Measuring Semantic Similarity between Words Using HowNet

Liuling DAI
Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, PRC
dailiu@bit.edu.cn

Bin LIU
School of Computer Science, Beijing Institute of Technology, Beijing 100081, PRC

Yuning XIA
Center for Speech and Language Technologies,
RIIT, Tsinghua University
Beijing 100084, PRC
yqxia@tsinghua.edu.cn

ShiKun WU
Central Radio and TV Tower, Beijing, 100036, PRC

Abstract

Semantic similarity between words is a fundamental issue for many natural language processing applications. The difficulty lies in that how to develop a computational method that is capable of generating satisfactory results close to how humans perceive. In this paper, a novel method is proposed to measure semantic similarity between words using HowNet, which is a renowned Chinese-English bilingual knowledge base. Furthermore, a Chinese thesaurus is used to improve the similarity measuring. Theoretically, our method can be used in many languages while in this case it is applied for English and Chinese. Experiments on English and Chinese word pairs show that our method are closest to human similarity judgments when compared to the major state-of-the-art methods.

1. Introduction

The study on semantic similarity measuring between words acts as an important role in natural language processing (*NLP*) and information retrieval (*IR*). It has been applied in many applications such as word sense disambiguation [1], and full text searching [2], etc.

For English, *WordNet* [1] is the most popular and valuable resource. In fact, there is another important knowledge base called *HowNet* [2] [3]. It is a Chinese-English bilingual knowledge base aiming to provide knowledge for *NLP* applications, especially in cross-lingual manner. It has attracted many researchers since it was released in 2000. But so far, the majority of the research based on *HowNet* is limited in Chinese language processing [3, 4].

Many semantic similarity measuring approaches have been developed in the past decades. These approaches can be roughly classified into two groups, i.e. distance-based approaches and corpus auxiliary approaches.

The distance-based approaches measure the semantic similarity between two words using the distance defined in lexicon or knowledge base. Some instances of this kind of approach are published by Rada [5], Leacock and Chodorow [6], Yang and Powers [7], Hirst and St-Onge [8] and Alvarez [9].

Generally speaking, a distance-based approach maps the semantic similarity between two words by a formula defined as follows.

$$Sim(w_1, w_2) = \varphi(dist(w_1, w_2)) \qquad (1)$$

where $dist(\cdot)$ returns distance between $w1$ and $w2$, and $\varphi(\cdot)$ is a function that transforms distance to similarity defined with various considerations.

The corpus auxiliary approaches measure word similarity by considering not only lexical information but also auxiliary information such as word co-occurrence frequency and other statistical information

[1] WordNet: http://wordnet.princeton.edu

[2] *HowNet*: http:// www.keenage.com

978-0-7695-3308-7/08 $25.00 © 2008 IEEE
DOI 10.1109/ICCSIT.2008.101

extracted from auxiliary corpus. Examples of this method include Resnik [10], Jiang [11], Lin [12] and Li et al [13].

According to the published results, corpus auxiliary approaches outperform distance-based approaches in some degree while are more complex. All of these aforementioned studies are based on *WordNet*.

In this paper, a novel method is proposed to measure the similarity between English words or Chinese words using *HowNet*. Our method not only refines the distance-based methodology, but also explores *HowNet* as knowledge base. In practice, we denote a concept of a word via a concept graph according to its definition in *HowNet*. Then we compute the similarity between concepts by measuring the similarity between concept graphs. As some words are missing in *HowNet*, serious deviation occurs on these words in *HowNet*-based similarity measurement. To handle this problem, a thesaurus is incorporated. We will compare our method against the above mentioned studies.

The rest of this paper is organized as follows. We firstly review related works in Section 2. Then we introduce *HowNet* and its advantages against *WordNet* briefly in Section 3. The details of our method are presented in Section 4. We evaluate the method in Section 5 and conclude this paper in Section 6.

3. Introduction to HowNet

HowNet is an on-line common-sense knowledge base unveiling the inter-conceptual relations and inter-attribute relations between concepts. Each concept is denoted by Chinese words and its English equivalents. *HowNet* covers plentiful semantic knowledge and world knowledge thus becomes an important resource for *NLP* and knowledge mining [3].

Two important notations in *HowNet* are notable. A ***concept*** describes a semantic sense of words. In nature language, one word may carry several concepts. This phenomenon is called polysemy. The second notation is ***sememe***. In *HowNet*, concepts are described by *Knowledge Description Language* (*KDL*). The basic element of *KDL* is *sememe* which is the basic unit to describe concepts. Different from *WordNet*, *HowNet* doesn't put all of the concepts into a tree directly, but describes them by a set of sememes. The exception is that in *HowNet,* hypernym-hyponym relation organizes sememes into several trees.

Differences and arguable advantage of *HowNet* over *WordNet* are described by Dong [3] as follows. (1) *WordNet* is human oriented while *HowNet* is computer oriented; (2) *WordNet* is word based but *HowNet* is concept based; (3) The atomic unit of *HowNet* is synset but that of *WordNet* is sememe; (4) *WordNet* is defined with natural language but *HowNet* is represented by structural markup language. These characteristics make *HowNet* friendly for computer systems including word similarity measuring.

As far as the problem of similarity measuring is concerned, another advantage is worth of noting when *HowNet* is used. As the concepts described in *HowNet* are represented in both English and Chinese, the methods based on *HowNet* are able to measure semantic similarity between words in different languages. This contributes a lot to cross-lingual applications such as machine translation and cross-lingual information retrieval.

4. Algorithm

As *HowNet* doesn't organize words directly into a tree, we are not able to measure similarity between words directly. As a precondition, we need to measure the semantic similarity between sememes.

4.1. Similarity between sememes

HowNet organizes all the sememes into several trees and each sememe is considered as a node of a tree. So we are able to count distance between any two sememes. We define the distance between sememes as the number of edges of the shortest path between them. If the two sememes are not on the same tree, the distance between them is set to infinite.

We adopt the distance-based method, which is frequently used on *senses* of *WordNet*, to measure semantic similarity between sememes. To do this, *Eq*.2 is used to calculate similarity between sememes as:

$$Sim(S_1, S_2) = \varphi(dist(S_1, S_2)) \qquad (2)$$

where S_1 and S_2 are two sememes, $Sim(S_1, S_2)$ is the semantic similarity between S_1 and S_2. $dist(S_1, S_2)$ is the distance between S_1 and S_2. We will try two strategies to implement function $\varphi(\cdot)$.

Strategy 1. We modify the strategy proposed by Liu [4] which assumes that the similarity between two sememes only depends on the distance between them. Intuitively, we know that the similarity varies according to the depth of sememes in the sememe tree. For example, the distance between *beast* and *animal* is the same as that between *thing* and *entity*. But the latter sememe pair is more abstract. So the similarity between *beast* and *animal* would be larger than that between *thing* and *entity*. To imply this impact, we modify *Eq*. 2 as:

$$Sim(S_1, S_2) = \frac{\alpha}{d+\alpha} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \qquad (3)$$

where d is the distance between S_1 and S_2, α is a parameter which means the distance when the similarity is 0.5. h is the dept of the first common parent node of the two sememes, β is a smoothing factor. In the right part of *Eq*. 3, the second part is the impact of depth.

Strategy 2. The second strategy is borrowed from literature [13] in which Li et al. thoroughly studied the word similarity using *WordNet* and other resources. We transplant their formula which draws the best results based on *HowNet*. As is:

$$Sim(S_1,S_2)=e^{-ad}\cdot\frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}} \tag{4}$$

where $\alpha \geq 0$ and $\beta \geq 0$ are parameters scaling the contribution of the distance and depth, respectively.

4.2. Similarity between concepts

In *HowNet*, a concept is defined by a *DEF* expression with nesting grammar. A *DEF* expression consists of sememes and a framework. The framework organizes the sememes into a complete definition. For example, the *DEF* of *doctor* is:

DEF={human|人:{own|有: possession={Status|身分:domain={education|教育}, modifier={HighRank|高等:degree={most|最}}},possessor={~}}}.

The primary sememe is *human*, which means a doctor is an instance of human. The following sememe *own* modifies the primary sememe. Furthermore, the sememe *own* is modified by sememe *Status*, *Status* is modified by *education* and *HighRank*, *HighRank* is modified by *most*. The framework consists of *possession*, *domain*, *modifier*, *degree*, and *possessor*.

We can denote a *DEF* with a concept graph. To make things simply, we take the primary sememe as the most important sememe. And the importance of a modifying sememe is independent with its position. Thus the concept graph can be reduced into only one layer. Again, use the definition of doctor as an example. *Figure* 1 illustrates the concept graph of *doctor*.

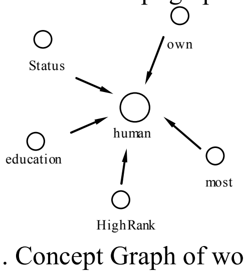

Figure 1. Concept Graph of word *doctor*

In *Figure* 1, *human* is the primary sememe, and the others are modifying sememes of *human*. Let P and Q be two concepts and the number of modifying sememes of P is less than that of Q. We use the formula below to measure the semantic similarity between them.

$$Sim(P, Q) = \alpha\cdot Sim\,(P',Q') + \beta\cdot\frac{\sum\limits_{0\leq i<|P|}\max\limits_{0\leq j<|Q|}(Sim(P_i,Q_i))}{|P|}+\gamma\frac{num(S,T)}{|S|+|T|} \tag{5}$$

In *Eq*. 5, *P'* and *Q'* is the primary sememe of P and Q, respectively. $|P|$ and $|Q|$ are the number of modifying sememes in their concept graphs. S and T are the descriptors sets of framework of P and Q. $num(S, T)$ is the number of the common descriptors of them. $|S|$ and $|T|$ are the number of their descriptors.*α, β and γ* is the parameters that scale the weight of the three parts.

4.3. Similarity between words

Similarity between words actually means the similarity between concepts associated with them. But in a natural language, a word can represent one or more concepts. Under this circumstance, we take the maximum similarity between the concepts as the similarity between the two words. Formally, it is defined as:

$$Sim(W_1, W_2) = max\ Sim(C_{1i}, C_{2j}) \tag{6}$$

where C_{1i}, C_{2j} is the i-th and j-th concept associated with W_1 and W_2 respectively. $Sim(C_{1i}, C_{2j})$ is the similarity between C_{1i} and C_{2j} that can be computed by *Eq*. 6.

4.4. Amendment with thesaurus

Our observation shows that some words are missing and *some DEFs* are too rough in in *HowNet.* This may cause serious deviations from human judgments when we are using *Eq*. 6 to measure similarity. To handle this problem, we use a Chinese thesaurus[3] named *Tongyici Cilin* to amend the deviations.

Cilin organizes the words in a tree liking how *WorldNet* does. So we can utilize the tree structure of *Cilin* to measure the similarity among words according to *Eq*. 1. In this paper, we use the following formula:

$$Sim(W_1,W_2)=\frac{\alpha}{d+\alpha} \tag{7}$$

where d is the distance between W_1 and W_2, α is a parameter which means the distance when the similarity is 0.5.

[3] HIT IR-Lab Cilin (Extended)

At last, we use the formula below to mix the similarities of using *Eq*. 6 and *Eq*. 7 to get the final results.

$$Sim(P,\ Q)\ = \begin{cases} \alpha\cdot Sim_1\ (P,Q)\ +\beta\cdot Sim_2(P,Q) \\ \qquad\qquad if\,(Sim_1 > Sim_2) \\ \gamma\cdot Sim_1\ (P,Q)\ +\eta\cdot Sim_2(P,Q) \\ \qquad\qquad if\,(Sim_1 \le Sim_2) \end{cases} \qquad (8)$$

where $Sim_1(\cdot)$ and $Sim_2(\cdot)$ are computed with *Eq*. 6 and *Eq*. 7 respectively. α, β, γ and η are the parameters to scale the weights of the two parts.

5. Evaluation

5.1. Dataset and measurement

Rubenstein and Goodenough [14] established *synonymy judgments* for 65 pairs of nouns. They invited 51 human judges to assign every pair a score between 0.0 and 4.0 to indicate semantic similarity. Miller and Charles [15] follow this idea and restricted themselves to 30 pairs of nouns selected from Rubenstein and Goodenough's list, divided equally amongst words with high, intermediate and low similarity.

To show our method being capable of measuring similarity between Chinese words, we translated *GR-65* dataset into Chinese manually and evaluate our algorithm on them.

We use the human measurements of RG's experiment and MC's experiment as the baselines. As many published work did, we compute the correlation coefficient [16] between the human judgments and the measures achieved by our method. In addition, we would like to compare our results to eight groups of measures that rely on *WordNet*. The measures are reported by: Hirst and St-Onge [8], Jiang and Conrath [11], Leacock and Chodorow [6], Lin [12], Resnik [10], Yang and Powers [7], Li [13], and Alvarez [9].

5.2. Experimental results

We measure the similarities between words using *Eq*. 8. In *Eq*. 3, α is set to 1.6, β is set to 0.16. In *Eq*. 4, α and β are set to 0.2 and 0.16 respectively. In *Eq*. 5, α , β and γ *are* set to 0.54, 0.36 and 0.1. In *Eq*. 8, α, β, γ and η *are* set to 0.95, 0.05, 0.05 and 0.95 on Chinese dataset, 0.95, 0.05, 0.45 and 0.55 on English dataset.

The semantic similarities of English word pairs are shown in *Figure* 2 and those of Chinese word pairs are shown in *Figure* 3. In them, we list four measurements for each word pair. The line labelled by *S1* and *S2* are the results achieved by our algorithm, using *Eq*. 3 and *Eq*. 4 to measure similarity between sememes respectively. The line labelled with *HAPI* is the result by directly calling the "*HowNet_Get_Concept_Similarity*" API of *HowNet* system [3].

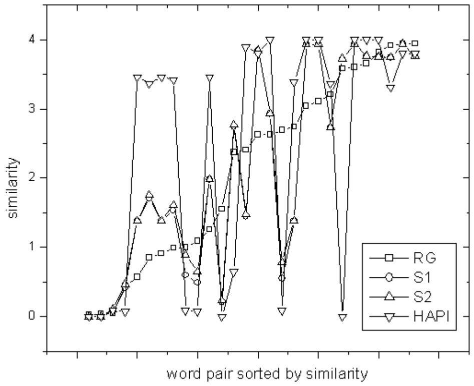

Figure 2. Similarities of RG's 28 pairs of words.

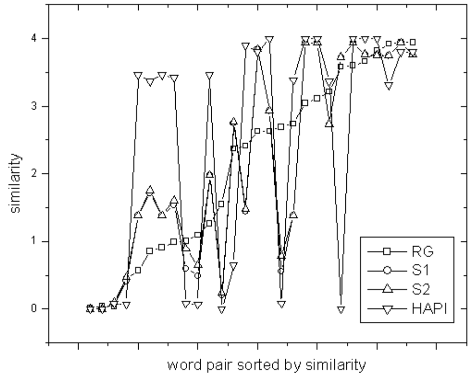

Figure 3. Similarities of 28 Chinese RG's pairs of words.

As shown in *Figure* 2 and *Figure* 3, the majority of results drawn by *S1* and *S2* are sound. And the results achieved by *S1* and S2 are very similar, which are much better than those achieved by *HAPI*. There are two words i.e., *madhouse* and *oracle*, are not enrolled in current version of *HowNet*. To cope with this problem, we replace them with their synonyms *mental hospital* and *prophet*, which are enrolled in current version of *Hownet*.

To evaluate our algorithm furthermore, we conducted experiments on Rubenstein and Goodenough's 65 word pairs and their translated word pairs in Chinese. In order to compare with published results, we listed the correlation coefficients of algorithms to the baselines. The correlation coefficients are shown in *Table* 1.

Table 1. Correlations coefficient of algorithms

Aproach	RG-28	MC-28	RG-65
Hirst-St.Onge	0.671	0.682	0.732
Jiang	0.670	0.682	0.732
Leacock	0.801	0.820	0.852

Lin	0.773	0.814	0.834
Resnik	0.706	0.763	0.800
Yang	0.889	0.921	0.897
Li	0.8914	0.882	*N/A*
Alvarez	0.900	0.913	*N/A*
S1-English	0.9238	0.9074	0.8764
S2-English	0.9286	0.9056	0.8744
HAPI-English	0.5371	0.5113	0.6089
S1-Chinese	0.8617	0.8401	0.8958
S2-Chinese	0.8679	0.8460	0.8950
HAPI-Chinese	0.5328	0.5001	0.6752

As can be seen from *Table* 1, the correlation coefficients between our algorithms (labelled by *S1* and *S2*) and baselines are inspiring. On English dataset, the correlation coefficients between our algorithms and baselines are about 0.92 on *RG-28*, 0.90 on *MC-28* and 0.87 on *RG-65*. These results are of the top grade when compared with those published earlier. For example, the highest correlation coefficient on *RG-28* is 0.9 achieved by Alvarez [9], while ours is 0.9238 with *S1* and 0.9286 with *S2*. On Chinese dataset, the correlation coefficients between our algorithms and baselines are about 0.86 on *RG-28*, 0.84 on *MC-28* and 0.90 on *RG-65*, which mean they are very consistent.

6. Conclusion and Future Works

In this paper, an algorithm is proposed to measure semantic similarity between two words. Different from previous studies, we use *HowNet* instead of *WordNet* as the underlying knowledge base. We denote concept of a word via a concept graph according to its definition in *HowNet*. And we compute similarity between concepts by measuring the similarity between their concept graphs. To handle the deviations caused by absence or roughness of words in *HowNet*, a thesaurus is adopted during similarity measuring. In a manner similar to earlier researches, we carried out experiments on a benchmark set of word pairs and took human similarity ratings as baselines to evaluate our method. Experimental results show that our algorithm is competitive with works based on *WordNet*. We are planning to find a boosting method that combines *HowNet* with other resources, such as *WordNet* or corpora, to get more accurate results in future researches.

Acknowledgement

This work is supported by Beijing Key Discipline Program, NSFC (60703051), and Basic Research Foundation of Beijing Institute of Technology (20071142004).

References

[1] P. Resnik, "Semantic Similarity in a Taxonomy: an Information-based Measure and Its Application to Problems of Ambiguity in Natural Language", *Artificial Intelligence Research*, Vol 11, 1999, pp. 95-130,

[2] R. K. Srihari, Z. F. Zhang, A. B. Rao, "Intelligent Indexing and Semantic Retrieval of Multimodal Documents", *Information Retrieval*, Vol 2, 2000, pp. 245-275.

[3] Z. Dong and Q. Dong. *HowNet and the Computation of Meaning*. World Scientific Publishing. 2006.

[4] Q. Liu, S. J. Li, “Word Semantic Similarity Computation based on HowNet”, *The 3rd Chinese lexical and semantic proseminar, Taipei*, 2005. (in Chinese)

[5] R. Rada, H. Mili, E. Bichnell, M. Blettner, "Development and Application of a Metric on Aemantic Nets", *IEEE Trans. Systems, Man, and Cybernetics*, Vol 9, No. 1, Jan, 1989, pp. 17-30.

[6] C. Leacock, M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification", *Proceeding of Fellbaum*, 1998, pp. 265-283.

[7] D. Yang, D.M.W. Powers, “Measuring semantic similarity in the taxonomy of WordNet”, *Proc. of ACCS-05*, pp. 315–322.

[8] G. Hirst, D. St-Onge, “Lexical chains as representation of context for the detection and correction of malapropisms”, In Fellbaum, *WordNet: An Electronic Lexical Database and Some of its Applications*, The MIT Press, Cambridge, MA, 1998, pp. 305-332.

[9] M. A. Alvarez, S. Lim, "A Graph Modeling of Semantic Similarity between Words", *Proc. of ICSC-07*, 2007.

[10] P. Resnik, "Using Information Content to Evaluate Semantic Similarity", *Proceeding of the 14th International Joint Conference on Artificial Intelligence*, Montreal, 1995, pp. 448-453.

[11] J.J. Jiang, D.W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", *Proc. ROCLINGX*, 1997.

[12] D. Lin, "An Information-theoretic Definition of Similarity", *Proc. of ICML-98 Madison*, Wisconsin, 1998.

[13] Y. Li, Z. Bandar, D. McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, July-August, 2003, pp. 871-882.

[14] H. Rubenstein, J.B. Goodenough, “Contextual correlates of synonymy”. *Communications of the ACM*, 1965, 8(10), pp. 627-633.

[15] G.A. Miller, W.G. Charles, “Contextual correlates of semantic similarity”, *Language and Cognitive Processes*, 1991, 6(1), pp. 1-28.

[16] A.L. Edwards, *An Introduction to Linear Regression and Correlation*. San Francisco: W.H. Freeman, 1976.