctions, and that it does hold across corpora, but that collocations vary from one corpus to other, following genre and topic variations. We showed that when two independent corpora share a related genre/topic, the WSD results are better.

Bibliografía

Escudero, G., L. Màrquez, y G. Rigau. 2000. On the Portability and Tuning of Supervised Word Sense Disambiguation Systems. En *Joint SIGDAT Conference EMNLP/VLC*, Hong Kong, China.

Leacock, C., M. Chodorow, y G. A. Miller. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. En *Computational Linguistics*, volumen 24.

Yarowsky, D. 1993. One Sense per Collocation. En *Proceedings of the 5th DARPA Speech and Natural Language Workshop.*