Differences in Semantic Relatedness as Judged by Humans and Algorithms

Nicolas Neubauer, Nils Haldenwang, Oliver Vornberger

University of Osnabrück, Institute of Computer Science
Albrechtstraße 28, 49076 Osnabrück, Germany
{nneubauer, nils.haldenwang, oliver}@uni-osnabrueck.de

Abstract

Quantifying the semantic relatedness of two terms is a field with vast amounts of research, as the knowledge provided by it has applications for many Natural Language Processing problems. While many algorithmic measures have been proposed, it is often hard to say if one measure outperforms another, since their evaluation often lacks meaningful comparisons to human judgement on semantic relatedness. In this paper we present a study using the BLESS data set to compare the preferences of humans regarding semantic relatedness to popular algorithmic baselines, PMI and NGD, which shows that significant differences in relationship-type preferences between humans and algorithms exist.

1. Introduction

Research regarding the *semantic relatedness* of two terms in natural language processing tasks tries to answer the question of, as El-Yaniv and Yanay (2012) quote from Wikipedia, "How much [...] term A [does] have to do with term B". This very general question covers a large amount of semantic relationships between terms while a related concept is *semantic similarity*: According to Budanitsky and Hirst (2006) "similar entities are semantically related by virtue of their similarity". Hence, two terms can very well be related, although not strictly similar ("pencil"-"paper") or even related through antonymy ("hot"-"cold").

The knowledge among semantic relatedness of terms can be applied to multiple tasks in natural language processing such as information extraction and retrieval, spelling error detection, question answering and many more. One example of an application employing automated text-comprehension is Apple's voice controlled digital assistant "Siri". The question "Will it rain today?" results in the presentation of a current weather report. Understanding the concept of "rain" in this case – or generally, what is actually meant with a given input, is crucial to this and many other applications.

Multiple so called *Measures of Semantic Relatedness* (MSRs) have been introduced and evaluated in the past, surveyed for instance in Budanitsky and Hirst (2006), with many more in related works (Landauer and Dumais, 1997; Turney, 2001; Cilibrasi and Vitanyi, 2007; Lindsey et al., 2007; Landauer and Dumais, 1997). However, a common problem is the comparability of performance. It is common to evaluate the models in "semantic tasks" (Baroni and Lenci, 2011), thus the question whether one measure is a good or a bad one, if one outperforms another, is answered indirectly. These evaluations often lack insight in *why* a specific model performs the way it does. Some tests make use of ratings made by humans (Rubenstein and Goodenough, 1965; Miller and Charles, 1991; Nelson et al., 2004), but often do not discuss how well and why a model mimics human judgement on semantic relatedness.

In this paper we present the results of a study we performed to measure how humans judge semantic relatedness on the basis of the recently released BLESS dataset by Baroni and Lenci (2011). We provide multiple analysis results and compare the human preferences to two popular and well-established baseline measures, the *Pointwise Mutual Information* (PMI) (Turney, 2001) and *Normalized Google Distance* (NGD) (Cilibrasi and Vitanyi, 2007), discovering significant differences.

2. Algorithmic Baselines

As stated in the introduction, we see "semantic relatedness" as a very general concept, describing a relationship between two terms, which is informally defined as a measurable amount of how much two terms *have to do with each other* - maybe based on a informal consensus among human judges: We would like to look at their relations with regard to the *meaning* of the two terms.

While we are not trying to give a better or more formal definition than the one given by El-Yaniv and Yanay (2012), we want to point out some features of a measure for semantic relatedness, that are in our opinion desirable.

First and foremost we want a measure to perform *human-like*, which means having a high correlation with a human judgement. This alone is quite hard, maybe even impossible to achieve as pointed out by Nelson et al. (2004), since different cultures rate relatedness with regard to their personal experience and background.

However, not only the relatedness itself but the quantification of it is hard to grasp. While most will agree that "rain" and "weather" are somehow related, so are "car" and "engine" - but which of the two pairs is stronger related? And even if a consistent answer could be found, *how large* is the difference in semantic relatedness?

Although it is certainly desirable for a MSR to provide a good overall correlation with human judgement, we believe a more important feature is the correlation within a group of term pairs with one fixed term. Within such a group like "car" - "engine", "car" - "wheel", and "car" - "door", a MSR should be able to provide a ranking of high correlation with a human ranking, which - within the group - should be more consistent than the overall rankings.

Before looking into the experimental results of a study we conducted to gain more insight into human preferences regarding these features, we would like to introduce two popular algorithmic MSRs which will serve as algorithmic baselines.

2.1. Pointwise Mutual Information

The PMI (Turney, 2001) has often been used as a MSR (Lindsey et al., 2007; Turney et al., 2010).

The measure is based on the probability of two terms occurring in the same window of text versus the probabilities of the two terms occurring separately, meaning it needs a reasonably large text corpus of background knowledge to compute its relatedness scores. For the evaluations in the following sections, we used a 2009 dump of the English Wikipedia and a window size of three sentences serving as *context*.

The measure uses the probability of two terms occurring in the same window of text versus the probabilities of the two terms occurring separately. The score of relatedness of two given terms x and y is computed as

$$PMI(x, y) = \log_2 \left(\frac{p(x, y)}{p(x) \cdot p(y)} \right)$$

where $p(x), p(y)$ is the probability of the the term x, respectively y, to occur in a text-window and $p(x, y)$ denotes the dependent probability that both terms x and y occur within the same text-window. When x and y are statistically independent, the probability of their co-occurrence is $p(x) \cdot p(y)$. If they are not independent, which means they often occur together, $p(x, y)$ will be greater than $p(x) \cdot p(y)$. Hence the ratio $\frac{p(x,y)}{p(x) \cdot p(y)}$ will yield higher values for dependent terms x and y and therefore PMI is a measure for the statistical dependence of the term's co-occurrence.

The probabilities can be easily computed from any given text corpus:

$$p(x) = \frac{\text{number of text-windows containing x}}{\text{number of text windows in the corpus}}$$

$$p(x, y) = \frac{\text{number of text-windows containing x and y}}{\text{number of text windows in the corpus}}$$

2.2. Normalized Google Distance

The NGD (Cilibrasi and Vitanyi, 2007) is another MSR, performing well for various tasks. Like PMI it harnesses the probability of co-occurrences in a text corpus. NGD is computed as follows

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where $f(x), f(y)$ denotes the number of text-windows including the term x, respectively y, $f(x, y)$ the number of windows including both, and M is the overall number of text-windows.

By this definition the NGD measures the distance between two given terms. As we want to measure the relatedness, we are using the *Normalized Similarity Score* (NSS) (Lindsey et al., 2007), which is computed as

$$NSS(x, y) = 1 - NGD(x, y).$$

3. Related Experiments and the BLESS Dataset

3.1. Rubenstein and Goodenough & Miller and Charles

One of the earliest test sets containing human judgement on semantic relatedness was created by Rubenstein and Goodenough (1965) as early as 1965 during an experiment with up to 36 college students, which were to rate the "similarity of meaning" of in total 65 pairs of nouns on a scale from 0 to 4. They used the data they obtained during the experiment to conclude "that a pair of words is highly synonymous if their contexts show a relatively great amount of overlap".

Later on, Miller and Charles (1991) used a subset of 30 of the original 65 noun pairs of Rubenstein and Goodenough's test set in another experiment, in which they obtained new ratings from 0 to 4 by 38 students, which they found to have an exceptionally high correlation with the original scores from 1965. This suggests that what we described as as consensus of human judges regarding semantic similarity actually exists.

These test sets are widely used when measuring the performance of a MSR. However, they were not designed to provide a benchmark for MSRs. Thus, the choice of word pairs does not provide a very good coverage of different types of semantic relations we would expect for a benchmark set, not to mention the very small size of the set of only 65 word pairs.

3.2. WordSim-353

The WordSim-353 dataset (Finkelstein et al., 2001) consists of 353 word pairs rated by 16 human subjects. It includes the 65 pairs used in Rubenstein and Goodenough's set. Given a word pair, the subjects had to assign a score between zero (totally unrelated) and ten (very related). To evaluate a MSR the authors of the dataset reported correlation scores with the human judgement.

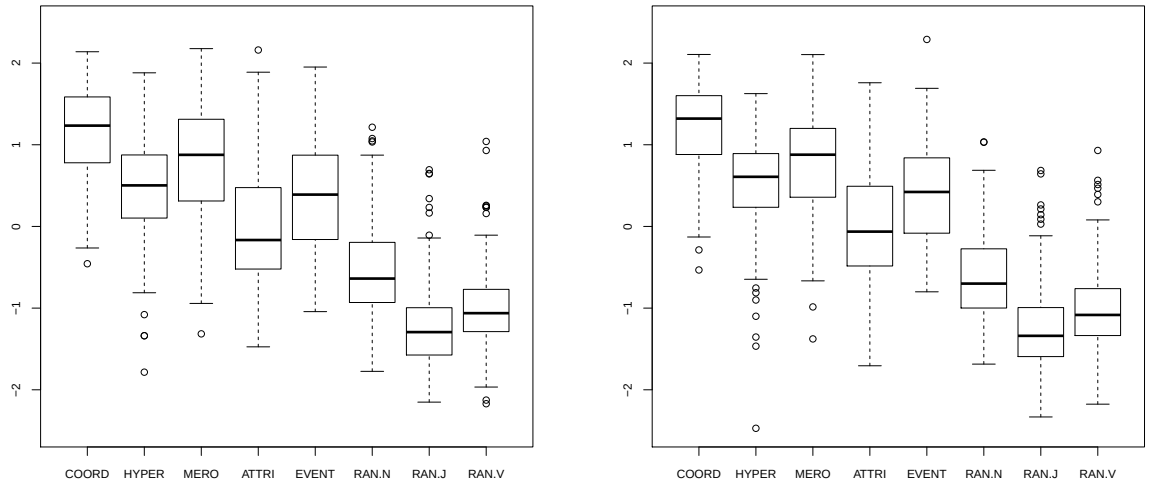
According to Baroni and Lenci (2011), WordSim-353 suffers mainly from the large variety of different types of semantic relations within the set. Evaluating a MSR favoring synonyms over meronyms may yield worse results than another one, which is not a better algorithm in general, but just favors different semantic relations.

Although considerably larger than the sets of Rubenstein and Goodenough, and Miller and Charles, the set still is fairly small and does not allow a meaningful evaluation when one wants to group words pairs with one common term.

3.3. BLESS

Baroni and Lenci (2011) have designed their BLESS dataset specifically with the evaluation of MSRs in mind.

The main problem they see with current testing of MSRs is that researchers often "compar[e] a single quality score", such as TOEFL accuracy or WordSim-353 correlation, to evaluate if a specific MSR *outperforms* others. While we do not think that it is a generally bad idea to quantify the performance of a MSR, we strongly agree that it is important to reflect on "how and why the models dif-



(a) Boxplot of PMI Results

(b) Boxplot of NSS Results

Figure 1: Distribution of z-transformed Scores Across Concept Types

fer". The BLESS set is designed specifically to give more insight into the latter.

The dataset consists of 200 nouns with 5 different *true* types of relations as well as 3 *random* types. With each of the 8 relation types containing on average about 133 related terms, the whole dataset consists of 26554 concept-type-relatum tuples. To assess a specific MSR's behavior regarding different types of semantic relations, the authors have chosen the 5 classes of semantic relations as co-hyponym or coordinate, hypernym, meronym, attribute, and event. The former three being nouns, attributes being adjectives and events being verbs. The three random types are divided by their word type, with nouns, adjectives and verbs as a controlled group of non-semantically related terms to the original nouns.

The suggested experiment setup is based on box plots of per-concept z-transformed similarity scores. For each concept the authors pick the best result, *nearest neighbor*, per relation type in order to average out model specific preferences (e.g. a preference towards technical terms). The resulting 8 transformed values are then collected for all 200 concepts and presented in a box plot. Significance of comparisons is then evaluated using a Tukey Honestly Significant Difference test.

In our opinion, the test set is very thoughtfully compiled and covers a wide range of terms. However, we would like to note that the set sometimes relies on the supplied *part-of-speech* tagging, which can confuse experiments not using this data. The set for example contains the tuple *dress-n - uniform-j* in the group of *random* associated adjectives. While this is obviously true, leaving the POS tagging out, it creates the tuple *dress - uniform*, which now is strongly related. We did not review these problems in-depth, but want users of the set to be aware of this fact.

Using the set, we noted that some relata have been added very often, such as *old* (associated to 157 of 200 concepts) or *new* (96 of 200) in the attribute group. While almost everything, even abstract concepts, can be considered old or new, we think the relation is most of the times very weak. In comparison, the random group of verbs contains the tuple *spear - take*. Thus, the set is not very strict regarding what is considered truly related and what is definitely not.

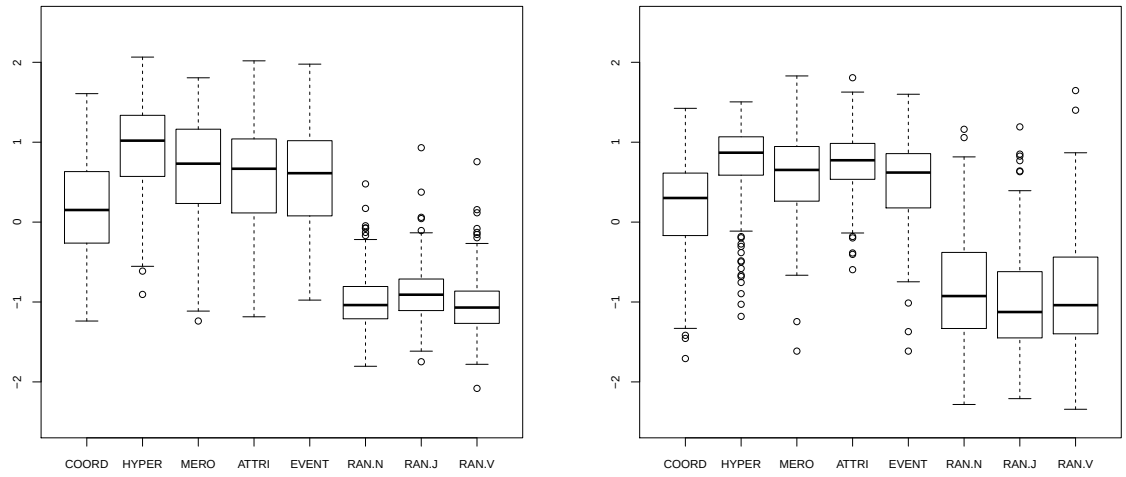
As displayed in figure 1, both algorithmic measures, PMI and NSS, are well able to discriminate the *true* relata from the random ones. Differences between each *true* relation type and the random types are all significant ($\alpha = 0.05$). Interestingly PMI and NSS produce an extremely similar output and both show a distinctive preference towards co-hyponyms.

Looking at the plot and pairwise comparisons for PMI and NSS, hypernyms and events are statistically indistinguishable, while both measures show significantly different distributions among all random group types.

4. Human Judgement on BLESS

As we stated before, the BLESS test set provides a means to evaluate what types of relations a specific MSR prefers and, with its controlled random set, it also allows statements about the ability of a MSR to discriminate *true* relata from random ones. Lacking human judgement on semantic relatedness, the set however can not be used to determine the correlation of human choices with the ones of a MSR. Therefore, we decided to take the BLESS test set as a basis for a human rating experiment.

The experiment was performed by showing the participants a website with a small instructional section followed by a concept and one randomly chosen relatum per relation type (see 3.3.). Thus, every page showed one concept and 8 relata, which the participants should rate regarding their semantic relatedness with the concept. The instructional section read:



(a) Boxplot of Human Choices (Median)

(b) Boxplot of Human Choices (Max)

Figure 2: Distribution of z-transformed Scores as Voted by the Experiment Participants Across Concept Types

> For the word in the headline below, please rate its *relatedness* to the words in the table using the sliders. Values from 0 (not related at all) to 10 (extremely related) can be assigned. Click on "Submit Ratings" after rating *all* words to save your ratings and show the next word.
>
> If you are not sure about the meaning of a word, hovering the dictionary icon brings up an explanation in English. If you don't understand the explanation or don't know the meaning of the word in the headline, please click the "Different Word..." button to rate something else. Please do not submit ratings based on guessed meanings.

They could choose a rating from 0 to 10 (with 0 meaning not related and 10 very related) on a continuous scale. Also, they were only able to rate using a continuos slider interface (no direct number input was allowed), see 3 for a screenshot of the rating interface. Since among the displayed relata have always been three terms from the random group, we believe that most participants had a similar feeling of what the 0-point of the scale meant. We also assumed that with the slider starting in the middle (5) position, a value below 5 would have a "not so much related" meaning while a value above would represent a "rather related" connection of the terms. We would like to point out that users had to move every slider in order to have their ratings submitted, which means that they were not able to leave it in the starting position and thus were forced to actively choose the rating for every word.

The experiment was performed in a semi-controlled, semi-crowd-sourced environment with the majority of ratings having been submitted by university students who received credits for their participation. The website, however, was available to other participants as well. All in all, we tracked 155 different rating sessions. It is noteworthy that all of the participants were non-native English speakers and were allowed to use a dictionary for relata term explanations[1], but the instructions stated not to submit ratings when the cue term was unknown to them (in which case they could skip the entire set). The experiment resulted in 13370 total ratings for term pairs and covered all of the concepts in the BLESS test set.

While the number of ratings is considerable, due to the enormous size of the BLESS data set, only about half of all term pairs have been rated. Thus, with term pairs most often being rated by only one person, we would not recommend to perform a correlation performance evaluation such as usually done with the WordSim-353 set (see section 3.2.). Still, when being aggregated, the data can very well be used to compare the preferences of humans regarding the different types of semantic relations included in the BLESS test set.

For figure 2(a) we chose the median of ratings per concept and relation type over all rated relata, since model specific preferences, the reason why the authors of the data set suggested to use the nearest neighbor, are not a problem here – but plainly wrong answers, due to lacking word comprehension, are. Nevertheless, we also generated a box plot from the nearest neighbor results in figure 2(b), using the exact same evaluation procedure as in section 3.3.[2].

As seen in figure 2, unsurprisingly humans show a very good separation of random type relata from true type relata. Unexpected to us, however, is the fact that our experiment participants show a clear preference towards hypernyms, with co-hypernyms being the least preferred group of relata. This is especially noteworthy, since all the MSRs

[1] We integrated explanations in English from *wordnik* `http://www.wordnik.com` into the interface.

[2] The score for a given term pair still was defined as the median of assigned scores. However, there were seldom more than two scores available.

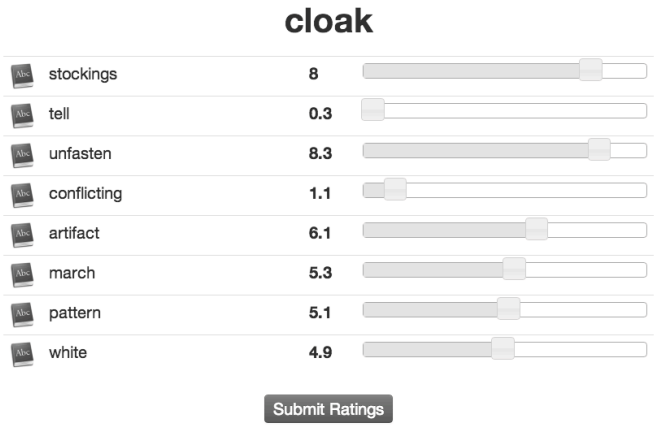


Figure 3: The Rating Interface used in the Experiment

we presented before significantly prefer co-hypernyms.

Another observation one can make is the fact that the different groups of relation types seem less differentiated than in the results for the algorithmic MSRs. Looking at the pairwise comparisons of distributions, the results show no significant difference between attributes and hypernyms, as well as attributes and meronyms, and meronyms and events. Also, no statistical difference between the different types of random relata can be detected.

5. Conclusion

In the former sections we motivated the importance, especially for information retrieval systems, of being able to determine the semantic relatedness between terms. With *PMI* and *NSS* we presented two state-of-the-art measures which we compared to the preferences of human judges by providing results of an experiment we conducted using the BLESS dataset.

We also revealed the fact that our human test subjects displayed a significant preference towards hypernyms over co-hypernyms, while both of the algorithmic MSRs' preferences where exactly the opposite.

6. Further Work

The results of our human judgement experiment suggest that further research regarding the preferences of MSRs is needed. For more definitive statements on human preferences, more experiments, for example with native-speakers in comparison to our test results, should be conducted. The results could then be used to refine parameters in MSRs to better mimic the human preferences.

Also, if enough data could be gathered, the complete BLESS test set could be annotated with human ratings and thus would allow to compare correlations of different MSRs on a very large and thought-through test set.

References

Baroni, M. and A. Lenci, 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics.

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

Cilibrasi, R.L. and P.M.B. Vitanyi, 2007. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383.

El-Yaniv, R. and D. Yanay, 2012. Supervised learning of semantic relatedness. *Machine Learning and Knowledge Discovery in Databases*:744–759.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*. ACM.

Landauer, T.K. and S.T. Dumais, 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, 104(2):211.

Lindsey, R., V.D. Veksler, A. Grintsvayg, and W.D. Gray, 2007. Be wary of what your computer reads: the effects of corpus selection on measuring semantic relatedness. In *8th International Conference of Cognitive Modeling, ICCM*.

Miller, G.A. and W.G. Charles, 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Nelson, D.L., C.L. McEvoy, and T.A. Schreiber, 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, 36(3):402–407.

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

Turney, P.D., 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*. Springer-Verlag.

Turney, P.D., P. Pantel, et al., 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.