SEMANTIC MEASURES BASED ON WORDNET USING MULTIPLE INFORMATION SOURCES

Mamoun Abu Helou and Adnan Abid
Dipartimento di Elettronica ed Informazione, Politecnico di Milano, Milan, Italy

Keywords: Semantic similarity, Lexical database, Information content, Corpus statistics, Word gloss.

Abstract: Recognizing semantic similarity between words is a generic problem for many applications of computational linguistics and artificial intelligence, such as text retrieval, classification and clustering. In this paper we investigate a new approach for measuring semantic similarity that combines methods of existing approaches that use different information sources in their similarity calculations namely, shortest path length between compared words, depth in the taxonomy hierarchy, information content, semantic density of compared words, and the gloss of words. We evaluate our measure against a benchmark set of human similarity ratings and the results show that our approach demonstrates better semantic measures as compared to the existing approaches.

1 INTRODUCTION

Similarity between words is often represented by similarity between concepts associated with the words (Li, Bandar, and McLean, 2003). Nowadays, the need to determine the degree of semantic similarity, or more generally relatedness between two lexically expressed concepts is a problem that pervades much of computational linguistics. The problem of formalizing and quantifying the intuitive notion of similarity has a long history in philosophy, psychology, and artificial intelligence; therefore, many different perspectives have been suggested (Budanitsky and Hirst, 2006).

The existing work provides a strong base for semantic relatedness; however, it is unclear how to assess the relative and absolute merits of the many competing approaches that have been proposed. Generally, these approaches can be classified into; *edge counting-based* methods, *corpus based* methods, and the *gloss based* methods. This paper explores the idea of joining the forces of these three categories of information sources for computing semantic similarity based on WordNet. We combine these lines of research by combining the different information sources in one metric.

Finally, our similarity measure is evaluated against the available benchmark set of human similarity ratings, and the results demonstrate that our proposed approach improves the similarity measures considerably.

The rest of the paper is organized as follows: section 2 briefly discusses the relevant literature about the semantic measures; section 3 presents the proposed approach; whereas the experiments and evaluation of the results are presented in section 4, which is followed by the conclusion.

2 SEMANTIC SIMILARITY BASED ON WORDNET

Given two words, *w1* and *w2*, the semantic similarity *sim*(*w1*; *w2*) can be calculated through the analysis of a lexical knowledge base, e.g. using WordNet which is developed at Princeton by a group led by Miler (Miller, 1995) it is an online semantic dictionary, partitioning the lexicon into nouns, verbs, adjectives, and adverbs. We apply well established semantic similarity measures originally developed for WordNet. The measures we use for computing semantic similarity fall into three broad categories.

Edge Counting-based Measures: These measures compute similarity as a function of the number of edges in the taxonomy along the path between two conceptual nodes. The simplest path-based measure is the straightforward edge counting method of (Rada, Mili, Ellen, Bicknell, and Blettner, 1989). (Leacock and Chodorow, 1998) *lch* proposes a normalized path-length measure which takes into

Abu Helou M. and Abid A..
SEMANTIC MEASURES BASED ON WORDNET USING MULTIPLE INFORMATION SOURCES .
DOI: 10.5220/0003101905000503
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 500-503
ISBN: 978-989-8425-28-7

account the depth of the taxonomy at which the concepts are found. (Wu and Palmer, 1994) *wup,* on the other hand, presents a scaled measure which takes into account the depth of the nodes together with the depth of their least common subsumer, *lcs*.

Corpus based Measures: The measure of (Resnik, 1995) *res* computes the similarity between the concepts as a function of their information content, given by their probability of occurrence in a corpus. (Jiang and Conrath, 1997) *JC* approach also uses the notion of information content, but in the form of the conditional probability of encountering an instance of a child-synset given an instance of a parent synset. (Lin, 1998) *lin* similarity measure uses the same elements as *JC*, but in a different fashion.

Gloss based Measures: This measure determines similarity between two words as a function of text (i.e. gloss) overlap. *Lesk* algorithm (Lesk, 1986) uses dictionary definitions (gloss) to disambiguate a polysemous word in the context of a sentence; mainly by counting the number of words that are shared between two glosses. An example of such approach is the extended gloss overlap measure of (Banerjee and Pedersen, 2003).

As introduced above, different methods use different information sources; thus, result in different levels of performance. The commonly used information sources in previous similarity measures are shortest path length between compared words, depth in the taxonomy hierarchy, information content, semantic density of compared words, and the word definition (gloss). A major problem with these similarity measures is that either information sources are directly used as a metric of similarity, or a method uses a particular information source without considering the contribution of others. However, as semantic similarity is influenced by a number of information sources that are interlaced with each other, we argue that semantic similarity depends not only on multiple information sources, but also that the information sources should be properly processed and combined similarly like (Li et al, 2003), they observe that the similarity measure can be improved by a suitable combination using the first two information sources categories, And we think the third one has its impact in computing the similarity if it is augmented to the other information sources.

3 NEW SEMANTIC MODEL

The idea is to develop a strategy that accumulates the advantages of all the aforementioned approaches by combining them in order to measure the semantic similarity between two words. We believe that, to achieve a good similarity measure, all of the information sources should be taken into account. After a careful analysis we consider that all three approaches use different methodology which are orthogonal to one another, therefore, we present our approach that focuses on combining these approaches in a single, thus, rather comprehensive approach in such a way that they augment one another. For that reason, the similarity, $S(\mathrm{w}_1, \mathrm{w}_2)$, between two words w_1 and w_2 can be defined as follows:

$$S(w_1, w_2) = f(d, ic ,g) \tag{1}$$

Where *d* is the edge counting-based method, *ic* is the information content (corpus based) method, and *g* is the gloss based method. Thereby, we try to find the best combination between available semantic measures that cover the three categories and by assigning the proper contribution to each measure which relate to each specific category, Table 1 shows the three categories in the first column and the semantic measure in the second column. Accordingly, our semantic function is the following:

$$f(d, ic ,g) = \alpha *(d) + \beta *(ic) + \gamma *(g) \tag{2}$$

We find and choose the best combination from all the variants, shown in Table 1, of the three groups of information sources by finding the best correlation while optimizing the values of coefficients *α, β* and *γ,* separately for each combination. Eventually, we find that **[***wup,res,lesk***]** is the best combination of all variants and use it on test data.

4 EXPERIMENTS

In order to investigate the effectiveness of our approach we carried out the experiments in two steps. Firstly, tuning the coefficients and finding the best combination using the training data set; and then, using the identified optimal coefficient and the best combination to calculate semantic similarity for word pairs in the testing data sets.

4.1 Data

There is a clear lack of standards for evaluation of lexical similarity. So the quality of a computational method for calculating word similarity can only be established by investigating its performance against human common sense; (Rubenstein and Goodenough, 1965) R&G 65 word pair, (Miller and Charles, 1991) *M&C* 30 word pair, and the full list of the *Word Similarity 353 Test Collection* (*353-TC*) (Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman, and Ruppin, 2002). We used *M&C* data set as a training data set with *R&G* and Word Similarity *353-TC* as a testing data set.

4.2 Tuning

We conducted the experimentation over seven similarity measures which cover the three information source categories, and we search for the suitable parameters α, β, and γ in order to assign the weight to the respective measure in the combined similarity result. The training data set is used to explore the role of α, β, and γ. And to find the semantic measure which represent each category in the similarity function *f (d, ic, g).*The interval of α, β, and γ is (0, 1], and α+ β+ γ =1. For each combination of variants we maximize the correlation result with discrete interval of 0.01 for α, β, and γ. The parameters resulting in the greatest correlation coefficient are considered as the optimal parameters. Finally, the identified optimal combination along with its coefficients is used to calculate semantic similarity for word pairs in the test data sets.

4.3 Evaluating Result

Table 1 shows the performance of similarity measures and the proposed approach in this paper. The correlation data is computed using Java WordNet::Similarity (Hope, 2008), and WordNet (3.0). The second column is the correlations with training data set *M&C* experiment, the correlations with *R&G* experiment, and the correlations with the *353-TC* experiment are listed in the third and the fourth column respectively. Best performance for data set is highlighted in bold, the strongest correlation of the new method reported in the last row which outperforms the individual subjects. It is worth to note that the coefficients might be reported based on only 28 out of the 30 *M&C* pairs because of a noun missing from an earlier version of WordNet. Moreover, nine pairs out of the *353-TC* data set containing at least one word not present as noun in WordNet, thus we remove them from the dataset. As a result of our experiment the similarity measure *wup*, *res*, and *lesk* measures will represent the edge counting based, the corpus based, and the gloss based categories, respectively. Therefore, our similarity function is defined as follow:

$$f(\, d,\, ic\, ,g) = \alpha\, {*}(d) + \beta{*}(ic) + \gamma{*}(g)$$

Where,
$\alpha{*}(d) = \alpha{*}Sim_{wup}$, $\beta{*}(ic) = \beta{*}Sim_{res}$, and $\gamma{*}(g) = \gamma{*}Sim_{lesk}$. S_i is the synset of word *i*, the *ic* is the information content, the *lcs* is the lowest common subsumer, the normalized value of *res* measure is Sim_{Resnik} *(s1,s2)*, and Sim_{lesk} is (Hope, 2008) implementation of the extended gloss overlap measure of (Banerjee et al, 2003).

Table 1: The correlation coefficients of different semantic measures along with human judgment data sets.

Information Source Categories	**Similarity Measure**	**Correlation**		
		M&C	***R&G***	***353-TC***
Edge Counting Based	*path length*	0,755	0,784	**0,385**
	lch	0,779	**0,839**	0,348
	wup	0,765	0,804	0,298
Corpus Based	*jc*	0,742	0,704	0,242
	res	**0,818**	0,834	0,376
	lin	0,739	0,726	0,301
Gloss Based	*lesk*	0,755	0,762	0,374
f(d, ic ,g)	**[***wup,res,lesk***]**	**0,839**	**0,856**	**0,398**

The best weight for the three parameters α, β, and γ; the contribution of the three category measures in the new similarity function, is ***0.2, 0.52*** and ***0.28*** the weight of *wup*, *res*, and *lesk* measures respectivly. All experimental data and results are avilable at http://home.dei.polimi.it/abuhelou/data.

4.4 Discussion

We can notice that WordNet performs extremely well on the small datasets *M&C* and *R&G*, its performance drastically decreases when applied to a larger dataset such as *353-TC*. This is not due to coverage, as in the *353-TC* dataset there are only 9 pairs containing at least one word not present as noun in WordNet (3.0). (Strube and Ponzetto, 2006) suggest that the problems seem to be caused rather by sense proliferation. In their experiment (Li et al, 2003), the best performance was obtained when they combined the shortest path length and the depth of subsumer nonlinearly with correlation coefficient of 0.8914. But they used the 28 word pairs common between *M&C* and *R&G* as their testing set, while the reaming 37 word pairs of R&G has been used as a training set. For a future work we are looking to repeat our experiment using the same data set division for better comparison.

The computational time for our approach is obviously more than the individual measures,

Table 2: Comparison of Time taken among the proposed and existing approaches.

Similarity Measures	*lesk*	*res*	*wup*	*sum*	*f(d,ic,g)*
Time(s)	60,984	0,0625	0,1125	61,159	61,259

which merge three individual measures; however the better result we attain can justify this cost. Table 2 shows the time for the individual measures compared with the new model. For the sake of fairness, we run the experiment 10 times and take the average response time for each measure, so we can notice that 100 ms is the extra cost that we pay to gain more accurate measure comparing it with the sum of the three measure, while we pay 275 ms comparing it with *lesk* measure which is the time consuming measure.

5 CONCLUSIONS

In this paper, we have introduced a new model to identify the similarity between words using WordNet. This model combines existing methods for semantic similarity calculation and finds a combination of three methods each from a different category of information sources. We argue that, in order to achieve better similarity measures all the information sources; shortest path length between compared words, depth in the taxonomy hierarchy, information content, semantic density of compared words, and the gloss definition of the words should be taken into account. We evaluate our method on widely used benchmarking datasets, such as *M&C* dataset, *R&G* dataset, and *353-TC*. The experimental results prove our assumption and fit particularly well in simulating human judgment on semantic similarity between words. In future work, we intend to use this similarity measure in real world applications such as word sense disambiguation.

ACKNOWLEDGEMENTS

This research is part of the “Search Computing” (SeCo) project, funded by the European Research Council (ERC), under the 2008 Call for “IDEAS Advanced Grants”, dedicated to frontier research.

REFERENCES

Banerjee, S., and Pedersen, T., 2003. *Extended gloss overlaps as a measure of semantic relatedness*. In Proceedings on Artificial Intelligence, Eighteenth International Joint Conference, Acapulco, pp.805–810.

Budanitsky, A. and Hirst, G., 2006. *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E., 2002. *Placing Search in Context: The Concept Revisited.* ACM Transactions on Information Systems, 20(1):116-131.

Hope, D., 2008. Java WordNet::Similarity http://www.cogs.susx.ac.uk/users/drh21/

Jiang, J. J. and Conrath, D., 1997. *Semantic similarity based on corpus statistics and lexical taxonomy*. In Proceedings of International Conference on Research in Computational Linguistics, Taiwan.

Leacock, C., and Chodorow, M., 1998. *Combining local context and WordNet similarity for word sense identification*. pp. 265–283.

Lesk, M., 1986. *Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone*. In Proceedings, Fifth International Conference on Systems Documentation (SIGDOC ’86), pages 24–26.

Li, Y., Bandar, Z. A., and McLean, D., 2003. *An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources*, IEEE Transactions on knowledge and & engineering, 15(4).

Lin, D., 1998. *An information-theoretic definition of similarity*. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI.

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

Miller, G. A., 1995. *WordNet: a lexical database for English*. Communications ACM, 38(11), pp. 39-41.

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

Resnik, P., 1995. *Using information content to evaluate semantic similarity*. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, Montreal.

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

Strube, M., and Ponzetto, S., 2006. *WikiRelate! Computing Semantic Relatedness Using Wikipedia*, American Association for Artificial Intelligence.

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