Test Case Selection for Evaluating Measures of Semantic Distance

Vladislav D. Veksler & Wayne D. Gray
Cognitive Science Department
Rensselaer Polytechnic Institute
[vekslv, grayw] @rpi.edu

With a growing number of measures of semantic distance (MSD), such as LSA, GLSA, PMI-IR, WordNet, Normalized Google Distance (NGD), etc., we find the need for a standard direct comparison of these algorithms. A one-click browsing task affords direct comparisons of two or more MSDs. In the simplest version of the task, subjects search for a target term, T, and have to choose between one of two links, A and B. A choice of link A over link B suggests that in the subject's mind, T is more similar to A than it is to B. Here we formalize a procedure to select test case triads T-A-B, such that a human choice of A over B would directly implicate that one of the MSDs being investigated accords better with human judgment than another.

Test Case Selection

Test case selection consists of two parts. First, a large number of individual terms are collected. Second, all possible test triads are constructed and filtered accordingly.

Term Collection

Step 1. Random Term Selection

If we do not wish to disqualify MSDs due to their vocabulary limitations, the advice here is to limit the target and link terms to single word proper nouns that all MSDs can work with. One may collect random proper nouns through various means. One of the better sources that allows for specification of word length and frequency among other features is the Paivio, Yuille & Madigan word list generator (http://www.math.yorku.ca/SCS/Online/paivio/).

Step 2. Term Group Expansion

To control for term proximity – to make sure that target and link terms are at least somewhat related to each other – it may be a good idea to divide terms into groups. Each group would contain terms that are somewhat related to each other. This may be achieved using the closest-neighbor functionality that is offered by some MSDs, like LSA and GLSA. Grouping terms also reduces the number of MSD calculations to be performed; e.g. for 100 terms there would be 9900 (100 x 99) semantic distance calculations per MSD, while for 10 groups of 10 terms each, there would be only 900 (10 x 10 x 9) such computations.

Step 3. Term Inclusion in MSD and Human Vocabulary

Next, the new set of terms must be filtered in accordance with MSD vocabulary limitations, and filtered again in accordance with human vocabulary limitations. One way of ensuring that human subjects are familiar with all terms is by selecting only the terms that have high hit counts on Google or other search engines.

Test Triad Filtration

Step 4. Construct Test Triads

For every pair of competing MSDs, for each group of k terms, there are $k \cdot (k-1) \cdot (k-2)$ possible T-A-B test triads. For each such triad, similarity scores between terms T and A, and between terms T and B must be gathered for each MSD used. Due to differences in MSD scales, all similarity scores should be converted to Z-scores.

Step 5. Opposite MSD Predictions

The cases of interest are those in which the paired MSDs make different choices (i.e., where one MSD picks link A, while the picks link B as being the most closely associated with the target term, T). To be more stringent, each MSD should be 'certain' of its choice where certainly is operationally defined as a high difference in the Z-score between T-A versus T-B.

Step 6.

Finally, in order to avoid major priming effects in human subjects, none of the terms should persist across test cases.

Summary

Six steps are used to select the T-A-B triads given to human subjects. Although not discussed, it should be obvious that the procedure eliminates a large number of potential triads. The triads that remain are those that provide the highest level of discrimination among alternative MSDs. Hence, the procedure described here is optimized to obtain judgments from human subjects that will allow us to determine which MSD best describes human semantic space.

References

Cilibrasi, R. & Vitanyi, P.M.B. (2006). Similarity of objects and the meaning of words. Proc. 3rd Conf. Theory and Applications of Models of Computation (TAMC), J.-Y. Cai, S. B. Cooper, and A. Li (Eds.), Lecture Notes in Computer Science, Vol. 3959, Springer-Verlag, Berlin.

Kaur, I. & Hornof, A.J. (2005). A Comparison of LSA, WordNet and PMI-IR for Predicting User Click Behavior. Proceedings of the Conference on Human Factors in Computing, CHI 2005, 51-60.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.