Measuring Semantic Relatedness Using People and WordNet

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

Abstract

In this paper, we (1) propose a new dataset for testing the degree of relatedness between pairs of words; (2) propose a new WordNet-based measure of relatedness, and evaluate it on the new dataset.

1 Introduction

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).

Furthermore, Budanitsky and Hirst (2006) noted that various applications tend to pick the same measures of relatedness, which suggests a certain commonality in what is required from such a measure by the different applications. It thus seems worthwhile to develop such measures intrinsically, before putting them to application-based utility tests.

The most popular, by-now-standard testbed is Rubenstein and Goodenough's (1965) list of 65 noun pairs, ranked by similarity of meaning. A 30-pair subset (henceforth, **MC**) passed a number of replications (Miller and Charles, 1991; Resnik, 1995), and is thus highly reliable.

Rubenstein and Goodenough (1965) view similarity of meaning as degree of synonymy. Researchers have long recognized, however, that synonymy is only one kind of semantic affinity between words in a text (Halliday and Hasan, 1976), and expressed a wish for a dataset for testing a more general notion of semantic relatedness.[1]

This paper proposes and explores a new relatedness dataset. In sections 2-3, we briefly introduce the experiment by Beigman Klebanov and Shamir (henceforth, **BS**), and use the data to induce relatedness scores. In section 4, we propose a new WordNet-based measure of relatedness, and use it to explore the new dataset. We show that it usually does better than competing WordNet-based measures (section 5). We discuss future directions in section 6.

2 Data

Aiming at reader-based exploration of lexical cohesion in texts, Beigman Klebanov and Shamir conducted an experiment with 22 students, each reading 10 texts: 3 news stories, 4 journalistic and 3 fiction pieces (Beigman Klebanov and Shamir, 2006). People were instructed to read the text first, and then go over a separately attached list of words in order of their appearance in the text, and ask themselves, for every newly mentioned concept, "which previously mentioned concepts help the easy accommodation of the current concept into the evolving story, if indeed it is easily accommodated, based on the common knowledge as perceived by the annotator" (Beigman Klebanov and Shamir, 2005); this preceding helper concept is called an *anchor*. People were asked to mark all anchoring relations they could find.

The rendering of relatedness between two concepts is not tied to any specific lexical relation, but rather to common-sense knowledge, which has to do with "knowledge of kinds, of associations, of typical situations, and even typical utterances".[2] The phenomenon is thus clearly construed as much broader than degree-of-synonymy.

Beigman Klebanov and Shamir (2006) provide reliability estimation of the experimental data using

[1] "...similarity of meaning is not the same thing as semantic relatedness. However, there is at present no large dataset of human judgments of semantic relatedness" (Hirst and Budanitsky, 2005); "To our knowledge, no datasets are available for validating the results of semantic relatedness metric" (Gurevych, 2005).

[2] according to Hirst (2000), cited in the guidelines

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 13–16,
New York, June 2006.

statistical analysis and a validation experiment, identifying reliably anchored items with their strong anchors, and reliably un-anchored items. Such analysis provides high-validity data for classification; however, much of the data regarding intermediate degrees of relatedness is left out.

3 Relatedness Scores

Our idea is to induce scores for pairs of anchored items with their anchors (henceforth, **AApairs**) using the cumulative annotations by 20 people.[3] Thus, an AApair written by all 20 people scores 20, and that written by just one person scores 1. The scores would correspond to the perceived relatedness of the pair of concepts in the given text.

In Beigman Klebanov and Shamir's (2006) core classification data, no distinctions are retained between pairs marked by 19 or 13 people. Now we are interested in the relative relatedness, so it is important to handle cases where the BS data might under-rate a pair. One such case are multi-word items; we remove AApairs with suspect multi-word elements.[4] Further, we retain only pairs that belong to open-class parts of speech (henceforth, **POS**), as functional categories contribute little to the lexical texture (Halliday and Hasan, 1976). The *Size* column of table 1 shows the number of AApairs for each BS text, after the aforementioned exclusions.

The induced scores correspond to cumulative judgements of a group of people. How well do they represent the people's ideas? One way to measure group homogeneity is leave-one-out estimation, as done by Resnik (1995) for MC data, attaining the high average correlation of $r = 0.88$. In the current case, however, every specific person made a binary decision, whereas a group is represented by scores 1 to 20; such difference in granularity is problematic for correlation or rank order analysis.

Another way to measure group homogeneity is to split it into subgroups and compare scores emerging from the different subgroups. We know from Beigman Klebanov and Shamir's (2006) analysis that it is not the case that the 20-subject group clusters into subgroups that systematically produced different patterns of answers. This leads us to expect relative lack of sensitivity to the exact splits into subgroups.

To validate this reasoning, we performed 100 random choices of two 9-subject[4] groups, calculated the scores induced by the two groups, and computed Pearson correlation between the two lists. Thus, for every BS text, we have a distribution of 100 coefficients, which is approximately normal. Estimations of μ and σ of these distributions are $\mu = .69 - .82$ (av. 0.75), $\sigma = .02 - .03$ for the different BS texts.

To summarize: although the homogeneity is lower than for MC data, we observe good average intergroup correlations with little deviation across the 100 splits. We now turn to discussion of a relatedness measure, which we will evaluate using the data.

4 Gic: WordNet-based Measure

Measures using WordNet taxonomy are state-of-the-art in capturing semantic similarity, attaining r=.85 −.89 correlations with the MC dataset (Jiang and Conrath, 1997; Budanitsky and Hirst, 2006). However, they fall short of measuring relatedness, as, operating within a single-POS taxonomy, they cannot meaningfully compare *kill* to *death*. This is a major limitation with respect to BS data, where only about 40% of pairs are nominal, and less than 10% are verbal. We develop a WordNet-based measure that would allow cross-POS comparisons, using glosses in addition to the taxonomy.

One family of WordNet measures are methods based on estimation of information content (henceforth, **IC**) of concepts, as proposed in (Resnik, 1995). Resnik's key idea in corpus-based information content induction using a taxonomy is to count every appearance of a concept as mentions of all its hypernyms as well. This way, *artifact#n#1*, although rarely mentioned explicitly, receives high frequency and low IC value. We will count a concept's mention towards all its hypernyms AND all words[5] that appear in its own and its hypernyms' glosses. Analogously to *artifact*, we expect properties mentioned in glosses of more general concepts to be less informative, as those pertain to more things (ex., *visible*, a property of anything that is-a *physical object*). The details of the algorithm for information content induction from taxonomy and gloss information (IC_{GT}) are given in appendix A.

To estimate the semantic affinity between two senses A and B, we average the IC$_{GT}$ values of the 3 words with the highest IC$_{GT}$ in the overlap of A's and B's expanded glosses (the expansion follows the algorithm in appendix A).[6]

[3] Two subjects were revealed as outliers and their data was removed (Beigman Klebanov and Shamir, 2006).

[4] See Beigman Klebanov (2006) for details.

[5] We induce IC values on (POS-tagged base form) words rather than senses. Ongoing gloss sense-tagging projects like eXtended WordNet (http://xwn.hlt.utdallas.edu/links.html) would allow sense-based calculation in the future.

[6] The number 3 is empirically-based; the idea is to counter-balance (a) the effect of an accidental match of a

Data	Size	Gic	BP	Data	Size	Gic	BP
BS-1	1007	.29	.19	BS-6	536	.24	.19
BS-2	776	.37	.16	BS-7	917	.22	.10
BS-3	1015	.22	.09	BS-8	529	.24	.12
BS-4	512	.34	.39	BS-9	509	.31	.16
BS-5	1020	.25	.11	BS10	417	.36	.19

Table 1: Dataset sizes and correlations of Gic, BP with human ratings. $r > 0.16$ is significant at $p < .05$; $r > .23$ is significant at $p < .01$. Average correlation (Av_{BS}) is r=.28 (Gic), r=.17 (BP).

If A^* (the word of which A is a sense) appears in the expanded gloss of B, we take the maximum between the $\mathrm{IC}_{GT}(A^*)$ and the value returned by the 3-smoothed calculation. To compare two words, we take the maximum value returned by pairwise comparisons of their WordNet senses.[7]

The performance of this measure is shown under **Gic** in table 1. Gic manages robust but weak correlations, never reaching the $r = .40$ threshold.

5 Related Work

We compare Gic to another WordNet-based measure that can handle cross-POS comparisons, proposed by Banerjee and Pedersen (2003). To compare word senses A and B, the algorithm compares not only their glosses, but also glosses of items standing in various WordNet relations with A and B. For example, it compares the gloss of A's meronym to that of B's hyponym. We use the default configuration of the measure in WordNet::Similarity-0.12 package (Pedersen et al., 2004), and, with a single exception, the measure performed below Gic; see **BP** in table 1.

As mentioned before, taxonomy-based *similarity* measures cannot fully handle BS data. Table 2 uses nominal-only subsets of BS data and the MC nominal similarity dataset to show that (a) state-of-the-art WordNet-based similarity measure **JC**[8] (Jiang and Conrath, 1997; Budanitsky and Hirst, 2006) does very poorly on the relatedness data, suggesting that nominal similarity and relatedness are rather different things; (b) Gic does better on average, and is more robust; (c) Gic yields on MC to gain performance on BS, whereas BP is no more inclined towards relatedness than JC.

Data	Gic	BP	JC	Data	Gic	BP	JC
BS-1	.38	.18	.21	BS-6	.25	.16	.22
BS-2	.53	.18	.37	BS-7	.23	.10	.04
BS-3	.21	.04	.01	BS-8	.32	.10	.00
BS-4	.28	.38	.33	BS-9	.24	.17	.27
BS-5	.12	.07	.16	BS10	.41	.25	.25
Av_{BS}	.30	.16	.19	MC	.78	.80	.86

Table 2: MC and nominal-only subsets of BS: correlations of various measures with the human ratings.

Table 3 illustrates the relatedness vs. similarity distinction. Whereas, taxonomically speaking, *son* is more similar to *man*, as reflected in JC scores, people marked *family* and *mother* as much stronger anchors for *son* in BS-2; Gic follows suit.

AApair	Human	Gic	JC
son – man	2	0.355	22.3
son – family	13	0.375	16.9
son – mother	16	0.370	20.1

Table 3: Relatendess vs. similarity

6 Conclusion and Future Work

We proposed a dataset of relatedness judgements that differs from the existing ones[9] in (1) size – about 7000 items, as opposed to up to 350 in existing datasets; (2) cross-POS data, as opposed to purely nominal or verbal; (3) a broad approach to semantic relatedness, not focussing on any particular relation, but grounding it in the reader's (idea of) common knowledge; this as opposed to synonymy-based similarity prevalent in existing databases.

We explored the new data with WordNet-based measures, showing that (1) the data is different in character from a standard similarity dataset, and very challenging for state-of-the-art methods; (2) the proposed novel WordNet-based measure of relatedness usually outperforms its competitor, as well as a state-of-the-art similarity measure when the latter applies.

In future work, we plan to explore distributional methods for modeling relatedness, as well as the use of text-based information to improve correlations with the human data, as judgments are situated in specific textual contexts.

single word which is relatively rarely used in glosses; (b) the multitude of low-IC items in many of the overlaps that tend to downplay the impact of the few higher-IC members of the overlap.

[7] To speed the processing up, we use first 5 WordNet senses of each item for results reported here.

[8] See formula in appendix B. We use (Pedersen et al., 2004) implementation with a minor alteration – see Beigman Klebanov (2006).

[9] Though most widely used, MC is not the only available dataset; we will address other datasets in a subsequent paper.

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. 2005. Guidelines for annotation of concept mention patterns. Technical Report 2005-8, Leibniz Center for Research in Computer Science, The Hebrew University of Jerusalem, Israel.

Beata Beigman Klebanov and Eli Shamir. 2006. Reader-based exploration of lexical cohesion. *To appear in Language Resources and Evaluation.* Springer, Netherlands.

Beata Beigman Klebanov. 2006. Using people and WordNet to measure semantic relatedness. Technical Report 2006-17, Leibniz Center for Research in Computer Science, The Hebrew University of Jerusalem, Israel.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

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

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of IJCNLP.*

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English.* Longman Group Ltd.

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

Graeme Hirst. 2000. Context as a spurious concept. In *Proceedings of CICLING.*

Jay Jiang and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics.*

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

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity-measuring the relatedness of concepts. In *Proceedings of NAACL.*

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.

A Gloss&Taxonomy IC (IC_{GT})

We refer to POS-tagged base form items as "words" throughout this section. For every word-sense W in WordNet database for a given POS:

1. Collect all content words from the gloss of W, excluding examples, including W^* - the POS-tagged word of which W is a sense.
2. If W is part of a taxonomy, expand its gloss, without repetitions, with words appearing in the glosses of all its super-ordinate concepts, up to the top of the hierarchy. Thus, the expanded gloss for *airplane#n#1* would contain words from the glosses of the relevant senses of *aircraft* , *vehicle*, *transport*, etc.
3. Add W's sense count to all words in its expanded gloss.[10]

Each POS database induces its own counts on each word that appeared in the gloss of at least one of its members. When merging the data from the different POS, we scale the aggregated counts, such that they correspond to the proportion of the given word in the POS database where it was the least informative. The standard log-frequency calculation transforms these counts into taxonomy-and-gloss based information content (IC_{GT}) values.

B JC measure of similarity

In the formula, IC is taxonomy-only based information content, as in (Resnik, 1995), LS is the lowest common subsumer of the two concepts in the WordNet hierarchy, and Max is the maximum distance[11] between any two concepts.

$$JC(c_1, c_2) = Max-(IC(c_1)+IC(c_2)-2\times IC(LS(c_1, c_2))$$

To make JC scores comparable to Gic's [0,1] range, the score can be divided by Max. Normalization has no effect on correlations.

[10] We do add-1-smoothing on WordNet sense counts.

[11] This is about 26 for WordNet-2.0 nominal hierarchy with add-1-smoothed SemCor database; see Beigman Klebanov (2006) for details.