Chinese-English Bilingual Word Semantic Similarity Based on Chinese WordNet

Yangyang Wu[1*], Siying Wu[1], Duansheng Chen[1]
[1] College of Computer Science & Technology, Huaqiao University, Xiamen, Fujian, China.

* Corresponding author. Tel.: +865926162528; email: wuyy@hqu.edu.cn
Manuscript submitted January 23, 2014; accepted July 25, 2014.

Abstract: Semantic similarity measurement of multilingual words is a challenging problem in data mining, information extraction, information retrieval, etc. This paper introduces an algorithm to measure the semantic similarity of Chinese-English bilingual words based on Chinese WordNet, an expansion of WordNet in Simplified Chinese. The algorithm not only measures the semantic similarity for Chinese and English words, but also measures Chinese-English cross-lingual word semantic similarity. It utilizes WordNet's hypernym / hyponym relationships between synsets and evaluates the similarity by measuring the distances between synsets, the local densities of synsets and the depths of the synsets on the entire hierarchy of WordNet. Most words have more than one meaning. Therefore, the algorithm sets up the weights of the combination pairs of the two words' synsets in an adaptive mode. Experimental results show that the similarities measured by our algorithm match with human common sense in general.

Key words: Word semantic similarity, Chinese WordNet.

1. Introduction

Multilingual words semantic similarity measurement is an important topic / problem in the fields of data mining, information retrieval, machine translation, etc. Data mining methods such as text clustering usually extracts words or phrases from text documents as features, represents documents as feature vectors, and then groups the documents into classes or clusters based on the similarity of the feature vectors. In a multilingual document collection, the extracted features will contain multilingual words. It is important to measure the similarity between words not only from the same language, but also from different languages. In information retrieval, a typical problem is to locate relevant documents in a document collection based on some keywords describing an information need or an example relevant document. Estimating semantic similarity precisely between words is critical for judging if a document is relevant to a user's need. Many information retrieval systems, such as on-line library catalog systems, web search engines, they all need to deal with multilingual documents and measure semantic similarity between words from different languages.

Despite multilingual word semantic similarity measurement required in many fields, most algorithms measure the semantic similarity between words from the same language. Eneko Agirre *et al*. pioneered cross-lingual similarity in 2009. In the literature [1], they explored WordNet-based and distributional similarity approaches, and combined these two methods. Finally, they applied their methods to cross-lingual similarity and showed the experimental results for two English-Spanish cross-lingual datasets.

This paper introduces an algorithm for Chinese-English bilingual word semantic similarity calculation, which was developed by us and has been used in a project on data spaces for three years. The concept of data spaces was proposed by Micheal Franklin, Alon Halevy and David Maier in 2005 [2]. A data space should contain all the information that is relevant to a particular organization regardless its format and location, and model a rich collection of relationships between data repositories. In order to model the relationships between data repositories in data spaces, data space support platforms need to measure semantic similarity and relatedness between data items. Our data space contains a great deal of Chinese-English bilingual data repositories. Constructing a semantic mapping requires a component, which is able to measure the semantic similarity between a given pair of Chinese words, English words or Chinese-English words.

The techniques used to measure semantic similarity between terms can be classified into two main categories: one relies on pre-existing knowledge resources such as the well known WordNet, the other inducing distributional properties of words from corpora [1], [3]-[5]. In our project on data spaces, it is important to monitor the contents of the data space to propose additional relationships over time. The data sources in a data space can be relational databases, XML repositories, text databases, web services, etc., they contain multiple subjects that users are interested in. Distributional similarities may be influenced by the corpus, some similar phrases only made sense in the context of the corpus used [6]. In many applications, corpus is not available. Based on the actual demand for evaluation of the semantic similarity between data items in data spaces, we chose the technique that only relies on pre-existing knowledge resource and present a method based on Chinese WordNet, which is an expansion of the expert crafted WordNet in Simplified Chinese. Most words have more than one sense; therefore, our method determines word similarity by combining some top conceptual similarities from all the concept pairs of two given words. It sets the weights of the concept pairs in an adaptive mode to make the weights vary with the number of the concept pairs of the given words.

The rest of this paper is organized as follows. Section II introduces related work. Section III presents our algorithm of Chinese-English bilingual word semantic similarity measure. Section IV gives some results of our experiments. The final section draws a conclusion and mentions the further work.

2. Related Work

The techniques that rely on the pre-existing knowledge resources are more useful than those which rely on corpora for semantic similarity [1], [7], although they face the problem of capturing new words and senses [8], [9]. For the reason mentioned above, here we explore semantic similarity measurement based on pre-existing knowledge resources, like WordNet.

The algorithms for measuring semantic similarity between words using a lexical taxonomy can be divided into distance–based methods and hybrid of information–based and distance-based methods [3], [10], [11].

The basic idea of the distance–based methods is to measure the similarity of two words w_1 and w_2 by the distance of the shortest path between concepts of w_1 and concepts of w_2 in a lexical taxonomy [5], [10], [11]. This approach assumes that the shorter the distance, the more similar the concepts are. The simple distance–based methods such as Rada's edge counting method [5], measure the length of the shortest path between two words only in the hierarchy. Some methods add some correcting factors to take into account, the depth in the hierarchy, the density of the sub hierarchies, descriptive glosses and so on [10]-[13]. Alvarez and Lim's method first built a rooted weighted graph by exploiting two input words w_1 and w_2 and their corresponding concepts, relationships, and descriptive glosses available in WordNet. Then it explored the concepts presented in the graph and selected the minimal distance between any two concepts c_1 and c_2 of w_1 and w_2 respectively. Their definition of distance between two concepts is the combination of the depth

of the nearest common ancestor between two concepts, the intersection of the descriptive glosses of two concepts and the shortest distance between two concepts [10]. Wu and Palmer's method calculated conceptual similarity between two concepts by considering the depths of the two concepts in the WordNet hierarchy, along with the depth of the lowest common super concept of the two concepts [12]. Qin *et al*. combined semantic distance between two words with feature information in DAG (Directed Acyclic Graph) to measure the semantic similarity between two words in the hierarchy of WordNet [11].

Hybrid methods evaluate the semantic similarity using external information, such as word frequencies and / or information content extracted from available corpora in addition to the hierarchical information related to the corresponding concepts in the underlying taxonomy. For example, Li *et al*. proposed a method for semantic similarity which combined path length, node depth, and local semantic density. Their path length and depth were derived from a lexical database, while the local semantic density was from a corpus [3]. Jiang and Conrath's similarity measurement is a combined approach that inherited the edge-based approach of the edge counting scheme, which was then enhanced by the node-based approach of the information content calculation [14]. The cornerstone of this kind of methods is the accurate estimation of the information content of concepts. Some information–based methods generate corpora using a web search engine or crawler [1], [4], [8].

The algorithms to measure semantic similarity of Chinese words [15]-[18] and Chinese-English Cross-Lingual words [19] usually base on HowNet. HowNet is an on-line common-sense knowledge base that unveils inter-conceptual relations and the inter-attribute relations of concepts connoting in the lexicons of the Chinese and their English equivalents [20]. Unlike WordNet, HowNet defines a word in a complicated multi-dimensional knowledge description language. The description of each word consists of a group of sememes. Sememes are the smallest basic semantic units to describe concepts. In HowNet, the semantic relationship between Chinese words is given by concept-sememe relationships and sememe-sememe relationships. Semantic relationships between sememes are hidden separately in eleven concept feature files. To quantify semantic similarity between Chinese words, Qun, L. and Sujian, L. rewrote the HowNet definition of a word in a structural format, and then gave an algorithm for measuring semantic similarity between Chinese words based on their structural format [15]. Yi, G., *et al*. built a sememe network from HowNet in advance, by which to extract the semantic paths between two words, and then computed their semantic similarity on the basis of quantifying the semantic paths [16]. Min J. *et al*. improved Qun, L. and Sujian, L.'s method by using sememes' depth information, the antonym and definition information of the sememe [17]. Li D. and Heyan, H. focused on English-Chinese cross-lingual scenarios in paper [19]. Their basic idea is to compute the similarity between words by exploring their attributes and relations. Given a word pair, they first utilized bi-lingual knowledge base HowNet to locate their attributes and concepts, and then computed similarities between their attributes by combining distance, depth and relation information. Finally they used a combination scheme to compute semantic similarity of the word pairs.

3. Algorithm

Semantic similarity between two words is often represented by the similarity between concepts associated with these two words. To measure Chinese-English bilingual word semantic similarity, we use a pre-existing knowledge resource Chinese WordNet.

WordNet is a well-known English lexical database designed for the use under program control. It was developed under the direction of George A. Miller [21]. English nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (called synsets), each expresses a lexicalized concept. A polysemous word will appear in one synset for each of its senses. Synsets are linked to each other by

numerous semantic relations like hyponymy (the semantic relation of being subordinated or belongs to a lower rank or class), meronymy (the semantic relation that holds between a part and the whole), and entailment (logical inference). Relations between concepts and words in WordNet are made explicit and are labeled so that users can select a specific relation to guide them from one concept to another. WordNet fits semantic similarity measures very well because its hyponymy relation between synsets organizes nouns and verbs into hierarchies of is–a relations. This is-a relation connects a hyponym (a noun or verb in a more specific synset) to a hypernym (a noun or verb in a more general synset). Previous work has proved that the minimum number of edges separating two concepts c_1 and c_2 is a metric that measures the conceptual distance of c_1 and c_2 in a semantic net of hierarchical relations [5]. Lexical taxonomy may have irregular densities of links between concepts. This no uniformity problem can be corrected by using the node depth in the hierarchy where the word is found, and the local density of the sub hierarchies and the type of link [3], [14].

Chinese WordNet is an expansion of WordNet in Simplified Chinese developed by the Department of Computer Science and Engineering at Southeast University of China and the Department of Computer Science at Vrije Universiteit Amsterdam of Netherlands [22]. It is translated from WordNet manually. The mapping between the Chinese part and the English part is established base on the synsets. A Chinese synset is corresponding to an English synset. Chinese WordNet has brought in approximate 118,000 Chinese words and 115,424 synsets to fulfill the Chinese-Chinese function in the short term. It has the function of English-English search and the special search, such as antonym, hypernym and so on.

A word has one or more senses. This means that a word belongs to one or more synsets in WordNet. To calculate the semantic similarity between the given words w_1 and w_2, we first search their corresponding synsets in WordNet, and then measure the similarity between any two synsets s_1 and s_2 of w_1 and w_2 respectively. Finally, we get the similarity between words w_1 and w_2 based on some top synset similarities of all pairs of synsets.

3.1. Similarity between Synsets

Hypernym-hyponym relations (also called super-subordinate or ISA relation) group synsets in WordNet into 11 tree-like hierarchical structures from many specific terms at the lower levels to a few generic terms on the top. In our algorithm, a directed and acyclic graph G_{ISA} is built to represent the hypernym-hyponym relations among synsets in WordNet. In the graph, a node corresponds to a synset and a directed edge <u, v> represents that u is a hypernym of v.

Our method to compute semantic similarity between synsets is based on following ideas:

1) In the graph of hypernym-hyponym relation of WordNet, the longer the distance between two synsets is, the less similar they are on semantic.
2) The higher synset sibling node density, in other words, the larger the number of synset sibling node, the less the shared information between sibling nodes. Therefore, the similarities between synset sibling nodes must be less.
3) Synsets at upper layers of the hierarchy have more general semantics and less similarity among them, while synsets at lower layers have more concrete semantics and more similarity.

Therefore, the following three factors, distance, density and depth, are introduced to calculate the similarity between two synsets s_1 and s_2.

- Distance factor: the distance factor is defined by (1).

$$\sigma = \begin{cases} \sqrt{1-\frac{dis^2}{\theta^2}}, & dis < \theta \\ 0, & dis \geq \theta \end{cases} \quad (1)$$

where, *dis* is the distance between synsets s_1 and s_2, namely the shortest path length between synsets s_1 and s_2. θ is a threshold.

The value of σ varies with the distance and is within the range of 0 to 1. The longer the distance, the smaller σ is. The value of σ will be 0, when distance is greater than θ.

- Density factor: the density factor of the pair of synsets s_1 and s_2 is defined by (2).

$$\varphi = (\varphi_1 + \varphi_2)/2 \tag{2}$$

where, φ_1 and φ_2 are the density factors of the synsets s_1 and s_2 respectively. They are defined by formula (3).

$$\varphi_i = \frac{1}{\ln LN_i + 1} \quad (i = 1, 2) \tag{3}$$

where, LN_i is computed by Eq. (4).

$$LN_i = LN_{i1} + \frac{LN_{i2}}{LN_{i1}} + \frac{LN_{i3}}{LN_{i2}} \quad (i = 1, 2) \tag{4}$$

where, LN_{i1}, LN_{i2} and LN_{i3} are the number of nodes on the level of the synset s_i, one level and two levels above the synset s_i (i=1, 2) respectively. If the parents of s_1 and s_2 are the same, then set LN_{i2} and LN_{i3} to zero. If the grandparents of s_1 and s_2 are the same, then only set LN_{i3} to zero.

The bigger the LN_i, the smaller φ_i is. The value of φ is greater than zero and less than one.

- Depth factor: the formula we use to compute depth factor for the pair of synsets s_1 and s_2 is shown as (5).

$$\omega = (\omega_1 + \omega_2)/2 \tag{5}$$

where, ω_1 and ω_2 are the depth factors of the synsets s_1 and s_2 respectively. They are defined by (6).

$$\omega_i = \begin{cases} \dfrac{\sqrt{dep_i - E_d}}{E_d}, & dep_i \geq E_d \\ -\dfrac{\sqrt{dep_i - E_d}}{E_d}, & dep_i < E_d \end{cases} \quad (i = 1, 2) \tag{6}$$

where, dep_i (i=1, 2) is the length of the path from the synset s_i (i=1, 2) to the root in the WordNet hierarchy, E_d is the average depth of the synsets in WordNet.

If the depth of the synset is greater than the average depth of all synsets, the ω_i will be positive. Otherwise, it would be negative.

Combining above 3 factors, the semantic similarity between synsets s_1 and s_2 is defined by (7).

$$sim(s_1, s_2) = \sigma + \alpha \times \varphi + \beta \times \omega \tag{7}$$

where, α and β are the weights of density factor and depth factor respectively.

We set the semantic similarity to one, if the $sim(s_1, s_2)$ is greater than one.

3.2. Semantic Similarity between Words

A word may have more than one meaning, namely belong to more than one synset. Therefore, our algorithm to calculate the semantic similarity between words includes the following steps, as shown in algorithm 1.

Given Chinese or English words w_1 and w_2,

1) Find all of synset_ids for each given word by search Chinese WordNet database.
2) Translate each synset_id of Chinese word to the corresponding English synset_ids.
3) Suppose w_1 has *m* synsets and w_2 has *n* synsets, there will be $m \times n$ synset pairs. Calculate the similarity for each pair of synsets.
4) Sort the similarities of all synset pairs in descending order. Let the weight of the first similarity be ρ, the weight of the second similarity be $(1-\rho)\times\rho$, namely the remaining weight multiply ρ, and so on. The weight of the ith similarity is (1-sum of first i-1 weights)× ρ. We set a threshold δ ($0<\delta<1$) and take the top $\delta\times100\%$ similarities of synset pairs to calculate the similarity between words.

To make the weight varies with the number of the combination pairs of the two words' synsets, we use (8) to determine ρ, and (9) to determine the weights of similarities of synset pairs.

$$\rho = \begin{cases} 0.9 & num \le 3 \\ 0.9 - 0.02 \times (num - 3) & 3 < num \le 23 \\ 0.5 & num > 23 \end{cases} \tag{8}$$

$$w_i = \begin{cases} (1 - \sum_{j=1}^{i-1} w_j) \times \rho & i > 1 \\ \rho & i = 1 \end{cases} \tag{9}$$

where, *num* is the number of pairs of two words' synsets.

The value of ρ will vary by *num*. The smaller the *num* is, the larger the ρ will be, and $\rho \in [0.5, 0.9]$.

Algorithm 1. Chinese-English bilingual word similarity measure
Input: word pair (w_1 , w_2);
hypernym-hyponym relation graph $G_{ISA;}$
the weights of density and depth factor α and β;
threshold of distance θ,
threshold of top synset similarity number δ
Output: a similarity score of (w_1, w_2)

```
S1 ← Get_synset(Chinese WordNet database, w1);
S2 ← Get_synset(Chinese WordNet database, w2);
for each word wi in (w1, w2) do
    if wi is Chinese word then
        for each synset_id ∈ Si do
      translate Chinese synset_id to a corresponding English synset_id;
        end for
    end if
end for
SIM ← {}
for each synset s1 ∈ S1 do
    for each synset s2 ∈ S2 do
        get_distance(GISA; s1, s2);
        SIM ←SIM ∪ { sim(s1, s2) };
    end for
```

15. **end for**
16. sort (SIM) in descending order;
17. weight the top δ×100% similarities of synset pairs by (8) and (9);
18. **return** weighted sum of top δ×100% similarities

4. Experiments

We set up three groups of experiments to show that our method is capable of measuring semantic similarity for Chinese word pairs, Chinese-English cross-lingual word pairs and English word pairs.

The quality of a method to calculate semantic similarity between words can only be established by comparing its results with human common sense. Most similarity researchers have published their results in a dataset of 30 English word pairs with human similarity ratings. The dataset was given by Miller and Charles, usually is referred to MC Dataset. In 1991, Miller and Charles gave 38 undergraduate students 30 word pairs and asked the students to rate the word pairs for "similarity in meaning" on a scale from 0 (no similarity) to 4 (perfect synonymy). The average rating of each pair represents a good estimate of how similar the two words are. Resnik replicated the task with a different group of students in 1995 and found a correlation between the two ratings of r=0.9011 for the 28 word pairs tested.

In our experiments of Chinese-English cross-lingual word semantic similarity and English word pair semantic similarity, we also compute word similarity on MC Dataset. In the experiments of Chinese-English cross-lingual word semantic similarity, we translated the second word in each pair in the MC Dataset into Chinese. We chose the first or second translation from the Oxford Advanced Learner's English-Chinese Dictionary (7th edition).

In the following experiments, we experientially set the following parameters:
$\theta = 7$ for Distance factors.
$\alpha = 0.1$ as the weight of density factor.
$\beta = 0.1$ as the weight of depth factor.
$\delta = 0.2$, namely 20% as the rate of selected similarities of pairs of synsets.

4.1. Similarity between Chinese Word Pair

We first measured some similarities between Chinese word pairs and compared to the results from our method with the results from the method proposed in paper [15]. Part of the results is shown in Table 1. In Table 1, the column labeled **method 1** is the results measured by the method proposed in paper [15] and the column labeled **method 2** is the results measured by our method proposed in this paper.

Table 1. Some Similarities between Chinese Words

Word 1	Word 2	Method 1	Method 2
论文(paper)	文章 (article)	1.000	0.9295
论文(paper)	文献(literature)	0.167	0.9131
论文(paper)	文集 (corpus)	0.167	0.2946
论文(paper)	文字(character)	1.000	0.7433
论文(paper)	书籍(book)	0.151	0.8809
文化(culture)	文学(Literary culture)	0.171	0.6796
文化(culture)	语言(language)	0.444	0.7179

The experiment shows that the similarities measured by method 2 are closer to human common sense. For example, in Table 1, both Chinese words "文章"(article) and "文献"(literature) are close to "论文"(paper) in meaning, the similarities of the word pairs "论文-文章"(paper-article) and "论文-文献"(paper-literature) computed by method 2 are 0.929 and 0.913 respectively, but the similarities computed by method 1 are quite different. The similarity of word pair "论文-文章" (paper-article) computed by method 1 is 1.000, but the similarity of pair "论文-文献" (paper-literature) is 0.167.

Table 2. Some Chinese-English Cross-Lingual Word Similarities

Word 1	Word 2	Similarity
父亲 (father)	father	1.0000
父亲 (father)	dad	0.9707
computer	计算机 (computer, calculator)	0.9847
computer	电脑(computer, electronic brain)	0.9117
car	电脑(computer, electronic brain)	0.2690
中国(China)	Asia	0.5337
中国(China)	Europe	0.4699
Asia	欧洲(Europe)	0.9396
cat	狗(dog)	0.7977
猫(cat)	animal	0.7100
狗(dog)	car	0.2661
文化(culture)	language	0.7176

Table 3. Chinese-English Cross-Lingual Word Similarities for the Translated MC Dataset

English Word Pair		Humans(on a scale from 0 to 4)		English-Chinese Word Pair		Similarity(on a scale from 0 to 1)
		Miller & Charles	Resnik			
automobile	car	3.92	3.90	automobile	轿车	0.8896
gem	jewel	3.84	3.50	gem	宝石	0.9993
journey	voyage	3.84	3.50	journey	航行	0.9502
boy	lad	3.76	3.50	boy	小伙子	0.9751
coast	shore	3.70	3.50	coast	岸	0.9648
asylum	madhouse	3.61	3.60	asylum	精神病院	1.0000
magician	wizard	3.50	3.50	magician	向导	0.9410
midday	noon	3.42	3.60	midday	正午	0.9922
furnace	stove	3.11	2.60	furnace	火炉	0.8393
food	fruit	3.08	2.10	food	成果	0.8184
bird	cock	3.05	2.20	bird	公鸡	0.7209
bird	crane	2.97	2.10	bird	鹤	0.4771
implement	tool	2.95	3.4	implement	用具	0.7472
brother	monk	2.82	2.4	brother	和尚	0.8833
crane	implement	1.68	0.30	crane	实现	0.4308
brother	lad	1.66	1.20	brother	小伙子	0.6985
car	journey	1.16	0.70	car	旅行	0.0313
monk	oracle	1.10	0.80	monk	神谕	0.0297
cemetery	woodland	0.95		cemetery	林地	0.653
food	rooster	0.89	1.10	food	公鸡	0.0310
coast	hill	0.87	0.70	coast	小山丘	0.4552
forest	graveyard	0.84	0.60	forest	坟墓	0.3881
shore	woodland	0.63		shore	林地	0.5934
monk	slave	0.55	0.70	monk	苦工	0.4923
coast	forest	0.42	0.60	coast	森林	0.2737
lad	wizard	0.42	0.70	lad	向导	0.7510
chord	smile	0.13	0.10	chord	微笑	0.0452
glass	magician	0.11	0.10	glass	巫师	0.0407
rooster	voyage	0.08	0.00	rooster	航行	0.0463
noon	string	0.08	0.00	noon	细绳	0.0393

4.2. Chinese-English Cross-Lingual Word Similarity

One of the main properties of our algorithm is Chinese and English bilingual similarity measurement. The algorithm can measure not only Chinese word semantic similarity, but also English word semantic similarity and Chinese-English cross-lingual word similarity. Table 2 gives some of Chinese-English cross-lingual word semantic similarities measured by our algorithm.

In Table 2, the semantic similarity between Chinese word "父亲" and English word "father" measured by our algorithm is 1, the maximum value, because both "父亲" and "father" not only are identical in meaning, but also belong to formal sort of words. The similarity between Chinese word "父亲" and English word "dad" is 0.9707, which is lower than the similarity between "父亲" and "father", because different from "父亲", "dad" is a colloquial word even though "父亲" and "dad" have the same meaning. For the same reason, the similarity between English word "computer" and Chinese word "计算机" is 0.9847. It is higher than the similarity between English word "computer" and Chinese word "电脑". The similarity between "computer" and "电脑" is 0.9117.

In Table 3, column "similarity" shows our results (on a scale from 0 to 1) for the translated MC Dataset. The first four columns list the English word pairs and their associated the human judgments (on a scale from 0 to 4) in the MC Dataset. The third and fourth columns are the average ratings given by Miller and Charles's students and Resnik's students respectively. Most results measured by our algorithm are closer to the human judgments.

4.3. Similarity between English Word Pair

We compare three methods with our method in the experiments of measuring semantic similarity between English words. Table 4 shows the results of each similarity measure for each word pair in MC Dataset. The first four columns list MC Dataset and the corresponding human ratings. Column "Sim_{WP}", "Sim_{L}" and "Sim_{QL}" show the semantic similarities measured by Wu and palmer's method, Lin's method, Qin and Lu's method respectively, which was given by paper [11]. In paper [11], Qin *et al.* used an independent software package developed by Ted Pedersen to calculate similarity based on WordNet2.1. The package involves similarity measures described by Lin, Wu and Palmer, etc. The last column in Table 4 is our results for the MC Dataset. They are match with human common sense in general.

Table 4. SimilaritIes of English Word Pairs in MC Dataset

Word Pair		Humans (on a scale from 0 to 4)		Algorithms (on a scale from 0 to 1)			
		Miller &Charles	Resnik	SimWP	SimL	SimQL	Our method
automobile	car	3.92	3.90	1.000	1.0000	1.0511	0.9203
gem	jewel	3.84	3.50	1.000	1.0000	1.0287	0.8846
journey	voyage	3.84	3.50	0.9565	0.8277	0.9280	0.9636
boy	lad	3.76	3.50	0.9333	0.7979	0.6826	0.9700
coast	shore	3.70	3.50	0.9231	0.9632	0.9643	0.7295
asylum	madhouse	3.61	3.60	0.9565	0.9813	1.0411	0.9426
magician	wizard	3.50	3.50	1.0000	1.0000	1.0018	0.9630
midday	noon	3.42	3.60	1.0000	1.0000	1.0000	0.9922
furnace	stove	3.11	2.60	0.5714	0.2294	0.4790	0.8141
food	fruit	3.08	2.10	0.4706	0.1559	0.8044	0.7092
bird	cock	3.05	2.20	0.9565	0.7881	0.9214	0.7209
bird	crane	2.97	2.10	0.8800	0.0000	0.8982	0.4771
implement	tool	2.95	3.4	0.9412	0.9146	0.9388	0.4980
brother	monk	2.82	2.4	0.9565	0.2097	0.4185	0.8833

crane	implement	1.68	0.30	0.7778	0.3327	0.5697	0.4567
brother	lad	1.66	1.20	0.7143	0.2400	0.5440	0.7631
car	journey	1.16	0.70	0.1905	0.0000	0.2929	0.0200
monk	oracle	1.10	0.80	0.5882	0.1828	0.4132	0.0297
cemetery	woodland	0.95		0.5000	0.1119	0.3459	0.6533
food	rooster	0.89	1.10	0.2857	0.0762	0.2865	0.0334
coast	hill	0.87	0.70	0.7143	0.7286	0.6944	0.4231
forest	graveyard	0.84	0.60	0.5000	0.1119	0.3565	0.6256
shore	woodland	0.63		0.6667	0.1220	0.4303	0.5934
monk	slave	0.55	0.70	0.7143	0.2011	0.4461	0.7756
coast	forest	0.42	0.60	0.6154	0.1181	0.3398	0.3437
lad	wizard	0.42	0.70	0.7143	0.2241	0.5073	0.7841
chord	smile	0.13	0.10	0.3750	0.3269	0.4270	0.0369
glass	magician	0.11	0.10	0.5333	0.1421	0.3979	0.0193
rooster	voyage	0.08	0.00	0.1481	0.0000	0.2502	0.0339
noon	string	0.08	0.00	0.3529	0.0923	0.3258	0.0396

5. Conclusion

An algorithm of calculating Chinese-English bilingual word semantic similarity based on Chinese WordNet is presented in this paper. Experimental results show that evaluating semantic similarity of Chinese-English bilingual words using Chinese WordNet is feasible and our method to calculate word semantic similarity is reasonable. Setting the weight of the synset pairs of two words in an adaptive mode and using the mapping between Chinese synsets and English synsets to implement the cross-lingual similarity measurement are the major advantages of our algorithm. We have applied this algorithm to data semantic similarity analysis in data spaces and will further evaluate it in more experiments.

Acknowledgment

This work is partially supported by the Key Projects of Science and Technology of Fujian Province of China under Grant No.2011H6016 and 2011H0028.

References

[1] Eneko, A., Enrique, A., Keith, H., Jana, K., Marius, P., & Aitor, S. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19-27). Boulder, Colorado: Association for Computational Linguistics.

[2] Michael, J. F., Alon, Y. H., & David, M. (2005). From databases to data spaces: A new abstraction for information management. *SIGMOD Record*, *34(4)*, 27–33.

[3] Yuhua, L., Zuhair, A. B., & David, M. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge And Data Engineering*, *15(4)*, 871-882.

[4] Rudi, L. C., & Paul, M. B. V. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, *19(3)*, 370-383.

[5] Rada, R., Mili, H., Bichnell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transations on Systems, Man, and Cybernetics*, *9(1)*, 17-30.

[6] Dekang, L. (1998). Automatic retrieval and clustering of similar words. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on*

Computational Linguistics (pp. 768-774). Montreal: Association for Computational Linguistics.

[7] Alexander, B., & Graeme, H. (2006). Evaluating WordNet based measures of lexical semantic relatedness. *Computational Linguistics*, *32(1)*, 13–47.

[8] Danushka, B., Yutaka, M., & Mitsuru, I. (2011). A web search engine-based approach to measure semantic similarity between words. *IEEE Transactions on Knowledge and Data Engineering*, *23(7)*, 977-990.

[9] Guangjun, H., Shuili, W., & Xiaoguo, Z. (2011). Query expansion based on associated semantic space. *Journal of Computers*, *6(2)*, 172-177.

[10] Marco, A. A., & SeungJim, L. (2007). A graph modeling of semantic similarity between words. *Proceedings of International Conference on Semantic Computing* (pp. 335-362). Irvine, California.

[11] Peng, Q., Zhao, L., Yu, Y., & Fang, W. (2009). A new measure of word semantic similarity based on WordNet hierarchy and DAG theory. *Proceedings of International Conference on Web Information Systems and Mining* (pp. 181-185). Shanghai.

[12] Zhibiao, W., & Martha, P. (1994). Verb semantics and lexical selection. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* (pp. 133–138). Las Cruces, New Mexico.

[13] Giannis, V., Epimenidis, V., & Paraskevi, R. (2005). Semantic similarity methods in WordNet and their application to information retrieval on the Web. *Proceedings of the 7th ACM International Workshop on Web Information and Data Management* (pp. 10-16). Bremen: ACM.

[14] Jay, J. J., & David, W. C. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference Research on Computational Linguistics* (pp. 1-15). Taiwan: ROCLING X.

[15] Qun, L., & Sujian, L. (2005). Word semantic similarity computation based on HowNet. *Proceedings of the 3rd Chinese Lexical and Semantic Proseminar* (pp. 59-76). Taibei.

[16] Yi, G., Jian, Z., *et al*. (2002). Quantifying semantic similarity of Chinese words from HowNe. *Proceedings of the 1st International Conference on Machine Learning and Cybernetics* (pp. 234-239). Beijing.

[17] Min, J., Shuicai, S., *et al*. (2008). An improved word similarity computing method based on HowNet. *Journal of Chinese Information Processing*, *22(5)*, 84-89.

[18] Peng, W., *et al*. (2012). Applications of text clustering based on semantic body for Chinese spam filtering. *Journal of Computers*, *7(11)*, 2612-2616.

[19] Lin, D., & Heyan, H. (2011). An English-Chinese cross-lingual word semantic similarity measure exploring attributes and relations. In Helena, H. G., & Minghui, D. (Eds.), *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation* (pp.467-476). Singapore: Nanyang Technological University.

[20] HowNet[EB/OL]. Retrieved from http://www.keenage.com/

[21] WordNet[EB/OL]. Retrieved from http://wordnet.princeton.edu/

[22] Chinese WordNet. Retrieved from http://www.aturstudio.com/wordnet/windex.php

Yangyang Wu received her B.S in computer science from Fuzhou University, 1982. She is a professor of the Department of Computer Science and Technology, Huaqiao University, China.

Prof. Wu is a senior member of China Computer Federation. She has authored or coauthored more than 70 scientific articles papers in Chinese or English Journals and Proceedings. Her research interests are data management and data mining.

Siying Wu is a graduate student of the Department of Computer Science and Technology, Huaqiao University, Xiamen, China.

Duansheng Chen received his bachelor's degree from Zhejiang University, 1982 and the PhD degree in information and communication engineering from University of Science and Technology of China, 2005. He is a professor of the Department of Computer Science and Technology, Huaqiao University, China.

Prof. Chen is a senior member of China Computer Federation. He has authored or coauthored more than 80 scientific articles papers in Chinese or English Journals and Proceedings. His research fields include information processing and machine learning.