Neighborhood lists and word proximities generated from a COALS semantic space built over a corpus of Polish

Marcin Tatjewski, Insitute of Computer Science, Polish Academy of Sciences

in collaboration with

Mirosław Bańko, University of Warsaw
Joanna Rączaszek-Leonardi, Faculty of Psychology, University of Warsaw

One way of representing word semantics in natural languages is based on the assumption that words similar in meaning occur in similar contexts. Now that large language corpora are available, the idea of representing a word's meaning as a class of the contexts in which it occurs has recently been implemented in a number of semantic models in which words are represented as vectors in high-dimensional spaces. The values of these vectors are defined through the contexts in which the corresponding words appear in a given corpus, and the semantic proximity of words is expressed by the Euclidean distance or a cosine between two vectors.

One of the models of vector semantics is the COALS model (Rohde, Gonnerman, Plaut 2005). In our study, on its basis, a semantic space was constructed over the National Corpus of Polish (NKJP), more specifically, over its balanced portion, comprising approx. 300 mil. segments (240 mil. words). The space can be downloaded from the APPROVAL website, along with its interface and a user manual in Polish (http://www.approval.uw.edu.pl/en_GB/przestrzenie_wyniki1). The interface can be run under Linux, Windows, or Mac OS operating systems. More information about the space can be found in Tatjewski et al. (2016, in print).

The NKJP semantic space – which to our knowledge is the first of this type constructed for the Polish language – can be used for a variety of purposes and the full scope of its applicability has yet to be investigated. Its task in the APPROVAL project was to provide data for the following purposes: 1) to compare the meaning of words analyzed within the project, 2) to validate other methods of assessing semantic proximity applied in the project, 3) to test certain hypotheses concerning the perception of loanwords as opposed to their native synonyms or better assimilated variants. Accordingly, the space was used to generate lists of the nearest neighbors of the words compared, as well as to measure the semantic distances between them.

Neighborhood lists

From among 42 word pairs investigated in the psychological part of the APPROVAL project, 6 pairs had to be left out because the frequency of one of their members was too low (the COALS model requires that words have some minimal frequency and the frequency threshold in the NKJP-based semantic space was set to 150 occurrences, including inflections). For each of the remaining 36 word pairs, neighborhood lists were generated and arranged in decreasing order of similarity to the word in question. The lists had 20 neighboring words each and were set in pairs so as to facilitate the comparison of the words for which they were produced (see Appendix 1).

In order to make comparison even easier, neighboring words common to both words compared were put on a grey background. Neighboring words which are related morphologically were marked the same way, unless the semantic difference between them was felt to be too large. For instance, the noun *nocleg* 'accommodation' and the derivative adjective *noclegowy* were marked grey in the neighborhood lists for *camping* and *kemping* (both meaning 'camping site'), but *hotel* and *hotelik* 'small and cozy hotel' were not. Thus, neighboring words which were not given a grey background are specific to one or the other of the words compared, and as such may be indicative of some meaning distinctions between them.

It is worth noting that the majority of words on a neighborhood list are related paradigmatically, but not syntagmatically to the word for which the list has been made. This is not surprising as the similarity in the semantic space depends on the similarity of contexts of word use. As such, neighborhood lists provide complementary data to those obtained from collocation analysis (see NKJP-based collocation images in the articles on particular words described in the APPROVAL project, http://www.approval.uw.edu.pl/en_GB/wybrane-ciagi).

A closer look at the neighborhood lists generated in the APPROVAL project reveals several drawbacks, some of which are inherent to the COALS model, others to its particular implementation. First, because of the frequency threshold, neighborhood lists can be produced for only the 42,200 most frequent lexemes represented in the NKJP corpus. Secondly, the disambiguation is insufficient, which makes it difficult to obtain a neighborhood list for a chosen meaning of a polysemous word. Third, lemmatization sometimes works poorly, which is obviously related to the problem of polysemy. Fourth, the abundance of proper names on some neighborhood lists indicates that the ability to leave out such names on demand would be a useful option the semantic space could provide.

The above observations can serve as guidelines for what may be changed in the next version of the semantic space built over the National Corpus of Polish.

Word proximities

Another use of the NKJP-semantic space, apart from generating neighborhood lists, is to measure the distance between any two chosen words. The distances between the members of the 36 word pairs included in Appendix 1 are listed in Appendix 2. One can compare them with the distances calculated on the basis of the free associations elicited in another part of the APPROVAL project (http://www.approval.uw.edu.pl/en_GB/percepcja_wyniki, report 1, appendix 4). As shown in Tatjewski et al. (2016, in print) the two word proximity measures are positively correlated on a statistically significant level, which makes this method worth considering as a viable alternative to the more costly elicitation methods. More research is needed, of course, to test their usefulness and the range of applicability.

References

Rohde D. L. T., Gonnerman L. M., Plaut D. C. 2005. *An improved model of semantic similarity based on lexical co-occurrence*. https://www.cnbc.cmu.edu/~plaut/papers/pdf/RohdeGonnermanPlautSUB-CogSci.COALS.pdf

Tatjewski M., Bańko M., Kucińska A., Rączaszek-Leonardi J. 2016, in print. *Computational distributional semantics and free associations: a comparison of two word-similarity models of synonyms and lexical variants*, [w:] *Language, Corpora and Cognition*, red. P. Pęzik, J. Waliński, K. Kosecki.

Published on 17/12/2015