Lindsey, R., Veksler, V. D., Grintsvayg, A., & Gray, W. D. (2007).
Be wary of what your computer reads: The effects of corpus selection on measuring semantic relatedness.
Proceedings of the 8th International Conference on Cognitive Modeling. Ann Arbor, MI.

Be Wary of What Your Computer Reads: The Effects of Corpus Selection on Measuring Semantic Relatedness

Robert Lindsey (lindsr@rpi.edu)
Vladislav D. Veksler (vekslv@rpi.edu)
Alex Grintsvayg (grinta@rpi.edu)
Wayne D. Gray (grayw@rpi.edu)

Rensselaer Polytechnic Institute, 110 8th Street
Troy, NY 12180 USA

Abstract

Measures of Semantic Relatedness (MSRs) provide models of human semantic associations and, as such, have been applied to predict human text comprehension (Lemaire, Denhiere, Bellissens, & Jhean-Iarose, 2006). In addition, MSRs form key components in more integrated cognitive modeling such as models that perform information search on the World Wide Web (WWW) (Pirolli, 2005). However, the effectiveness of an MSR depends on the algorithm it uses as well as the text corpus on which it is trained. In this paper, we examine the impact of corpus selection on the performance of two popular MSRs, Pointwise Mutual Information and Normalised Google Distance. We tested these measures with corpora derived from the WWW, books, news articles, emails, web-forums, and encyclopedia. Results indicate that for the tested MSRs, the traditionally employed books and WWW-based corpora are less than optimal, and that using a corpus based on the New York Times news articles best predicts human behavior.

Keywords: Measures of Semantic Relatedness, semantic similarity, training corpus, corpus comparison, Pointwise Mutual Information, PMI, Normalised Google Distance, NGD, computational linguistics, natural language processing.

Introduction

Adding a text-comprehension component to cognitive models is a worthy goal, but it is a goal with many obstacles standing in its way. Although grammar parsing is still a major problem in computational linguistics, we are close to being able to accurately approximate relative meanings of words and documents. Using statistical techniques known as Measures of Semantic Relatedness (MSRs), we can automatically extract word definitions and relationships from large text corpora.

MSRs have been used in modeling language acquisition (Landauer & Dumais, 1997), human web-browsing behavior (Fu & Pirolli, 2007), text comprehension (Lemaire et al., 2006), semantic maps (Veksler & Gray, 2007) and many other modeling applications. In more applied domains, MSRs have been used to develop a wide variety of applications such as augmented search engine technology (Dumais, 2003) and automated essay-grading algorithms for the Educational Testing Service (Landauer & Dumais, 1997). MSRs have a wide range of practical applications and are potentially useful to any cognitive model or AI agent dealing with text (Veksler, Grintsvayg, Lindsey, & Gray, Submitted).

MSR performance depends on the corpus on which it is trained. Imagine if a child learning the English language were only allowed to read Shakespeare. Although the child would certainly learn English, he or she would undoubtedly encounter a number of communications problems. A conversation with this child would be difficult because they learned a very out-of-style form of English. Many of the words in the text the child learned from are used less often nowadays, some of those words are used more often, and some maybe not at all. Moreover, many of the words would have acquired new meanings, or would be used in different contexts than in Shakespeare's day. In addition, a child exposed exclusively to Shakespeare might be able to converse about love and war, but not about how to hail a taxi or how to reboot a computer. All in all, the choice of a set of learning material, or text corpora, for children has a profound impact on how well they comprehend English. The same concept applies to MSRs.

MSRs try to learn word relations the same way children do (Landauer & Dumais, 1997; Newport & Aslin, 2004; Newport, Hauser, Spaepen, & Aslin, 2004), and consequently their effectiveness is dependent on the text corpus from which they glean information. Whereas children's exposure to speech and text may, to some degree, be considered open to many sources, MSRs are strictly bound by their training corpora. MSRs calculate the probability of the co-occurrence of two query words in order to ascertain their semantic relatedness value. This probability varies greatly from one corpus to another, so the output of MSRs trained on different text corpora also varies greatly. There are many corpora commonly used to train MSRs and each produces different semantic relatedness values.

Landauer and Dumais (1997) claim that because children do not hear most of their lexicon, they must gain their vocabulary through books. Consequently, MSRs are often trained on books in the hopes of gaining knowledge from the same source as children. To our knowledge, this corpus choice has never been objectively validated and rigorously examined in comparison with other corpus types.

Certain MSRs may take as their corpus the entire World Wide Web (Turney, 2001). A naive assumption might be that such an overwhelmingly large amount of text will result in properly trained MSRs, and that although some web pages will not accurately represent the semantic relations of our language, those few unhelpful websites are statistically insignificant. To our knowledge, the use of the World Wide

Web as a training corpus is just as unfounded as the use of any given corpus of books.

Using books, the internet, or any other text corpus that has not been studied rigorously for its effect on the performance of MSRs may seriously compromise MSR-based applications. Just as children must be trained on proper material, an MSR must likewise be trained on the proper corpus in order to accurately model human lexical knowledge. In this paper, we examine the impact of corpus selection on the performance of two popular MSRs, Pointwise Mutual Information and Normalised Google Distance. We tested these measures in combinations with WWW, books, news articles, emails, web-forums, and encyclopedia-like corpora. All MSR-corpus pairs were evaluated as to their ability to represent human lexical knowledge based on data from a large-scale free-association psychological study (Nelson, McEvoy, & Schreiber, 1998).

The Evaluation Challenge

Deciding how to evaluate the goodness of MSRs presents a daunting challenge. First, there are at least a dozen MSRs in the published literature and more are being invented each year. Second, as we will show, the goodness of an MSR depends at least partially on the corpus on which it is trained. Third, it may well be that different MSRs capture human semantic relatedness more so in some tasks (e.g., deciding what link to click on next) than in others (e.g., deciding if the content of a paragraph provides the answer to a sought-after question).

Clearly, we do not have room in this small paper to exhaustively explore the problem space implied by the combination of these three factors. Rather, as discussed below, we choose two MSRs, a small set of large corpora, and one criteria task on which reliable and valid measures of human performance exist. However, our work is ongoing, and we intend this paper to be an exemplar, not an exhaustive, evaluation of MSRs.

Measures of Semantic Relatedness

MSRs give computers the ability to quantify the meaning of text. MSRs define words in terms of their connection strengths to other words, and they define connection strengths in terms of word co-occurrence. In other words, two terms are related if they often occur in the same contexts. Two terms are synonymous if their contexts are identical.

Pointwise Mutual Information (PMI)

PMI is a well-established and successful measure for approximating human semantics (Turney, 2001). PMI is based on the probability of finding two terms of interest (t1 and t2) within the same window of text versus the probabilities of finding each of those terms separately:

$$PMI(t1,t2) = \log_2\left(\frac{P(t1,t2)}{P(t1)\times P(t2)}\right)$$

where P(t1) and P(t2) are the probabilities of finding a window of text in the corpus containing the term t1 or t2 respectively; and P(t1,t2) is the probability of finding a window of text in the corpus containing both t1 and t2. Please see Turney (2001) for a more expanded discussion of PMI.

Window-size is a free parameter in PMI and most-all other MSRs. For web-based corpora window size is typically set to be a webpage; however, it can also be any grouping of text – a sentence, an email, a webpage, or some other organizational group.

Normalised Google Distance (NGD)

NGD is another popular MSR (Cilibrasi & Vitanyi, 2007) that measures the similarity between two terms by using the probability of co-occurrences as demonstrated by the following equation:

$$NGD(t1,t2) = \frac{\max\{\log f(t1),\log f(t2)\} - \log f(t1,t2)}{\log M - \min\{\log f(t1),\log f(t2)\}}$$

where M is the total number of searchable Google pages, and *f(x)* is the number of pages that a Google search for *x* returns.

Although NGD was originally based on the Google search engine, this formula may be used in combination with other text corpora just as well. That Google's entire document-base is a better text corpus for this MSR is exactly the premise that we wish to challenge in the current work.

In order to use NGD as a relatedness measure, rather than a measure of distance, we convert NGD scores into similarity scores by subtracting NGD from 1 (1 being the maximum NGD score). From this point forth we will refer to the similarity score based on the NGD formula as the *Normalized Similarity Score* (*NSS*).

Corpus Issues

A text corpus used to train an MSR may suffer from a variety of problems that impair its effectiveness. The content may be too old to accurately represent the semantic relatedness of words, as modern language uses words more or less frequently than in the past. Thus, classic literature may not be the ideal training corpus for MSRs.

Text corpora may also be too biased to be useful. For example, a corpus comprised of writings from a single political party will likely lead an MSR to calculate an overly strong relatedness between words like “axis” and “evil”. Likewise, a biased corpus may calculate a weak relatedness between words in situations where it should be higher. We may find that the internet has a commercial (or some other) bias and thus will not make a good overall training corpus.

Additionally, text corpora may be too impoverished or contain bad examples of language. A log of instant messaging conversations, for example, may provide a poor source of the English language. Using poorly written text as

the training material would be similar to learning English from someone who does not speak English.

Text corpora may be too structured. For example, a dictionary or an encyclopedia may turn out to be a poor training source for MSRs.

By the same token, text corpora may be too unstructured. We presume that web forums contain conversational English and would thus make a great MSR training source, but the lack of structure in such corpora may make these suboptimal, as well.

Additionally, a text corpus may be computationally expensive to use. If it is excessively large, many MSRs will take a long time to produce a result, and some MSRs will not be able to produce the result at all.

Corpus Evaluation

In order to select an optimal training corpus for an MSR, many corpora must be tested and have their performances compared. We studied two MSRs, PMI and NSS, and evaluated their performance on six unique corpora. The following sections describe the method by which we performed our evaluations.

MSRs

PMI and NSS were the two MSRs used in this study. These are two popular MSRs that can handle all of the corpus types that we were considering in our research. Other MSRs, e.g. LSA (Landauer & Dumais, 1997), GLSA (Matveeva, Levow, Farahat, & Royer, 2005), ICAN (Lemaire & Denhiére, 2004), simply cannot handle large corpuses (e.g. WWW).

For five of the corpora the text-window size was a webpage. For the sixth corpus, the Enron Email Corpus, the text-window size was an email.

Corpora

Google Corpus

This corpus is an extremely large collection of text (the World Wide Web), and is a popular choice for a training corpus. One major advantage of this corpus is that MSRs run extremely fast on it. Counting the number of hits returned by a search takes an inconsequential length of time.

Wikipedia Corpus

Wikipedia is the largest, free-content encyclopedia on the internet. We chose to study this corpus because it represents a great wealth of human knowledge. In order to use this corpus, we count the hits returned by a Google search for the terms after restricting our results to "site:wikipedia.org".

New York Times Corpus

New York Times is a news source that we chose to study as a corpus because of their large collection of online articles. We access this corpus the same way we access Wikipedia, by restricting Google searches to "site:nytimes.com".

Project Gutenberg Corpus

Books make a popular choice as an MSR training corpus (e.g. Landauer & Dumais, 1997). Project Gutenberg is an online collection of over 20,000 books. This corpus represents one of the largest online collections of books available. In order to use this corpus, we count the hits returned by Google searches restricted to "site:gutenberg.org/files".

Google Groups Corpus

Google Groups is a subdivision of Google's website that hosts online discussions and forums. This corpus was chosen because it represents a large collection of informal conversational language. We use this corpus in the same way we use Wikipedia, New York Times, and Project Gutenberg – by restricting our searches on Google to "site:groups.google.com".

Enron Email Corpus

Some time ago, a large collection of emails from Enron Corporation's top management personnel was released to public-domain. We chose to study this collection of emails as a training corpus because it is one of the largest collections of emails available. The hypothesis is that emails may make for an excellent corpus choice because they contain modern conversational language. In order to use this corpus, we imported all email bodies into a database, and ran queries on this database to find out the probability/frequency information for each PMI/NSS request.

Limitations

Each of the corpora we chose represents a sampling from the set of all possible corpora. It is unclear to us how representative or non-representative our selection is of this larger set. Indeed, it is unclear to us how to formally characterize our selected corpora or the larger set of corpora so as to answer this question. Hence, our only claim for our current work is that we compare each of our two selected MSRs on each of our six corpora. These comparisons should allow us to begin to characterize the ways in which these two MSRs predict human performance when provided with equal training. (We view our effort as the first study of its kind, not the last.)

Evaluation

Our evaluation method is based on a comparison between the performance of an MSR trained on particular text corpus and semantic relatedness data collected from a large-scale free-association experiment (Nelson et al., 1998). In this experiment, subjects were given a stimulus word, *cue*, and were then asked what word first came to mind, *target*. The target word that first came to mind is considered to be the most semantically related word for that cue, for that participant. More than 6,000 participants produced nearly three-quarters of a million responses to 5,019 cue words.

In order to find out whether the MSRs, trained on the six provided corpora, agreed with human judgments of word relatedness, we checked that the MSRs picked target words for each cue as more relevant to that cue than other random words. To do this, we added a list of n random nouns to the list of n target words for each cue, resulting in a list of $2n$ words (n was limited to a maximum of 5 for cue words that

were associated with more than 5 targets); this list of words was then sorted by MSRs according to word-cue relatedness. If a given MSR perfectly agreed with human judgments, the top *n* words in the sorted list would be all of the human-picked targets for that cue. If half of the targets were found by the MSR to be less relevant to the cue than half of the random words, the MSR performance on that cue would be considered 50%. The average percentage of targets found in the top *n* MSR-sorted words was used as the overall MSR performance score.

Results & Discussion

We evaluated PMI and NSS on the following corpora: Project Gutenberg, Google Groups, Google, Enron, Wikipedia, and New York Times. PMI performed best on the New York Times corpus with an average score of 67.3%. PMI performed the worst on the Project Gutenberg corpus, the massive online collection of books, with an average score of 43.2%. NSS performed best on the Wikipedia corpus with an average score of 65.4%. NSS performed worst on the Project Gutenberg corpus with an average score of 54.2%. A two-factor ANOVA revealed a significant main effect of Corpus, $F(5,25080) = 824.45$, $p < .001$, a significant main effect of MSR, $F(1,5016) = 1094.67$, $p < .001$, and a significant effect of the Corpus by MSR interaction, $F(5,25080) = 229.80$, $p < .001$. PMI's performance showed a high dependence on the text corpus used, while NSS varied less from corpus to corpus.

NSS performed better than PMI on all but the New York Times corpus (mean NSS performance = 60.2%; mean PMI performance = 55.0%), and the overall performances of the two MSRs were highly correlated across the six corpuses (r-square = .82).

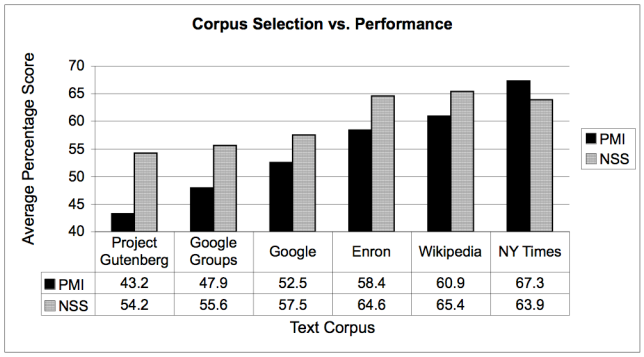

Figure 1. Corpus comparison for PMI and NSS. Standard error bars are too small to be displayed.

We were surprised that the New York Times corpus performed the best out of all the corpora we tested on PMI. It is not nearly as extensive as the Google corpus, nor as structured as Wikipedia, nor does it contain as much conversational English as the Enron Email Corpus or Google Groups. Yet it clearly had the highest score. Also surprisingly, Project Gutenberg, which is a large collection of online books, was the worst of these corpora. These findings have serious implications for the significant portion of MSR research and applications using books as the training corpus.

Our results show that corpus selection has a significant impact on an MSR's performance. One need look no further than at the difference in average scores between PMI using the Project Gutenberg corpus and PMI using New York Times corpus to see this fact. The fact that NSS scores do not vary nearly as much as PMI across different corpus selections indicates the presence of an MSR by corpus interaction effect. Further evidence of this effect lies in the fact that the New York Times corpus, our best corpus for PMI, did not perform as expected on NSS, our best-performing MSR. This MSR by corpus interaction effect is something in need of further investigation.

Another question that inevitably arises is why the Google corpus, which gives access to the World Wide Web as a corpus, is a suboptimal choice. Both of the MSRs that we tested, PMI and especially NSS, were designed to account for the format of the World Wide Web, and rely on its abundance of information (Cilibrasi & Vitanyi, 2007; Turney, 2001). According to our results, however, it appears that both PMI and NSS may be better served by a smaller corpus.

Summary & Conclusions

How "good" a text corpus is for an MSR is not an intuitive matter. We found that the Project Gutenberg corpus, a large collection of books, did a poor job of modeling the human lexicon. Had we been intending to use PMI or NSS in an application such as a cognitive model and had chosen the Project Gutenberg corpus, we would have selected the worst choice possible and our cognitive model's ability to understand text as humans do would have been seriously impaired.

Our study is still ongoing. Rather than evaluating just one or two MSRs trained on a variety of corpora, we would like to test many more MSRs, on many more corpora, using various evaluation techniques (Veksler & Gray, 2006). Ultimately, we would like to find a text corpus that would be the optimal choice for all MSRs. If we knew the optimal choice for a text corpus, when using a semantic relatedness component in ACT-R (Anderson et al., 2004), C-I (Kintsch, 1988), or some other cognitive architecture, we could tell researchers exactly what corpus to train their MSR on. Researchers could rest easy knowing that their semantic relatedness component was performing at the highest level possible. Rather than worrying about the details of their MSR, we hope to allow researchers to be able to focus their attention on the actual MSR-based applications and cognitive models.

Acknowledgments

We would like to thank Stephane Gamard for his contributions to our research. We would also like to thank Dr. Wallace of RPI for providing with the Enron email corpus. The work was supported in part by the Disruptive

Technology Office, ARIVA contract N61339-06-C-0139 issued by PEO STRI. The views and conclusions are those of the authors, not of the U.S. Government or its agencies.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglas, S., Lebiere, C., & Quin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036-1060.

Cilibrasi, R., & Vitanyi, P. M. B. (2007). The Google similarity distance. *Ieee Transactions on Knowledge and Data Engineering, 19*(3), 370-383.

Dumais, S. (2003). Data-driven approaches to information access. *Cognitive Science, 27*(3), 491-524.

Fu, W.-T., & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human-Computer Interaction, in press*.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163–182.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Lemaire, B., & Denhiére, G. (2004). Incremental construction of an associative network from a corpus. In K. D. Forbus, D. Gentner & T. Regier (Eds.), *26th Annual Meeting of the Cognitive Science Society, CogSci2004*. Hillsdale, NJ: Lawrence Erlbaum Publisher.

Lemaire, B., Denhiere, G., Bellissens, C., & Jhean-Iarose, S. (2006). A computational model for simulating text comprehension. *Behavior Research Methods, 38*(4), 628-637.

Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005). *Term representation with generalized latent semantic analysis.* Paper presented at the 2005 Conference on Recent Advances in Natural Language Processing.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. : http://www.usf.edu/FreeAssociation/.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*(2), 127-162.

Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance - II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology, 49*(2), 85-117.

Pirolli, P. (2005). Rational analyses of information foraging on the Web. *Cognitive Science, 29*(3), 343-373.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P. Flach (Eds.), *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)* (pp. 491-502). Freiburg, Germany.

Veksler, V. D., & Gray, W. D. (2006). *Test case selection for evaluating measures of semantic distance.* Paper presented at the 28th Annual Meeting of the Cognitive Science Society, Vacouver, BC.

Veksler, V. D., & Gray, W. D. (2007). *Mapping semantic relevancy of information displays.* Paper presented at the CHI 2007, San Jose, CA.

Veksler, V. D., Grintsvayg, A., Lindsey, R., & Gray, W. D. (Submitted). *A proxy for all your semantic needs*.