# Combining Knowledge and Corpus-Based Measures for Word-to-Word Similarity

**Dan Ştefănescu, Vasile Rus, Nobal B. Niraula and Rajendra Banjade**
Department of Computer Science, Institute for Intelligent Systems, The University of Memphis
{dstfnscu, vrus, nbnraula, rbanjade}@memphis.edu

**Abstract**

This paper shows that the combination of knowledge and corpus-based word-to-word similarity measures can produce higher agreement with human judgment than any of the individual measures. While this might be a predictable result, the paper provides insights about the circumstances under which a combination is productive and about the improvement levels that are to be expected. The experiments presented here were conducted using the word-to-word similarity measures included in SEMILAR, a freely available semantic similarity toolkit.

## Introduction

Semantic similarity is a fundamental concept which appears recurrently in many Natural Language Processing (NLP) tasks. For example, the goal in Information Retrieval (IR) is to find documents that are most relevant to an user query and, most often than not, this actually means finding documents that have *similar* content to the query. Information Extraction implies searching for linguistic constructions that match or are *similar* to certain expert-defined patterns that generalize for certain types of information and have precisely defined semantics. In Machine Translation, the goal is to automatically find the translation of a text from a source language into a target language. This implies that a text must be produced in the target language in such a way that has a *similar* meaning with the source text.

More specifically, two important NLP tasks dealing directly with semantic similarity are Textual Entailment and Paraphrase Identification. These are critical components in many applications such as Intelligent Tutoring Systems (ITS) with natural language interaction. For example, the conversational ITS DeepTutor (Rus et al. 2013a) uses scripts defined by experts to guide students in their construction of solutions to Newtonian physics problems. Such a system assesses the correctness of student answers through sentence level semantic similarity (paraphrase similarity between the student response and the expert response, which is deemed correct). If the student response is semantically similar to the expert response the student response is deemed correct too.

Word-to-word similarity is the foundation on which the semantic similarity between longer texts can be built (Rus et al., 2013b), and therefore, there is a constant push to improve it, so as to match human judgments as close as possible. This is why the team behind the DeepTutor system recently released SEMILAR (Rus et al. 2013b), a semantic similarity toolkit which contains a collection of various known word-to-word and sentence similarity measures and models. This paper presents our experiments on combining SEMILAR's word-to-word similarity measures in an effort to improve the agreement with human judgments.

In the second section of the paper we will review the literature on word-to-word similarity measures. The third section will present SEMILAR and an evaluation of the measures it contains, while the fourth will present our combining experiments and results. The paper ends with a conclusions section.

## Word-to-Word Similarity Measures

Research literature contains an impressively high number of measures for computing word-to-word similarity. Some of the measures, which are considered to be knowledge-based, exploit the structure of semantic networks or ontologies (e.g. is-a hierarchy in Princeton WordNet [PWN; Fellbaum 1998]) and work on distance-based measures on the network's paths (Rada et al. 1989; Lee et al. 1989; Leacock and Chodorow 1998; Wu and Palmer 1994). These can further be improved by using the Information Content of the lowest common subsumer in the hierarchy and corpus statistics (Resnik 1995; Jiang and Conrath

1997; Lin 1998). Moreover, the PWN gloss overlap measure can be used for inferring similarity (Banerjee and Pedersen 2003). Such methods are implemented in the WordNet::Similarity package (Pedersen et al. 2004) and also included in the SEMILAR toolkit (Rus et al. 2013b).

Another category of word-to-word similarity measures relies on a corpus to compute a similarity score. For example, Latent Semantic Analysis (LSA; Landauer et al. 1998), Explicit Semantic Analysis (ESA; Gabrilovich and Markovitch, 2007), or Latent Dirichlet Allocation (LDA; Blei et al. 2003) exploit the distributions of words in large collections of documents. LSA and ESA generate semantic models or spaces in which words are represented as vectors, the values of which being, for instance, weighted frequencies of occurrences within given documents. On the other hand, LDA models documents as topic distributions and topics as distributions over words in the vocabulary. In this case, each word can be represented as a vector encoding its contribution to the LDA generated topics. As such, for all these methods (LSA, ESA and LDA), the similarity between words can be, and usually is, computed in terms of cosine similarity between corresponding vectors.

Another class of methods uses IR engines to gather co-occurrence statistics based on which similarity scores are computed. Pointwise Mutual Information (Bollegala et al. 2007) is typically used in such cases.

One explanation for the existence of so many methods is the fact that instead of being directly defined by a formula, a similarity measure is rather "derived from a set of assumptions about similarity" (Lin 1998). This is very important in distinguishing between the existing various measures and choosing the right one, given a certain task. Lin (1998) states 3 principal general intuitions that should be considered when defining a similarity measure. Given two objects (in our case words) and a clear definition for commonality (often expressed in terms of information content) these are: (i) the more commonality they share, the more similar they are; (ii) the more differences they have, the less similar they are; (iii) maximum similarity is reached when the two items are identical. Many of the existing measures are focusing on the first and the third intuitions, ignoring the second one, but most of the differences between them are largely given by the way in which the commonality is measured.

All the automatic measures for similarity make certain assumptions about commonality, which are consistently preserved no matter the word pairs under scrutiny. On the other side, humans are not always aware of the commonality assumptions they are making, and we speculate that sometimes they even change or merge assumptions depending on the given word pairs, judging similarity from multiple angles. This is why we think that combining different types of similarities would better simulate the way humans think.

## SEMILAR

SEMILAR[1] (Rus et al. 2013b) is a recently released semantic similarity toolkit which comprises a multitude of components for project management, data view browsing-visualization, natural language preprocessing (e.g., collocation identification, part-of-speech tagging, phrase or dependency parsing, etc.), semantic similarity methods for both word and sentence level, classification components (naïve Bayes, Decision Trees, Support Vector Machines, and Neural Network), kernel based methods (sequence kernels, word sequence kernels, and tree kernels), and other functionalities designed to help researchers in the process of choosing the right similarity model that would suit their needs.

Among the available word-to-word similarity measures included in this package we list all the WordNet based similarity measures that are also present in the WordNet::Similarity package (Pedersen et al. 2004) and multiple LSA and LDA semantic models generated based on the Touchstone Applied Science Associates (TASA) corpus (Ivens and Koslin 1991; Landauer et al. 1998). Recently, there have been added new LSA and ESA semantic models constructed on the whole English Wikipedia (a January 2013 version).

We started by evaluating the performances of these measures against human judgment. To do this, we turned to the WordSimilarity-353 Test Collection (Finkelstein et al., 2001), the most used test set for word-to-word similarity in the literature. It contains 353 word pairs along with similarity scores which were manually assigned by more than 13 different subjects. The average values can be seen as good estimators for the similarity between these pairs, and can be used as a gold standard when evaluating different measures against human judgment. Such evaluations are traditionally done by computing Spearman rank-order (*rho*) correlations (Agirre et al. 2009; Gabrilovich and Markovitch 2007), but in what concerns our results, we also provide Pearson (*r*) correlation values. Table 1 presents evaluations on SEMILAR's word-to-word similarity measures compared to other existing measures in the literature.

Interestingly, the computed *r* values show that the results obtained with WordNet methods are more correlated with those obtained using the LSA model built over Wikipedia than the one built over TASA (see Table 2). However, *rho* values seem to indicate the opposite.

[1]Available at http://www.semanticsimilarity.org/

| Method | Pearson | Spearman |
|---|---|---|
| **Results obtained using SEMILAR** | | |
| WordNet based Similarity (WN) | 0.187-0.380 | 0.196-0.381 |
| ESA Wiki | 0.542 | 0.568 |
| Best LSA Wiki Model | **0.589** | **0.603** |
| Best LSA TASA Model | 0.576 | 0.591 |
| Best LDA TASA Model | 0.345 | 0.326 |
| **Results reported in other papers** | | |
| PWN (Jarmasz, 2003) | | 0.33-0.35 |
| Roget's Thesaurus (Jarmasz, 2003) | | 0.55 |
| LSA (Finkelstein et al., 2002) | | 0.56 |
| Wikipedia (Strube & Ponzetto, 2006) | 0.19-0.48 | |
| ESA Wiki (Gabrilovich & Markovitch, 2007) | | 0.75 |
| ESA ODP (Gabrilovich & Markovitch, 2007) | | 0.65 |
| PWN 3.0 (Agirre et al., 2009) | | 0.56 |
| PWN 3.0 + glosses (Agirre et al., 2009) | | 0.66 |
| Context Windows (CW) (Agirre et al., 2009) | | 0.57-0.63 |
| Bag of Words (Agirre et al., 2009) | | 0.64-0.65 |
| Syntactic Vectors (Syn) (Agirre et al., 2009) | | 0.55-0.62 |
| CW + Syn (Agirre et al., 2009) | | 0.48-0.66 |
| SVM (Agirre et al., 2009) | | 0.78 |

*Table 1: Comparison between the results obtained using SEMILAR and those reported by others (in terms of correlation with humans)*

## Combining Measures

Given the multitude of word-to-word similarity measures available in SEMILAR, a logical step would be to combine them into a better function. We consider this to be a very useful feature that should be available in the SEMILAR toolkit. It has been shown that in order to obtain better results via the combination of two or more methods, two conditions must be satisfied (Dieterich, 1998): (i) the methods should be different but, (ii) they should have comparable performance scores. In other words, their overall performance should be comparable, while making different mistakes. Some of the methods in the toolkit meet these requirements. Table 2 shows the correlations among the measures selected for our experiments. It is important to observe that all WordNet (WN) measures have a relatively low correlation with human judgment. This is why we selected for our experiments only one WN measure, the one having the best correlation with humans (i.e. a Lesk type measure [Banerjee and Pedersen 2002]).

The most straightforward way to combine the available measures is by fitting linear regression models. The results obtained using two by two combinations, are presented in Table 3.

Looking at the values in Table 1, one can see that the correlation between humans and the best LSA Wiki model is 0.60 (*rho*), while the correlation between humans and LSA TASA model is 0.59 (*rho*). But, the correlation between the two LSA models is 0.60 (*rho*). In other words, the efficiency of the two models is comparable, but they are not making so many different mistakes. Consequently, we could guess that combining the two would increase the agreement versus humans, but not by much. The values in table 3, confirm this intuition: the correlation of the combined model is 0.65 (*rho*).

| | LSA TASA | LDA | ESA | WN |
|---|---|---|---|---|
| **LSA Wiki** | 0.594<br>0.600 | 0.354<br>0.249 | 0.599<br>0.633 | 0.304<br>0.224 |
| **LSA TASA** | | 0.635<br>0.374 | 0.585<br>0.584 | 0.282<br>0.248 |
| **LDA** | | | 0.407<br>0.249 | 0.278<br>0.168 |
| **ESA** | | | | 0.314<br>0.292 |

*Table 2: Pearson (top) and Spearman (bottom) correlations among the selected measures*

On the other hand, the low correlation values between WN and the other methods show that they are making almost completely different mistakes. Although this can show a high potential for combinations, we should be tempered by the fact the WN methods have a low correlation score comparing to humans.

| | LSA TASA | LDA | ESA | WN |
|---|---|---|---|---|
| **LSA Wiki** | **0.632**<br>0.646 | 0.591<br>0.623 | 0.643<br>0.658 | 0.621<br>0.647 |
| **LSA TASA** | | 0.559<br>0.576 | 0.640<br>0.661 | 0.618<br>0.642 |
| **LDA** | | | 0.594<br>0.637 | 0.479<br>0.501 |
| **ESA** | | | | 0.627<br>**0.672** |

*Table 3: Correlations with humans obtained by linearly combining the selected measures, two by two (Pearson – top; Spearman - bottom).*

Table 3 presents the results obtained by combining the selected measures two by two. For example, the correlation values between the measure that combines the two LSA measures and human judgment are 0.632 (*r*) and 0.646 (*rho*) (see the top left cell). The values in Table 3 are obtained by conducting a ten-fold cross-validation: we selected 10 consecutive chunks of 31 word-pairs as test sets, and used the rest for training. Moreover, we investigated all the possible linear combinations between subsets of these selected methods. The best combinations of 3 measures are LSA Wiki + ESA + WN, with correlation values of 0.673 (*r*) and 0.704 (*rho*) and LSA TASA + ESA + WN with 0.673 (*r*) and 0.712 (*rho*). The best results can be obtained by combining LSA Wiki, LSA TASA, ESA and WN, with correlation values of **0.676** (*r*) and **0.723** (*rho*).

Looking at the measures that produce the best result, we realize that the observed increase in correlation should be expected, because the different measures involved are capturing different similarity aspects that can occur between

words. For example, WN-based measures cannot capture the co-occurrence facet of similarity but, in turn, this is properly addressed by the LSA and ESA. On the other hand, the latter models have no way of exploiting the similarity given by the generality or specificity encoded into an ontological hierarchical structure. This is clearly showing the potential of combining Knowledge and Corpus-based measures for word to word similarity.

## Conclusions

Word-to-word similarity is the foundation on which semantic similarity measures for longer texts (i.e. sentences, paragraphs, even documents) are built. It is therefore important that such measures are as close as possible to humans when assessing word-to-word similarity. This paper shows that a good way for achieving this objective is to combine measures that are capturing different types of similarity. Knowledge-based measures are exploiting semantic relationships in ontologies to judge similarity, while corpus-based ones rely on co-occurrence statistics to do it. Combining the two is a natural way to capture similarity on multiple facets and our experiments confirmed that this is a good strategy to get a better agreement with human judgment. Such conceptually different measures are bound to make different mistakes. If they also have similar performances, the ideal requirements for a productive combination are met.

In the near future we will continue our experiments on combining word-to-word similarity measures, also considering methods such as weighted average or bagging.

## Acknowledgements

This research was supported in part by Institute for Education Sciences under awards R305A100875. Any opinions, findings, and conclusions or recommendations expressed in this material are solely the authors'.

## References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In Proceedings of HLT: The 2009 Annual Conference of the NAACL, 19–27.

Banerjee, S., and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness, in: IJCAI, 3: 805–810.

Blei, D., M., Ng, A., Y., and Jordan, M., I. 2003. Latent Dirichlet Allocation, the Journal of ML research 3: 993–1022.

Bollegala, D., Matsuo, Y., and Ishizuka, M. 2007. Measuring semantic similarity between words using web search engines. www.

Dietterich, T. G. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 10(7): 1895–1924.

Fellbaum, C. 1998. WordNet: an electronic lexical database. Cambrige, MIT Press, Language, Speech, and Communication.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. 2001. Placing search in context: The concept revisited. In Proceedings of the 10th international conference on World Wide Web, 406–414. ACM.

Gabrilovich, E., and Markovitch, S. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In IJCAI 7: 1606–1611.

Islam, A., and Inkpen, D. 2006. Second order co-occurrence PMI for determining the semantic similarity of words. In Proceedings of the International Conference on Language Resources and Evaluation, Genoa, Italy, 1033–1038.

Ivens, S. H., and Koslin, B. L. 1991. Demands for Reading Literacy Require New Accountability Methods. Touchstone Applied Science Associates.

Jarmasz, M. 2012. Roget's thesaurus as a lexical resource for natural language processing.arXiv preprint arXiv: 1204.0140.

Jiang, J., J., and Conrath, D., W. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In International Conference on Research on Computational Linguistics, pages 9008+.

Landauer, T. K., Foltz, P. W., and Laham, D. 1998. An introduction to latent semantic analysis. Discourse processes, 25(2-3): 259–284.

Leacock, C., and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., ed., WordNet: An electronic lexical database. MIT Press, 265–283.

Lee, J. H., Kim, M. H., and Lee, Y. J. 1989. Information retrieval based on conceptual distance in is-a hierarchies. Journal of Documentation 42(2): 188–207.

Lin, D. 1998. An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98), Jude W. Shavlik (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 296–304.

Pedersen, T., Patwardhan, S., and Michelizzi, J. 2004. WordNet::Similarity: measuring the relatedness of concepts. In Demonstration Papers at HLT-NAACL 2004, 38–41. ACL.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. 1989. Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics, 19(1): 17–30.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In Chris S. Mellish (Ed.). Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI'95) (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA) 1: 448–453. doi: 10.1.1.41.6956.

Rus, V., D'Mello, S., Hu, X., and Graesser, A. 2013a.Recent Advances in Conversational Intelligent Tutoring Systems. AI Magazine, 34(3): 42–54.

Rus, V., Lintean, M., Banjade, R., Niraula, N., and Ştefănescu, D. 2013b. SEMILAR: The Semantic Similarity Toolkit. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, August 4–9, 2013, Sofia, Bulgaria.

Strube, M., and Ponzetto, S. P. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In AAAI 6: 1419–1424.

Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In 32nd Annual Meeting of the Association for Computational Linguistics, 133–138.