Mathematical and Computational Applications, Vol. 14, No. 1, pp. 55-63, 2009.

AN APPROACH FOR MEASURING SEMANTIC RELATEDNESS BETWEEN WORDS VIA RELATED TERMS

Mehmet Ali Salahli
Department of Computer Engineering
Canakkale On Sekiz Mart University, 17100
Canakkale, Turkey
msalahli@comu.edu.tr

Abstract-In this paper we propose a new approach for measuring semantic relatedness between words. The semantic relatedness between words are not measured directly, but are computed via set of words highly related to them, which we call the set of determiner words. Our approach for evaluating relatedness belongs to web page counting based measurement methods. We take into account some information, which contains hierarchical and other type of relations between the words. The experimental results demonstrate the effectiveness of proposed method.

Keywords-semantic relatedness, semantic similarity, information based measurement, information content

1. INTRODUCTION

Measures of relatedness or similarity are used in a variety of applications, such as information retrieval, automatic indexing, word sense disambiguation, automatic text correction. Semantic similarity and semantic relatedness are sometimes used interchangeable in the literature. These terms however, are not identical. Semantic relatedness indicates degree to which words are associated via any type (such as synonymy, meronymy, hyponymy, hypernymy, functional, associative and other types) of semantic relationships. Semantic similarity is a special case of relatedness and takes into consideration only hyponymy/hypernymy relations. The relatedness measures may use a combination of the relationships existing between words depending on the context or their importance. To illustrate difference between similarity and relatedness, Reznik [1] provides the widely used example of *car* and *gasoline*. These terms are not very similar; they have only few features in common. But they are more closely related in a functional context; namely that *cars* use *gasoline*. A number of researchers use distance measure as opposite of similarity.

In this work we propose a new approach for measuring semantic relatedness between words. Main idea of the approach is that the semantic relatedness between words is not measured directly, but is determined via a set of words high related to them, which we call the set of determiner words. Our approach for evaluating relatedness belongs to web pages counting based measurements methods. But we take into account some information, expressing hierarchical and other type relations between the words. Comparison the experimental results with a benchmark set of human similarity ratings show the effectiveness of the proposed approach.

The paper is organized as follows. Section 2 presents related work. In section 3 motivations on proposed method is given. The method for evaluating semantic

relatedness between the words is discussed in section 4. In this section the implementation results are presented also. Our conclusions and future work are presented in the final section.

2. RELATED WORK

A number semantic similarity method has been developed. Generally these methods can be classified into two main categories: edge counting methods and information content methods. Edge counting methods, also known as path based methods define the similarity of two words as a function of the length of the path linking the word and on the position of the terms in the taxonomy. The work of Rada et al. [2] deals with the basis of edge counting based methods. They compute semantic relatedness in terms of the number of edges between the words in the taxonomy. In Leacock and Chodorow [3] measure takes into account depth of the taxonomy in which the words are found: $lch(c_1,c_2) = -\log(length(c_1,c_2)/2D$, where $length(c_1,c_2)$ is the number of nodes along the shortest path between the two nodes. D is the maximum depths of the taxonomy. The Wu and Palmer similarity metric measures the depth of two given words in the taxonomy, along with the depths of the least common subsume (LCS): $sim_{wup} = (2*depth(LCS)/(depths(word_1)+depth(word_2))$. [4]

Information content methods, also known as corpus based methods measure the difference in information content of two words as a function of their probability of occurrence in a corpus. The method first proposed by Resnik[1]. According to Resnik similarity of two words is equal to information content (IC) of the least common subsumer: $sim_{rez} = IC(lsc(c_1,c_2))$. However, because many words may share the same LCSr, and would therefore have identical values of similarity, Resnik measure may not be able to obtain fine grained distinctions .[5] Jiang and Conrath [6] and Lin [7] have developed measures that scale the information content of the subsuming concept by the information content of the individual concepts. Lin does this via a ratio, and Jiang and Conrath with a difference.

Gloss based methods define the relatedness between two words as a function of gloss overlap. [8] Banerjee and Pedersen [9] have proposed the method that computes the overlap score by extending the glosses of the words under consideration to include the glosses of related works in a hierarchy.

Many of these measures were initially defined using the context of the WordNet ontology [10]. WordNet is a lexical reference system that was created by a team of linguists and psycholinguists at Princeton University. WordNet may be distinguished from traditional lexicons in that lexical information is organized according to word meanings, and not according to word forms. As a result of the shift of emphasis toward word meanings, the core unit in WordNet is something called a synset. Synsets are sets of words that have the same meaning, that is, synonyms. A synset represents one concept, to which different word forms refer. For example, the set {car, auto, automobile, machine, motorcar} is a synset in WordNet and forms one basic unit of the WordNet lexicon. Although there are subtle differences in the meanings of synonyms, these are ignored in WordNet.

Some researchers define the semantic relatedness between the words using Web. Danushka Bollegala an et al. [11] has proposed a method that exploits page counts and

text snippets returned by a Web search engine to measure semantic similarity between words. Rudi L and et al. [12] developed the method that defines the relatedness between the words via Google Similarity Distance. They use the World Wide Web as the database, and Google as the search engine. An approach to computing semantic relatedness using Wikipedia is proposed in [13]. Michael Strube and Simone Paolo Ponzetto also investigated the use of Wikipedia for computing semantic relatedness measures [14] Yhua Li and et all [15] has determined the semantic similarity by a number of information sources which consist of structural information from a taxonomy and information content from a corpus.

Some similarity measure based on applications of fuzzy sets theory. Particularly, the new fuzzy similarity measure with better performance compared with conventional similarity methods have been proposed in [16].

3. MOTIVATION

In this section we briefly focus on drawbacks of Web oriented and WordNet oriented approaches to motivate our method. First we look on Web oriented approach. Two linguistic factors negatively affect the results obtained from web based relatedness computing. These factors are synonymy, when many word are referring to same concept (for example, *car* and *automobile*), and polysemy, when many concepts are expressed by the same word (for example, Oracle). The impact of synonymy is that if a document consists of synonym word, then the other synonym of the word usually is not used in this document; authors prefer to use same word to expressing same meaning. For this reason, similarity degree between synonym words, computing via only web based methods, gets less value than as it is. For example, Google Search for "*journey*" returns 114000000 hits. (For calculating NGD distance the following site was used: http://digitalhistory.uwo.ca/cgi-bin/ngd-calculator.cgi). The number of hits for "voyage" is 113000000. The numbers of pages where "*journey*" and "*voyage*" are occurred are 1670000. Using these data we obtain a normalized Google Distance between the highly semantic similar words "*journey*" and "*voyage*" as

$$NGD(journey, voyage) \approx 0.90808$$

If we believe that this result is reliable, we must say that there is not any similarity between the "*journey"* and "*voyage*".

Polysemy gives opposite effect, causing documents that use the same word in different senses to be considered related when they should not be. For example, the word "*cord*" may be used in various means (rope*, automobile, rock group, spinal cord*…). A Google Search for "*cord"* returns 61400000 articles. But if we are interested only "*spinal cord"* meaning of the word, approximately 148000 articles will meet our interest.

Namely, for these reasons measuring semantic similarity based on large search engine don't give expected results. Certainly, without any alternatives web contains sufficient information about words and their relations. But the main problem is to find the ways that allow us to extract only useful, related information from the huge information storage.

Now we will give a sample that clearly indicates the drawbacks of Wordnet based methods. The similarity values between the "*student*"and"*examination*", have computed by methods based on Wordnet ontology are given in the Table 1 (For calculating similarity please refer to: http://marimba.d.umn.edu/cgi-bin/similarity.cgi). As it is seen from Table 1, the similarities on *hco, lin* and *res* methods are equal to null. Other methods return little similarity value between the words.

Table 1. Similarity between the word *student* and *examination*

similarity method	hco	jcn	wup	path	lin	lesk	res	lch	V_p	v
Similarity value	0	0.0666	0.25	0.0769	0	20	0	1.0726	0.3163	0.3568

Another example of similarity data between the words "*student"* and "*animal"* are given *in* the table 2.

Table 2. Similarity between the word *student* and *animal*

similarity method	hco	jcn	wup	path	lin	lesk	res	lch	V_p	v
Similarity value	3	0.1207	0.75	0.2	0.3 615	28	2.3 447	2.0281	0.0041	0.2338

Comparing the values from the tables above we can conclude that "*student*" and "*animal*" have more similarity than "*student*" and "*examination*". The samples clearly indicate the difference between the "relatedness" and "similarity". To strengthen the idea, both approaches are not sufficient for measuring relatedness between the words separately. Before measuring relatedness we must clearly determine what we expect from relatedness and measure methods we should choose according to our expectation.

In the next section we propose the relatedness measure which may be useful for applications on information retrieval.

To solve the problems we encountered a method, determining the similarity of words via related those terms (like keywords for articles) which we call determiner words. For every word it is not difficult to find closely related terms. For example, if we say "*student*", the words "*examination*", "*university*", "*instructor",* and *"young people*" comes into the mind. We think that using a set of related words allows us to define a word more preciously.

4. THE METHOD

Let W_1 and W_2 be words, which we want to measure relatedness between them.

The method determines the following steps:

1. Determine the pairs of sets of the related words on W_1 and W_2.

Let $D_1=\{d_{11}, d_{12}, d_{13},\ldots, d_{1n})$ and $D_2=\{d_{21}, d_{22}, d_{23},\ldots, d_{2m})$, these are the sets of determiner words of W_1 and W_2 respectively. Next we form the set of common determiner words D as:

$$D = D_1 \cup D_2$$

We call the elements of D as d to avoid of complexity.

$$D==\{d_1, d_2, d_3,\ldots, d_k),$$

where k is equal to or less than (n+k).

2. Calculate the normalizing values of relatedness between the determiners and W_1 (W_2):

$$rel(d_i, W_1) = \text{freq (d, W1)/maxfreq}_1$$

$$rel(d_i, W_2) = \text{freq (d, W1)/maxfreq}_2$$

Where freq (d_i, W1) - is a number of pages where d_i and W_1 are occurred together. Analogically, freq(d_i,W_2)- is defined.

$$\text{maxfreq}_1 = \max\{rel(d_1, W_1), rel\ (d_2, W_1), \ldots, rel\ (d_k, W_1)\}$$
$$\text{maxfreq}_2 = \max\{rel(d_1, W_2), rel\ (d_2, W_2), \ldots, rel\ (d_k, W_2)\}$$

We consider that if a determiner word is highly related to the word, then the probability of the determiner occurring in the pages where the word's appearance is high. In a special case, if d_i is synonymy, or nearly synonymy to W_1 (W_2) we take $rel(d_i, W_1)=1$ ($rel(d_i, W_2)=1$).

3. Calculate the relatedness between the words:

$$rel\ (W_1, W_2) = \left(\sum_{i=1}^{k}\left(\frac{\alpha_i R_i}{1+R_i}\right) + syn\right)/(1+syn)$$

Here

$$R_i = \frac{\min\{rel(d_i, W_1), rel(d_i, W_2)\}}{\max\{rel(d_i, W_1), rel(d_i, W_2)\}}$$

α_i is called the co-occurrence factor, and is defined as

$$\alpha_i = \begin{cases} 2, & d_i \text{ is occurred in both words } W_1 \text{ and } W_2 \\ 1, & \text{otherwise} \end{cases}$$

syn is called synonymy factor and is defined as

$$\text{syn} = \begin{cases} 1, & W_1 \text{ and } W_2 \text{ are synonymy or nearly synonymy} \\ 0, & \text{otherwise} \end{cases}$$

The sample.

To explore our method we use the pair of words of *(car, train*) from Rubenstein-Goodenough set [17]. We take W1= car; W_2= train. Determiner words of W1 and W_2 are

D_1 = {*rail, transport, vehicle, freight, passenger*)
D_2 = {*automobile, motor, wheel, passenger, vehicle*)

Thus *automobile* is a synonymy (or nearly synonymy) of car, we count that all the pages in which *car* occurred, automobile is occurred also. In other words, hits of (*car, automobile*) are equal to 1. The words *vehicle* and *passenger* are determiners of the both words. For these determiners co-occurrence factor is equal to 2. Data about the number of hits are given in Table 3.

Table 3. Determiners of the word *train* and car and theirs hits

determiners	Rail	Transport	Vehicle*	Freight	Passenger*	Automobile	motor	wheel
train	3.85	29.8	2.94	2.2	3.03	2.41	2.49	2.56
car	18.1	129	90.9	2.47	12.1	129**	90	42.3

* indicates that related words are determiners of both words.

** indicates that related value is not real numbers of pages, where automobile and car are co-occurred. Thus these words are synonymy, as the hits numbers we take the maximum of hits for car on the determiner words.

According to the formulae we obtained that relatedness between train and *car* is 0.54182719 apposite of 6.31 on FC (note that FC measuring gives a number between 0 and 10).

5. IMPLEMENTATION

For realization our method we used WordNet and Wikipedia as information sources. As we mentioned above WordNet is a lexical database, developed at Princeton by Miller and freely available. On 2006 the WordNet database contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs. However WordNet does not include some named entities and specialized concepts. Wikipedia is a multilingual, web-based, free content encyclopedia project, operated by the Wikimedia Foundation, a non-profit organization. On July 20, 2007, Wikipedia has approximately 7.8 million articles in 253 languages, 1.893 million of which are in the English edition.

For evaluating proposed method we used Miller-Charles dataset [10]. Miller-Charles dataset consists of 30 word-pairs rated by a group of 38 human subjects. The word pairs are rated on a scale from 0 (no similarity) to 4 (perfect synonymy). The dataset is considered as a reliable benchmark for evaluating semantic similarity

measurements. Most researchers have used only 28 word pairs of the Miller-Charles set. These pairs have been used in our experiments also. In table 4 the result of experiments implementing on Miller Charles dataset is presented.

Table 4. Semantic Similarity of Human Ratings and Baselines on Miller-Charles dataset

Word Pair	Miller-Charle's	Jaccard	proposed
Cord-smile	0.13	0.102	0.137
Rooster-vojage	0.08	0.011	0.208
Noon-string	0.08	0.117	0.052
Glass-magician	0.11	0.181	0.107
Monk-slave	0.55	0.862	0.463
Coast-forest	0.42	0.016	0.649
Monk-oracle	1.1	0.072	0.223
Lad-wizard	0.42	0.068	0.293
Forest-graveyard	0.84	0.012	0.345
Food-rooster	0.89	0.963	0.738
Coast-hill	0.87	0.444	0.559
Car journey	1.16	0.071	0.443
Crane-implement	1.68	0.189	0.635
Brother-lad	1.66	0.189	0.713
Bird-crane	2.97	0.235	0.877
Bird-cock	3.05	0.153	0.857
Food-fruit	3.08	0.753	0.685
Brother-monk	3.82	0.261	0.752
Asylum-madhouse	3.61	0.024	0.863
Furnace-stove	3.11	0.401	0.887
Magician-wizard	3.5	0.295	0.653
Journey-voyage	3.84	0.415	0.879
Coast-shore	3.7	0.786	0.902
Implement-tool	2.95	1	0.762
Boy-lad	3.76	0.186	0.916
Automobile-car	3.92	0.654	0.939
Midday-noon	3.42	0.106	0.876
Gem-jewel	3.84	0.295	0.836
correlation	**1**	**0.692**	**0.953**

The correlation derived on the proposed method (0.953) shows high effectiveness of the proposed method.

6. CONCLUSION

A new approach for measuring the relatedness between the words has been presented in this paper. The approach is based on using determiner words. The experimental results show the effectiveness of the method. But there are some problems with application of the method. Main problem is to choose the determiner words. For this purpose, articles from Wikipedia may be used. Using common words as determiner is not recommended. Although is not limit to numbers of determiners, we think that 5-10 determiners for per words are sufficient. The main baseline of our future studies is design of the algorithm, allowing selecting of determiners from information sources automatically.

REFERENCES

1. Philip Resnik, Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research,* **11**, 95–130, 1999.
2. R.Rada, H.Mili,E.Bichnell, and M.Blettner, Development and Application of a Metric on Semantic Nets, *IEEE Trans. Systems, Man, and Cybernetics,* **9,** 1-30,1989.
3. C. Leacock and M. Chodorow, Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet, *An Electronic Lexical Database,* 265—283, MIT Press, 1998.
4. Wu Z, and Palmer, M. Verb semantics and lexical selection, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 133-138, Las Cruces, New Mexico, 1994
5. Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan and Christopher G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, *Journal of Biomedical Informatics*, **40**, 288-299, 2007
6. Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. *In: Proceedings of the 10th international conference on research in computational linguistics,* 19–33,Taipei, Taiwan, 1997
7. D.Lin, An information-theoretic definition of similarity, *Proceedings of the 15th International Conference on Machine Learning,* , 296–304, Madison , Wisconsin USA, 1998
8. Michael Lesk, 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*, 24–26, Toronto, 1986
9. Banerjee S, Pedersen T, An adapted Lesk algorithm for word sense disambiguation using WordNet, *Proceedings of the third international conference on intelligent text processing and computational linguistics*. Mexico City, Mexico, 136–45, 2002
10. G.A.Miller, WordNet: A lexical Database for English, *Comm. ACM*, **38**, 39-41,1995
11. Danushka Bollegara, Yutaka Matsuo, and Mitsuru Isizuka, Measuring Semantic Similarity between Words Using Web Search Engines, *Proceedings of the 16th International World Wide Web Conference (WWW2007),* 757-766, Banff, Alberta, Canada, 2007

12. Rudi L. Cilibrasi and Paul M.B. Vitanyi, The Google Similarity Distance, *IEEE Transactions on Knowledge and Data Engineering,* **19**, 370-383, 2007
13. Evgeniy Gabrilovich and Shaul Markovitch, Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 1606-1611, Hyderabad, India, 2007
14. Michael Strube and Simone Paolo Ponzetto, WikiRelate! Computing Semantic Related-ness Using Wikipedia, *Proceedings of the 21st National Conference on Artificial Intelligence,* 1419-1424, Boston, Mass, 2006
15.Yuhua Li, Zuhair A.Bandar, and David McLean, An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources, *IEEE Transactions on Knowlege and Data Engineering*, **15**, 871-882, 2003
16. Rıdvan Saraçoğlu, Kemal Tütüncü and Novruz Allahverdi, A fuzzy clustering approach for finding similar documents using a novel similarity measure, *Expert Systems With Applications*, **33**, 600-605, 2007.
17. H.Rubenstein and J.B.Gooodenough, Contextual Correlates of synonymy, *Communications of the ACM,* **8,** 627-633, 1965