Measuring Word Semantic Relatedness Using WordNet-Based Approach

Tingting Wei[1], Huiyou Chang[2*]

[1] Sun Yat-Sen University/Information Science and Technology, Guangzhou, China.
[2] Sun Yat-Sen University/Software, Guangzhou, China.

* Corresponding author. Email: isschy@mail.sysu.edu.cn
Manuscript submitted January 10, 2015; accepted May 8, 2015.
doi: 10.17706/jcp.10.4.252-259

Abstract: Word semantic relatedness measure plays an important role in many applications of computational linguistics and artificial intelligence. In recent years the measures based on WordNet have shown its talents and attracted great concern. Many measures have been proposed to achieve the best expression possible for the degree of semantic relatedness of words. In this paper, we consider two different modified measures for computing the semantic relatedness between two words based on the path-based approach. The first measure introduces the maximum node path into the classical path-based method to compute the relatedness of words from ontology hierarchy; it mainly exploits edge-counting technique. The second one takes the definition and semantic relationships of synsets into account; it is based on the assumption that the explicit and implicit semantic relationships between synsets impose equally importance factors in the word relatedness measure. The experimental results using the proposed methods on common datasets show that our measures yields into better levels of performance compared to several classical methods. In addition, the second approach performed better than the first one.

Key words: Word semantic relatedness, WordNet, semantic relationships.

1. Introduction

Word semantic relatedness measure plays an important role in many applications of computational linguistics and artificial intelligence such as information retrieval, word sense disambiguation [1], text clustering [2] and so on. WordNet [3], which is one of the most widely used thesauruses for English, has shown its talents and attracted great concern. Many measures have been proposed to achieve the best expression possible for the degree of semantic relatedness of words based on WordNet. In general, all the measures can be grouped into four classes: path length based measures, information content based measures, feature based measures, and hybrid measures. An exhaustive overview of these approaches can be found in [4].

In this paper, we propose two different modified measures for computing the semantic relatedness between two words based on the path-based approach. Previous works have showed that exploiting the structural information of WordNet can improve the accuracy of word semantic relatedness measurement; therefore, the proposed measures mainly exploit structural aspects of WordNet. The first measure is a combination of node position and path. It exploits edge-counting technique. On the other hand, as the effects of adding textual data to structural information are still not very extensively researched, we propose the second approach that integrates the structural (taxonomic parameters) with the textual (gloss) of WordNet.

It is based on the assumption that the explicit and implicit semantic relationships between synsets impose equally importance factors in the word relatedness measure. We focus on exploring if the combination of the structural information and the glosses of synsets can provide a more accurate assessment for the semantic relatedness between words in our study.

The rest of the paper is organized as follows: Section 2 reviews some related works. Section 3 presents the modified measures based on WordNet. In Section 4, we describe the experiments that evaluate our methods and the analysis of results. Finally, we conclude this work and show its implications in Section 5.

2. Related Works

The semantic relatedness quantification is based on lexical resources by exploiting the knowledge existing inside these resources. Some of the most popular measures are implemented and evaluated using WordNet as the underlying ontology. In order to understand our work better, some relevant works related to our interests will be introduced and the limitations of the described approaches will be presented as well.

2.1. WordNet

WordNet is one of the most widely used and largest lexical databases of English. In general as a dictionary, WordNet covers some specific terms from every subject related to their terms. WordNet interlinks not just word forms strings of letters, but specific senses of words. It maps all the stemmed words from the standard documents into their specifies lexical categories. In our study the WordNet 2.1 is used which contains 155,327 terms, 117,597 senses, and 207,016 pairs of term-sense. It groups nouns, verbs, adjectives and adverbs into sets of synonyms called synsets. The synsets are organized into senses, giving thus the synonyms of each word, and also into hyponym/hypernym (i.e. is-a), and meronym/holonym (i.e. part-of) relationships, providing a hierarchical tree-like structure for each term. Strictly speaking, the noun synsets consist of eleven separate hierarchies covering distinct conceptual and lexical domains. Meanwhile, it tries to establish different standards between concepts and convey different semantic relationships. The lower hierarchy concept reserves the overall property of the higher hierarchy concept. And thus the abstract concept has turning into concept tree and one can carry out connectional reasoning and calculation. WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity [5].

2.2. Path-Based Measures

Following the cited overview [4], we focus on measures that are related to our work.

Rada *et al.* [6] proposed an approach based on MeSH ontology to improve text retrieval. It computed semantic similarity straightforwardly in terms of the number of edges between terms in the hierarchy. Their assumption of this approach is that the number of edges between terms in ontology is a measure of conceptual distance between terms. Wu and Palmer [7] defined a measure of similarity between concepts based on path lengths (in number of nodes), common parent concepts, and distance from the hierarchy root. Leacock and Chodorow [8] proposed a metric based on the count of link numbers between two set of terms or synonyms representing the same concept, and Jarmasz and Szpakowicz [9] used the same approach with Roget's Thesaurus while Hirst and St-Onge [10] applied a similar strategy to WordNet.

In this paper we utilize the Wu and Palmer measure and take into account the glosses of terms for quantifying word semantic relatedness. Some of the above described metrics also have been implemented for a comparison with our measure.

2.3. Word Sense Disambiguation

Word sense disambiguation (WSD) is a process that replacing the original terms in a document by the

most appropriate sense as dictated by the surrounding context of a document. Typically, many semantic similarity measures are used for calculating the relatedness among senses. Early work varied between counting word overlaps between definitions of the word [11]-[14] to finding distances between concepts following the structure of the LKB [15]. As an alternative, graph-based methods have gained much attention in recent years [16]-[21]. Graph-based techniques are performed over the graph underlying a particular knowledge base; they first consider all the sense combinations of the words in a given context and then try to search for the relations among senses based on the whole graph. The main disadvantage of graph-based methods is their computational expense [21].

3. Improving Word Semantic Relatedness Using WordNet

In this section, we first describe related semantic relatedness measures in detail and then lead to the connected use of these methods in our measures.

3.1. Traditional Semantic Relatedness Measures

Wu and Palmer [6] computed the similarity between two senses by finding the least common subsumer (LCS) node that connects their senses. For example, we can see from the red rectangle boxes in Fig. 1, the LCS of **canine** and **chap** is the lowest common node between the paths of these two senses from the root of WordNet hierarchy, **organism**.

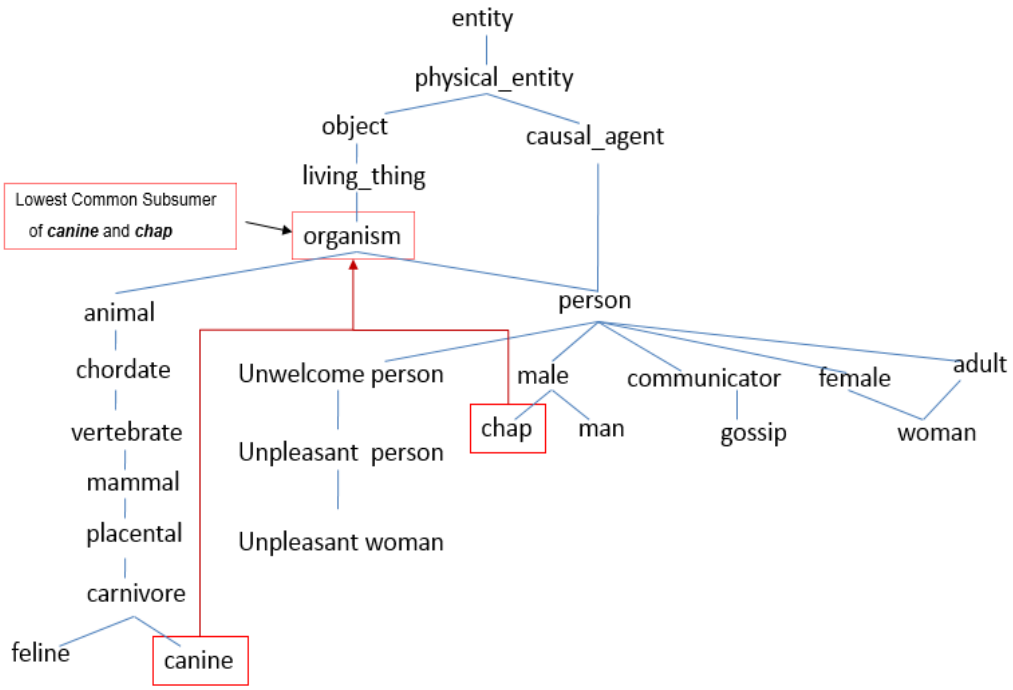

Fig. 1. A sample WordNet hierarchy.

Once the LCS has been identified, the distance between two senses is computed by

$$\delta_{\text{Wu\_Palmer}}(c_p, c_q) = \frac{2d}{L_p + L_q + 2d}. \quad (1)$$

where d is the depth of the LCS from the root, L_p is the path length between c_p and LCS, and L_q is the path length between c_q and LCS.

However, this measure is based only on the explicit semantic relations that assuming the links between concepts represent distances; but such links do not cover all possible relations between synsets. For

example, WordNet encodes no direct link between the synsets ***car*** and ***tire***, although they are clearly related.

Thus, different from Wu-Palmer measure, Banerjee and Pedersen [14] presented a new measure of semantic relatedness between concepts that is based on the number of shared words (overlaps) in their definitions (glosses). When measuring the relatedness between two input synsets, this method not only looks for overlaps between the glosses of those synsets, but also between the glosses of the hypernym, hyponym, meronym, holonym and troponym synsets of the input synsets, as well as between synsets related to the input synsets through the relations of attribute, similar–to and also–see. For purposes of illustration, we introduce two definitions in terms of Banerjee and Pedersen (2003).

Definition 1 (Description of a synset)

Let $C=\{c_1, c_2,\ldots,c_k\}$ be the set of synsets in a document, $c_i \in C$. Let Lemma (c_i) be the set of words that constitute a synset c_i. Let Gloss (c_i) be the definition and examples of usages of c_i. Let Related (c_i) be the union of the hypernym, hyponym, meronym, holonym and troponym synsets of c_i, as well as synsets related to the c_i through the relations of attribute, similar–to and also–see. Then the description of c_i is defined as

$$\mathrm{DES}(c_i) = (\mathrm{Lemma}(c_i) \cup \mathrm{Gloss}(c_i) \cup \mathrm{Gloss}(\mathrm{Related}(c_i))) \cap \neg \mathrm{stopwords}. \quad (2)$$

In this study, we add the lemma of synset to the description set on the basis of Banerjee and Pedersen (2003).

Based on Definition 1, we now define the scoring function of similarity.

Definition 2 (Similarity between two synsets)

Let $C=\{c_1, c_2,\ldots,c_k\}$ be the set of synsets in a document. Given two sets of description of synsets c_i and c_j (c_i, $c_j \in C$), which denoted by DES (c_i), DES (c_j) respectively, and they can be regarded as two input strings; the longest overlap between these two strings is detected first, then removed and in its place a unique marker is placed in each of the two input strings; the two strings thus obtained are then again checked for overlaps, and this process continues until there are no longer any overlaps between them. Let N be the number of continuous words that overlapped, and let k be the number of iterations they had detected. Then the similarity between two synsets is defined by

$$\mathrm{Score}(\mathrm{DES}(c_i), \mathrm{DES}(c_j)) = \sum_k N^2. \quad (3)$$

The score mechanism assigns an N continuous words overlapped the score of N^2, which gives an N–word overlapped a score that is greater than the sum of the scores assigned to those N words if they had occurred in two or more phrases, each less than N words long. This measure next assigns each possible sense a score by some other mechanisms; and sense with the highest score is judged to be the most appropriate sense for the target word.

The measure of Banerjee assumes that synsets description pair with more common words and less non-common words are more similar. However it can't work well when there is not an overlap description set.

3.2. Improved Semantic Relatedness Measures

3.2.1. Path-only based measure

A simple method to calculate the similarity between two concepts in ontology is to count the number and directions of edges between two concept nodes, as used in [6]. However, without considering the levels of nodes in the ontology hierarchy, the similarity of this method could be misleading. We define a modified measure based on Wu and Palmer [6] method that takes into account of positions of concept nodes in the ontology.

$$\delta_{path\_only}(c_p, c_q) = \frac{2d + S_1}{L_p + L_q + 2d + S_1} \tag{4}$$

$$S_1 = \log \frac{2 * \max(f(p)) * \max(f(q))}{\max(f(p)) + \max(f(q))} \tag{5}$$

where d is the depth of the LCS, L_p and L_q are the path lengths between semantic node c_p, c_q and the LCS respectively, max($f(p)$) is the maximum path length of semantic node c_p, max($f(q)$) is the maximum path length of semantic node c_q, S_1 is the logarithm of the harmonic mean between max($f(p)$) and max($f(q)$).

Formula (4) represents that the distance between any two concept nodes c_1 and c_2 is reduced if their parent node lies in the lower level of the hierarchy.

3.2.2. Combination of path and glosses measure

In the previous section, we have described the traditional term similarity measures; however, they have various weaknesses. For example, WordNet provides explicit semantic relations between synsets, such as through the is–a or has–part links, but links do not cover all possible relations between synsets; while overlaps provide evidence that there is an implicit relation between those uncovered synsets. Thus, we define an improved term similarity measure with ontology that takes into account the explicit and implicit semantic information in the ontology. The improved measure is a combination of the Wu and Palmer measure [7] and the Banerjee and Pedersen measure [14], which is defined as

$$\delta_{path\&gloss}(c_p, c_q) = \frac{2d + S_2}{L_p + L_q + 2d + S_2}. \tag{6}$$

where S_2=log (Score (DES (c_p), DES (c_q))+1), and the other parameters are similar with formula (1) and (2). This method not only reflects structure information of synsets, such as distance, but also incorporates content meaning of synsets in the ontology. It integrates well with explicit and implicit semantic between synsets in ontology, leading to better performance, and which can be verified in our later comparative experiments.

4. Experiments Results

We test four methods on three data sets R&G (65 pairs) [22], M&C (30 pairs) [23] and WordSim353 (353 pairs) [24]. These three data sets for the English word similarity task have been extensively used in the past. They consist of pairs of words accompanied by average similarity scores assigned by human subjects to each pair. The correlation with human judgments of similarity is measured using the Spearman rank correlation coefficient ρ as well as the Pearson correlation coefficient r between the computation values for each pair and the human judgments. The results are shown in Table 1, and its corresponding intuitive representation as shown in Fig. 2 and Fig. 3.

Table 1. A Comparison of Several Word Relatedness Measures

Dataset	W&P [6]		B&P [13]		Path_only		Path&Glosses	
	ρ	r	ρ	r	ρ	r	ρ	r
M&C (30 pairs)	0.512	0.536	0.602	0.635	0.551	0.579	**0.632**	**0.681**
R&G (65 pairs)	0.468	0.507	0.571	0.624	0.514	0.542	**0.619**	**0.692**
WordSim353 (353 pairs)	0.24	0.271	0.35	0.383	0.267	0.284	**0.387**	**0.481**

Path_only is our proposed method that takes into account of positions of concept nodes in the ontology. Path&Glosses is our proposed method that combines path and glosses of synsets. The improvement is

significant according to the paired-sample t-test at the level of $p < 0.05$. As expected, we can see that using the path and glosses (Path&Glosses) of synsets tend to generate a higher correlation with the human rating than the path only (W&P, Path_only) and glosses only (B&P). This means that the explicit and implicit semantic relations together can reveal hidden similarity between terms, when we add the gloss to the path-based method, it pays a positive role to the correlation. From the two Figures, we found that the method using glosses only consistently performs better than the other two path only measures, which indicates that glosses of synsets give more information to compute the relatedness between words. Even so, our modified method Path_only performed better than W&P approach without considering node location information. This experimental result indicates that the two modified measures produce the results which are closer to the human ratings compared to classical methods.

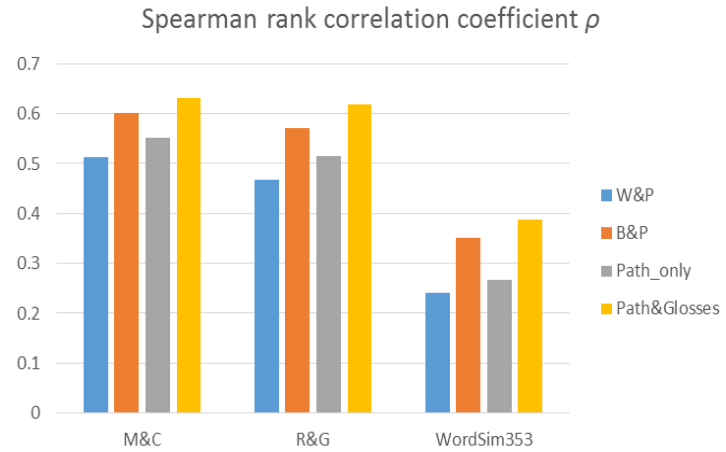

Fig. 2. Spearman rank correlation ρ.

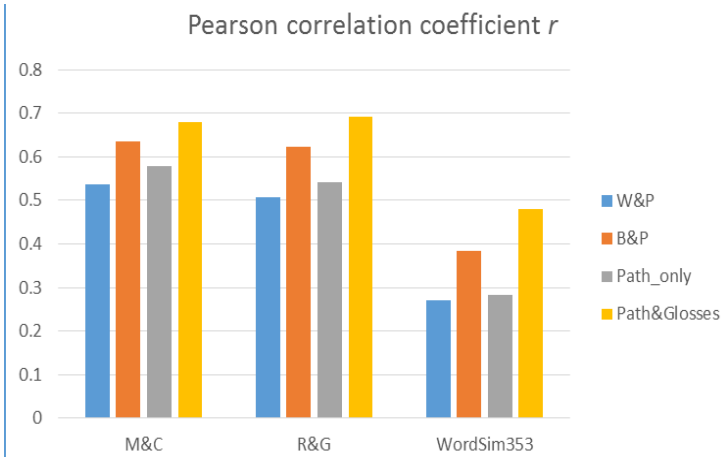

Fig. 3. Pearson correlation coefficient r.

The major reason for performance enhancement of Path&Glosses and glosses only measures compared to path only method is that they effectively capture the latent semantic association between words. Also, it is notable that all of the methods performed the worst in WordSim353 compared to the other two data sets, because there are too many feature vectors have zero values(e.g. words cannot find in WordNet).

5. Conclusions

WordNet provides very useful semantic information for word relatedness computation. However, some of them have not been fully investigated in practice. In this paper, we proposed two modified WordNet-based

word semantic relatedness measures that make use of location information of concept nodes in the ontology hierarchy and their glosses. We performed an experimental evaluation on three public data sets and compared the two modified measures against their corresponding classical methods. The combination of path and glosses clearly outperformed all of the comparison partners in terms of the Spearman rank correlation coefficient ρ as well as the Pearson correlation coefficient r. We argue that the reason for the consistent improvement of the combination method is that it better captures the explicit and implicit semantic relations between words. For the other modified path_only method which makes use of location information of concept nodes in WordNet hierarchy, also gain better results than the classical Wu and Palmer measure without considering concept node location information. Most notably, as the glosses only method is consistently performs better than any of the path only measures in the comparison partners, we believe that the implicit semantic information is more helpful than the explicit information in terms of word semantic relatedness assessment based on WordNet.

However, there still are some shortcomings in our research. We assign zero to the word which is not contained in WordNet. Ontology is always incompleteness, for the future work, we can extract knowledge source from Wikipedia to extend the ontology maybe a good research direction.

Acknowledgment

This research is supported by National High Technology Research and Development Program of China (863 Program) under Grant No. 2012AA101701 and the Science and Technology Plan Projects of Guangdong Province under Grant No. 2012A020100008.

References

[1] Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics, 32(1)*, 13-47.

[2] Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications, 42(4)*, 2264-2275.

[3] Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38(11),* 39-41.

[4] Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology, 6(1),* 1-12.

[5] Pan, D. (2013). A study of english word sense disambiguation base on wordnet. *Chinese Lexical Semantics* (pp. 166-174).

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

[7] Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*.

[8] Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database, 49(2),* 265-283.

[9] Jarmasz, M., & Szpakowicz, S. (2012). Roget's thesaurus and semantic similarity. *ArXiv Preprint ArXiv:1204.0245*.

[10] Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An Electronic Lexical Database, 305*, 305-332.

[11] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 5th Annual International Conference on Systems Documentation.*

[12] Cowie, J., Guthrie, J., & Guthrie, L. (1992). Lexical disambiguation using simulated annealing.

Proceedings of the 14th conference on Computational linguistics: *Vol. 1*.
[13] Kilgarriff, A., & Rosenzweig, J. (2000). English senseval: Report and results. *Proceedings of LREC.*
[14] Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *Proceedings of IJCAI.*
[15] Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. *Computational linguistics and Intelligent Text Processing,* 241-257.
[16] Mihalcea, R. (2005). Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.
[17] Navigli, R., & Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(7)*, 1075-1086.
[18] Sinha, R. S., & Mihalcea, R. (2007). Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. *Proceedings of ICSC.*
[19] Agirre, E., & Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.*
[20] Ponzetto, S. P., & Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.*
[21] Navigli, R., & Lapata, M. (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(4),* 678-692.
[22] Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8(10),* 627-633.
[23] Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6(1)*, 1-28.
[24] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., *et al.* (2001). Placing search in context: The concept revisited. *Proceedings of the 10th International Conference on World Wide Web.*

Tingting Wei is currently working toward the PhD degree in Information Science and Technology College in Sun Yat-Sen University, China. Her research interests include semantic web, data mining, and natural language processing.

Huiyou Chang was born in Heilongjiang province in China in 1962. He received B.Sc, M.Sc, and PhD degrees in computer science and technology from Harbin Institute of Technology in 1982, 1985, and 1988, respectively. Now he is a chief professor in the School of Software in Sun Yat-Sen University, China. His research interests include data mining, workflow management coalition, and e-business.