Word Independent Context Pair Classification Model for Word Sense Disambiguation

Cheng Niu, Wei Li, Rohini K. Srihari, and Huifeng Li
Cymfony Inc.
600 Essjay Road, Williamsville, NY 14221, USA.
{cniu, wei, rohini,hli}@cymfony.com

Abstract

Traditionally, word sense disambiguation (WSD) involves a different context classification model for each individual word. This paper presents a weakly supervised learning approach to WSD based on learning a word independent context pair classification model. Statistical models are not trained for classifying the word contexts, but for classifying a pair of contexts, i.e. determining if a pair of contexts of the same ambiguous word refers to the same or different senses. Using this approach, annotated corpus of a target word *A* can be explored to disambiguate senses of a different word *B*. Hence, only a limited amount of existing annotated corpus is required in order to disambiguate the entire vocabulary. In this research, maximum entropy modeling is used to train the word independent context pair classification model. Then based on the context pair classification results, clustering is performed on word mentions extracted from a large raw corpus. The resulting context clusters are mapped onto the external thesaurus WordNet. This approach shows great flexibility to efficiently integrate heterogeneous knowledge sources, e.g. trigger words and parsing structures. Based on Senseval-3 Lexical Sample standards, this approach achieves state-of-the-art performance in the unsupervised learning category, and performs comparably with the supervised Naïve Bayes system.

1 Introduction

Word Sense Disambiguation (WSD) is one of the central problems in Natural Language Processing. The difficulty of this task lies in the fact that context features and the corresponding statistical distribution are different for each individual word. Traditionally, WSD involves training the context classification models for each ambiguous word. (Gale et al. 1992) uses the Naïve Bayes method for context classification which requires a manually annotated corpus for each ambiguous word. This causes a serious *Knowledge Bottleneck*. The bottleneck is particularly serious when considering the domain dependency of word senses. To overcome the *Knowledge Bottleneck*, unsupervised or weakly supervised learning approaches have been proposed. These include the bootstrapping approach (Yarowsky 1995) and the context clustering approach (Schütze 1998).

The above unsupervised or weakly supervised learning approaches are less subject to the *Knowledge Bottleneck*. For example, (Yarowsky 1995) only requires sense number and a few seeds for each sense of an ambiguous word (hereafter called *keyword*). (Schütze 1998) may only need minimal annotation to map the resulting context clusters onto external thesaurus for benchmarking and application-related purposes. Both methods are based on trigger words only.

This paper presents a novel approach based on learning word-independent context pair classification model. This idea may be traced back to (Schütze 1998) where context clusters based on generic Euclidean distance are regarded as distinct word senses. Different from (Schütze 1998), we observe that generic context clusters may not always correspond to distinct word senses. Therefore, we used supervised machine learning to model the relationships between the context distinctness and the sense distinctness.

Although supervised machine learning is used for the context pair classification model, our overall system belongs to the weakly supervised category because the learned context pair classification

Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL),
pages 33–39, Ann Arbor, June 2005.