Clustering of word senses

Eneko Agirre

University of the Basque Country, Donostia 20.080, Spain,
eneko@si.ehu.es,
WWW home page: http://ixa.si.ehu.es

WordNet does not provide any information about the relation among the word senses of a given word, that is, the word senses are given as a flat list. Some dictionaries provide an abstraction hierarchy, and previous work has tried to find systematic polysemy relations [3] using the hierarchies in WordNet.

In [1, 2] we apply distributional similarity methods to word senses, in order to build hierarchical clusters for the word senses of a word. The method uses the information in WordNet (monosemous relatives) in order to collect examples of word senses from the web. In the absence of hand-tagged data, those examples constitute the context of each word sense. The contexts are modeled into vectors using different weighting functions, e.g. χ^2 or $tf.idf$. The similarity between the word senses can thus be obtained using any similarity function, e.g. the cosine. Once we have a similarity matrix for the word senses of a given word, clustering techniques are applied in order to obtain a hierarchical cluster.

The evaluation shows that our hierarchical clusters are able to approximate the manual sense groupings for the nouns in Senseval 2 with purity values of 84%, comparing favorably to using directly the hand-tagged data available in Senseval 2 (purity of 80%). The results are better than those attained by other techniques like confusion matrixes from Senseval 2 participating systems or multilingual similarity.

The primary goal of our work is to tackle the fine-grainedness and lack of structure of WordNet word senses, and we will be using the clusters to improve Word Sense Disambiguation results. We plan to make this resource publicly available for all WordNet nominal word senses, and we expect for the similarity measure to be valuable in better acquiring the explicit relations among WordNet word senses, including specialization, systematic polysemy and metaphorical relations.

References

1. E. Agirre and O. Lopez de Lacalle. Clustering wordnet word senses. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP'03)*. Borovets, Bulgary. 2003.
2. E. Agirre, E. Alfonseca and O. Lopez de Lacalle Approximating hierarchy-based similarity for WordNet nominal synsets using Topic Signatures. In *Proceedings of the 2nd Global WordNet Conference*. Brno, Czech Republic. 2004
3. W. Peters and I. Peters. Lexicalized systematic polysemy in WordNet. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece. 2000.