

Sense Proximity versus Sense Relations

Julio Gonzalo

Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
E.T.S.I Industriales, Ciudad Universitaria s/n, 28040 Madrid, SPAIN
Email: julio@lsi.uned.es WWW: <http://sensei.lsi.uned.es/~julio/>

It has been widely assumed that sense distinctions in WordNet are often too fine-grained for applications such as Machine Translation, Information Retrieval, Text Classification, Document clustering, Question Answering, etc. This has led to a number of studies in sense clustering, i.e., collapsing sense distinctions in WordNet that can be ignored for most practical applications [1,5,6]. At the UNED NLP group, we have also conducted a few experiments in sense clustering with the goal of improving WordNet for Information Retrieval and related applications [4,3,2].

Our experiments led us to the conclusion that annotating WordNet with a typology of polysemy relations is more helpful than forming sense clusters based on a notion of sense proximity. The reason is that sense proximity depends on the application, and in many cases can be derived from the type of relation between two senses. In the case of metaphors, senses often belong to different semantic fields, and therefore a metaphor can be a relevant distinction for Information Retrieval or Question & Answer systems. For Machine Translation applications, however, the metaphoric sense extensions might be kept across languages, and therefore the distinction might not be necessary to achieve a proper translation.

In the panel presentation, we will summarize the experiments that led us to hold this position:

- In [2] we compared two clustering criteria: the first criterion, meant for Information Retrieval applications, consists of grouping senses that tend to co-occur in Semcor documents. The second criterion, inspired by [7], groups senses that tend to receive the same translation in several target languages via the EuroWordNet Interlingual Index (*parallel polysemy*). The overlapping of both criteria was between 55% and 60%, which reveals a correlation between both criteria but leaves doubts about the usefulness of the clusters. However, a classification of the sense groupings according to the type of polysemy relation clarifies the data: all homonym and metaphor pairs satisfying the parallel polysemy criterion did not satisfy the co-occurrence criterion; all generalization/specialization pairs did satisfy the co-occurrence criterion; finally, metonymy pairs were evenly distributed between valid and invalid co-occurrence clusters. Further inspection revealed that the type of metonymic relation could be used to predict sense clusters for Information Retrieval.
- In [3] we applied Resnik & Yarowsky measure to evaluate the Senseval-2 WordNet subset for sense granularity. We found that the average proximity was similar to the Senseval-1 sense inventory (Hector), questioning the idea that WordNet sense distinctions are finer than in other resources built by lexicographers. We also found that Resnik & Yarowsky proximity measure provides valuable information, but should

be complemented with information about polysemy relations. There are, for instance, a significant fraction of homonyms that receive a non-null proximity measure, and the average proximity for metaphoric pairs is higher than would be expected for sense pairs belonging to different semantic fields. We believe that the classification of such pairs as homonyms is more valuable and has more predictive power than the quantitative measure of proximity.

In WordNet, the different senses of a word can be implicitly connected through the semantic relations between synsets. But these connections are too vague to understand the relations holding between senses: for instance, it is hard to decide when two senses of a word are homonyms, an information that is essential for Language Engineering applications, and can be found in other, more conventional lexicographic resources. We believe that, to achieve the full potential of wordnets as a de facto standard for lexical resources in computational applications, the relations between senses of polysemous words should be explicitly annotated. In the panel discussion, we will briefly discuss a proposal for a simple typology of polysemy relations, and the exploratory annotation of the senses for a thousand nouns in WordNet using this typology.

References

1. E. Agirre and O. Lopez de Lacalle. Clustering Wordnet word senses. In *Proceedings RANLP 2003*, 2003.
2. I. Chugur, J. Gonzalo, and F. Verdejo. A study of sense clustering criteria for information retrieval applications. In *Proceedings of Ontolex 2000*, 2000.
3. I. Chugur, J. Gonzalo, and F. Verdejo. Polysemy and sense proximity in the senseval-2 test suite. In *Proceedings of the ACL-2002 Workshop on "Word Sense Disambiguation: recent successes and future directions"*, 2002.
4. Julio Gonzalo, Irina Chugur, and Felisa Verdejo. Sense clustering for information retrieval: evidence from Semcor and the EWN InterLingual Index. In *Proceedings of the ACL'00 Workshop on Word Senses and Multilinguality*, 2000.
5. R. Mihalcea and D. Moldovan. Automatic generation of a coarse grained wordnet. In *Proceedings of NAACL Workshop on WordNet and Other LEXical Resources*, 2001.
6. W. Peters and I. Peters. Automatic sense clustering in EuroWordnet. In *Proceedings of LREC'2000*, 2000.
7. P. Resnik and D. Yarowsky. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 1999.