

Using Sense Clustering for the Disambiguation of Words

Henry Anaya-Sánchez, Aurora Pons-Porrata, and Rafael Berlanga-Llavori

Abstract—Clustering methods have been extensively used in the solution of many Information Processing tasks in order to capture unknown object categories. This paper presents an approach to Word Sense Disambiguation based on clustering. The underlying idea is that the clustering of word senses provides a useful way to discover semantically related senses. We evaluate our proposal regarding both fine- and coarse-grained disambiguation. Experimental results over Senseval-3 all-words, SemCor 2.0 and SemEval-2007 corpora are presented. Promising values of precision and recall are obtained.

Index Terms—Word sense disambiguation, clustering.

I. INTRODUCTION

THE task of Word Sense Disambiguation (WSD) consists of selecting the appropriate sense for a particular contextual occurrence of a polysemous word. This task can be specialized according to the sense definitions. For instance, word sense induction refers to the process of discovering different senses of an ambiguous word without prior information about the inventory of senses [21]. On the other hand, there are two major approaches for the disambiguation when predetermined sense definitions are provided: data-driven (or corpus-based) and knowledge-driven WSD. Data-driven methods are supervised because they require a learning model built from hand-tagged samples to disambiguate words. Instead, knowledge-driven methods exploit word relationships provided by a background knowledge source, avoiding thus the use of samples. Currently, lexical resources like WordNet [14] constitute the referred source in most cases.

WSD can be seen as a categorization problem consisting of assigning a category label (predefined sense) to each word. In this way, data-driven approaches can be regarded as supervised categorization methods, whereas knowledge-driven ones as unsupervised.

Clustering is one of the most accepted unsupervised categorization methods. It has been explicitly used in WSD for two main purposes. The first one consists of clustering textual contexts to represent different senses in corpus-driven WSD (e.g. [17]) and to induce word senses (e.g. [18], [3]). The other

purpose has been the clustering of fine-grained word senses into coarse-grained ones for reducing the polysemy degree of words (e.g. [13], [1]). However, clustering has not been used as categorization method for WSD, that is, as a way to identify sets of word senses that are semantically related.

In this paper, we present a knowledge-driven approach to WSD based on sense clustering. Basically, our proposal uses sense clustering to capture the reflected cohesion among the words of a textual unit. More specifically, starting from an initial clustering of all the possible senses for a textual unit, clusters of senses with a high cohesion w.r.t the textual context are selected. The senses belonging to the selected clusters are grouped and selected again until all words are disambiguated.

The rest of the paper is organized as follows. First, Section II presents our proposal for the disambiguation of words. Section III describes some experiments carried out over Senseval-3 all-words, SemCor 2.0 and SemEval coarse-grained corpora. Finally, Section IV is devoted to offer some considerations and future work as conclusions.

II. WORD SENSE CLUSTERING

In this section we address the problem of disambiguating a finite set of words $W = \{w_1, \dots, w_n\}$ w.r.t its textual context T . The underlying idea of sense clustering is that meaningful word senses must be associated by means of a certain complex relation, which is non-relevant for our purposes because we are only interested in the senses it links. Hence, we propose to identify cohesive groups of senses which are assumed to represent different meanings for the set of words W . Finally, those clusters that fit in with the context T contain the suitable senses.

Algorithm 1 shows the general steps of our proposal. In the algorithm, *clustering* represents the basic clustering algorithm which groups word senses and, *filter* denotes the filtering process which selects the clusters that allow the disambiguation of words in W . The filtering process is described in Algorithm 2. Next paragraphs describe in detail the whole process.

a) Topic signatures: In our approach word senses are represented as topic signatures [12]. Thus, for each word sense s we define a vector $\langle t_1 : \sigma_1, \dots, t_m : \sigma_m \rangle$, where each t_i is a WordNet term highly correlated to s with an association weight σ_i . The set of signature terms for a word sense includes all its WordNet hyponyms, its directly related terms (including coordinated terms) and their filtered and lemmatized glosses.

Manuscript received November 4, 2008. Manuscript accepted for publication August 28, 2009.

Henry Anaya-Sánchez and Aurora Pons-Porrata are with Center for Pattern Recognition and Data Mining, Universidad de Oriente, Santiago de Cuba, Cuba (henry@cepramid.co.cu, aurora@cepramid.co.cu).

Rafael Berlanga-Llavori is with Department of Languages and Computer Systems, Universitat Jaume I, Castelló, Spain (berlanga@lsi.uji.es).

