Fast Semantic Relatedness

WordNet::Similarity versus Roget's Thesaurus

Alistair Kennedy
EECS, University of Ottawa
Ottawa, Ontario, Canada
akennedy@eecs.uottawa.ca

Stan Szpakowicz
EECS, University of Ottawa
Ottawa, Ontario, Canada
and
ICS, Polish Academy of Sciences
Warsaw, Poland
szpak@eecs.uottawa.ca

ABSTRACT

A Measure of Semantic Relatedness (MSR) automatically determines how close two words are in meaning. MSRs are used in such Natural Language Processing (NLP) problems as word-sense disambiguation or text summarization. To solve such problems may require millions of relatedness scores, but MSR run-time, clearly a major concern, has rarely been considered in NLP research. To evaluate an MSR, one often assigns relatedness scores to word pairs and measures the correlation with human-assigned scores. The WordSimilarity-353 test collection [1] is a known evaluation set of 353 word pairs with given relatedness scores. Spearman's correlation can be calculated, while Fisher's r-z transformation can be used to measure significance. We evaluate run-time performance of eleven MSRs previously evaluated with respect to correlation [2].

Resources like *WordNet* and *Roget's Thesaurus* have often been used to create MSRs. Ten MSRs are implemented in the *WordNet::Similarity* package [3], and one MSR uses the 1911 *Roget's Thesaurus* [2]. *Roget's* MSR calculates relatedness between two word-senses (appearances of a word in the *Thesaurus*) in constant time after finding them in an index. *WordNet*-based MSRs vary in complexity. A few could be implemented to run comparably fast to *Roget's* MSR, but *WordNet::Similarity* seldom does it. Our comparison of two popular MSR packages aims to inform researchers and developers who work on time-sensitive NLP applications.

BODY

The fastest & best-correlated WordNet MSRs, respectively, take 82 & 182 times longer than Roget's MSR, yet all are statistically equivalent.

REFERENCES

[1] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing Search in Context: the Concept Revisited. In *WWW'01 – Proc. of the 10th International World Wide Web Conference*, pages 406–414. ACM Press, 2001.

[2] A. Kennedy and S. Szpakowicz. Evaluating Roget's Thesauri. In *ACL-08: HLT – 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 416–424, 2008.

[3] T. Pedersen, S. Patwardhan, and J. Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts. In *AAAI-04 – Proc. of the 19th National Conference on Artificial Intelligence*, pages 1024–1025, 2004.

Volume 1 of Tiny Transactions on Computer Science