Semantic Relatedness: Computational Investigation of Human Data

Beata Beigman Klebanov
The Selim and Rachel Benin School of Engineering and Computer Science
The Hebrew University, Jerusalem, Israel
beata@cs.huji.ac.il

Abstract

We develop computational measures of semantic relatedness using data collected in psycholinguistic and annotation-style experiments. Datasets and measures are briefly described, and correlations with human data are presented.

Estimating the degree of semantic relatedness between words in a text is deemed important in numerous applications: word-sense disambiguation (Banerjee and Pedersen, 2003), story segmentation (Stokes et al., 2004), error correction (Hirst and Budanitsky, 2005), summarization (Barzilay and Elhadad, 1997; Gurevych and Strube, 2004).

To develop relevant software, human data is used; the most popular testbed is a list of 65 nouns ranked for degree-of-synonymy (Rubenstein and Goodenough, 1965) – henceforth, **RG**. A 30-pair subset of this dataset (**MC**) passed a number of replications (Miller and Charles, 1991; Resnik, 1995), and thus features highly reliable ratings. Additional datasets include 27 verb pairs ranked for similarity (Resnik and Diab, 2000) – **RD**, and 350 noun pairs ranked for a somewhat broader notion of relatedness (Finkelstein et al., 2002) – **FG**.

A much larger dataset was created as a result of an experiment that addressed patterns of lexical cohesion in texts (Beigman Klebanov and Shamir, 2006). Subjects were asked to mark common-knowledge based connections between pairs of words in 10 texts. This dataset represents a much broader rendering of the notion of semantic relatedness than in the other datasets, as there were neither restrictions on the part-of-speech nor on the type of semantic relation that holds between the members of the pair.

Equating the number of subjects who marked the pair with the pair's relatedness score results in a dataset of about 7000 pairs in total, with scores ranging from 1 to 20 (henceforth, **BS**); about half of the pairs are cross-part-of-speech. Beigman Klebanov (2006) estimates the stability of the resulting rankings at $r = .69 - .82$ for the 10 BS texts, averaging $r = .75$.

We explore 5 datasets – RG, MC, RD, FG, BS – with 3 computational measures of relatedness, based on WordNet, syntactic and text-based distribution, respectively. We propose a new measure of relatedness based on WordNet glosses and taxonomy called **GIC**, designed to handle cross-POS cases (Beigman Klebanov, 2006). GIC compares favorably with another WordNet-based measure that is capable of such comparisons (Banerjee and Pedersen, 2003). Additionally, we develop a syntax-based measure that estimates the salience of one word in the syntactic dependency relations of the other (**DEP**); we also use Latent Semantic Analysis (Deerwester et al., 1990) (**LSA**) as a measure of raw-text-based distributional relatedness. The three measures are scaled and combined by a simple additive procedure (**Com**). Table 1 summarizes the performance of the measures on the 5 datasets, along with state of art, if available.

Data	GIC	DEP	LSA	Com	S.-of-Art
MC	.78	.53	.73	.90	.85-.89
RG	.83	.44	.64	.87	.82-.84
RD	.55	.28	.08	.54	.67
FG	.47	.29	.55	.59	.54-55
BS	.28	.22	.28	.39	–

Table 1: Correlations with human rankings; GIC, DEP, LSA and Com vs. State-of-Art for MC/RG/RD/FG datasets. $r > .23$ is significant at $p < .01$. State-of-Art figures for the datasets are quoted from: RG/MC (Li et al., 2003; Jarmasz and Szpakowicz, 2003; Budanitsky and Hirst, 2006); RD (Resnik and Diab, 2000); FG (Finkelstein et al., 2002; Jarmasz and Szpakowicz, 2003).

To summarize our main findings:

- A system that handles nominal synonymy very well (Com on RG/MC) has difficulties with modeling a broader notion of relatedness and/or cross-pos data embodied in the BS dataset.[1] Since higher-level applications tend to employ relatedness measures developed mainly on nominal similarity data, the service these applications receive from the measures is expected to fall short of the target. Thus, we advocate the introduction the BS dataset for development of relatedness measures.

- WordNet, syntax and raw-text distributional similarity provide complementary information – a simple combination outperforms any of the single measures on all datasets apart from the verb similarity data (RD), where complete failure of LSA only allows the combination to nearly match the best performing single measure. This dataset is the only one where Com is below state-of-the-art.

Our current work is concerned with improving the correlations with BS dataset, by (a) devising better combination schemes by analyzing patterns of errors of the different measures; (b) employing an additional resource - the arrangement of the words in the given text, as BS data is based on specific texts. For example, a possible text-based predictor is the number of occurrences of a given content word in the text, under the assumption that repeatedly mentioned items are important and thus exert more influence on the textual cohesion, and on people's perception thereof.

Another direction for future work is empirical exploration of the utility of the more general notion of relatedness for language processing applications.

References

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of IJCAI*.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of ACL Intelligent Scalable Text Summarization Workshop*.

Beata Beigman Klebanov and Eli Shamir. 2006. Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, accepted for publication. Springer, Netherlands.

Beata Beigman Klebanov. 2006. Measuring semantic relatedness using people and WordNet. In *Proceedings of NAACL Short Papers*, to appear.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *To appear in Computational Linguistics.*

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society For Information Science*, 41:391–407.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of COLING*.

Graeme Hirst and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of RANLP*.

Yuhua Li, Zygaur Bandar, and David McLean. 2003. An approach to measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.

George Miller and Walter Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Philip Resnik and Mona Diab. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pages 399–404.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*.

Herbert Rubenstein and John Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Nicola Stokes, Joe Carthy, and Alan F. Smeaton. 2004. SeLeCT: A lexical cohesion based news story segmentation system. *Journal of AI Communications*, 17(1):3–12.

[1] Beigman Klebanov (2006) shows that state-of-art similarity measures cannot handle even the noun-only subset of BS.