Using textual association measures and minimum edit distance to discover morphological relations

Marco Baroni* Johannes Matiasek* Harald Trost†

Collocations Workshop
July 23, 2002

We present an unsupervised, knowledge-free algorithm that takes an unannotated corpus as its input, and returns a ranked list of probable morphologically related pairs as its output. For example, when run with the Brown corpus as its input, our system returns a list with pairs such as *price/prices* and *reduce/reduced* at the top.

The algorithm is based on the simple idea that a combination of formal and semantic cues can be exploited to identify morphologically related pairs.

Orthographical similarity is computed using the well-known minimum edit distance measure.

Following a hint from [Brown et al. 1990], we measure semantic similarity by applying one of three textual association measures (AMs) ([Evert 2001]) to long-distance non-directional bigrams. The three AMs are raw co-occurrence frequency, mutual information and log-likelihood ratio.

Experiments with German and English input indicate that the algorithm is able to identify a variety of valid morphological relations. We report about these results in [Baroni et al. 2002]. Here, we concentrate on the comparison of the performance achieved with each of the AMs.

In the task at hand, log-likelihood ratios do not out-perform raw co-occurrence counts. Mutual information produces better results than the other two measures, but only in German.

A more detailed analysis of the top pairs indicates that there is large overlap between the results obtained with co-occurrence frequency and the results obtained with log-likelihood ratio, whereas the pairs discovered by mutual information (coming from the low end of the frequency spectrum) are different from the ones discovered using either of the other two AMs.

Thus, in future research we plan to combine evidence on semantic relatedness found using co-occurrence frequency (or log-likelihood ratio) and evidence found using mutual information.

References

[Baroni et al. 2002] Baroni M., Matiasek J., Trost H.: Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. Proceedings of the Workshop on Morphological and Phonological Learning of ACL-2002: 48-57; http://www.ai.univie.ac.at/~marco/.

[Brown et al. 1990] Brown P.F., Della Pietra V.J., DeSouza P.V., Lai J.C., Mercer R.L.: Class-based n-gram models of natural language, Computational Linguistics, 16(4):467-479.

[Evert 2001] Evert, S.: On lexical association measures, http://www.collocations.de/EK/.

*Austrian Research Institute for Artificial Intelligence, Vienna
†Department of Medical Cybernetics and Artificial Intelligence, University of Vienna