A Corpus-based Computational Approach to Sense Clustering: A Pilot Study of Semi-automatic Identification of Words with Semantic Change in a Chinese Newspaper Archive

Zhao-Ming Gao  Hong-Min Chu
National Taiwan University

Sense discrimination is important to linguistics, applied linguistics, and natural language processing (NLP). In historical linguistics and lexical semantics, for example, researchers are interested in investigating words which have undergone semantic changes. However, the identification of such words is very laborious and time-consuming. It would be highly desirable if there is a tool that can help linguists semi-automatically identify words which have undergone semantic changes. As there are very few sense-tagged corpora available, supervised machine learning approaches are untenable.

This paper presents a pilot study of semi-automatic identification of words with semantic change using unsupervised machine learning, which does not require a training corpus. We use sense clustering based on a small Chinese newspaper corpus. Our approach is based on the distributional hypothesis proposed by John Firth six decades ago which states that words that appear in the same contexts share semantic meaning. A modern counterpart of the distributional hypothesis is Yarowsky (1993)’s generalization which states that there is one sense per collocation. In other words, the same word with a different meaning has a different collocation.

As with previous studies on computational approaches to sense discrimination, the context of a word can be represented by a vector. Our proposed method to model context is similar to Levey and Goldberg (2014), which includes the words which bear syntactic dependency relations with the target word as well as their parts-of-speech. Since there is no delimiter between words in Chinese, word segmentation and Chinese part-of-speech tagging program from Academia Sinica is used. To derive the dependency relations in a sentence, we train a Chinese dependency Chinese parser from Sinica Chinese Treebank using SVM. The Chinese dependency parser can automatically analyze dependency relations such as subject/object, verb/noun, and modifier/noun of a sentence. After word segmentation, part-of-speech tagging, and dependency parsing, we then use the toolkit word2vec to cluster instances with the same meanings. In our approach, the target word in each instance is represented as a distinct lexis using a different special symbol. In order to evaluate the performance of the proposed method, we manually annotate the meaning of 10 ambiguous words in 4500 example sentences.

Our preliminary result indicates that the F measure is 73.5% (with average purity 0.949 and average random index 0.61), suggesting that the proposed approach has the potential of semi-automatically finding some words with lexical semantic changes. We will discuss other applications of such tool than historical linguistics and lexical semantic research.