

An English-Chinese Cross-lingual Word Semantic Similarity Measure Exploring Attributes and Relations *

Lin Dai and Heyan Huang

Department of Computer, Beijing Institute of Technology
Beijing 100081, China.
{dailiu,hhy63}@dailiu@bit.edu.cn

Abstract. Word semantic similarity measuring is a fundamental issue to many NLP applications and the globalization has made an urgent request for cross-lingual word similarity measure. This paper proposed a word semantic similarity measure which is able to work in cross-lingual scenarios. Basically, a concept can be defined by a set of attributes. The basic idea of this work is to compute the similarity between words by exploring their attributes and relations. For a given word pair, we first compute similarities between their attributes by combining distance, depth and relation information. Then word similarity are computed through a combination scheme. The algorithm is implemented based on an English-Chinese bilingual ontology *HowNet*. Experiments show that the proposed algorithm results in high correlation against human judgments, which encourages its broad application in cross-lingual applications.

Keywords: Word semantic similarity, Cross-lingual, Natural Language Processing, Computing linguistics.

1 Introduction

Word semantic similarity measure plays a fundamental role in many natural language processing (NLP) applications such as information retrieval (IR), machine translation (MT) and word sense disambiguation (WSD). Many research efforts have been made in the past decades to improve the effectiveness of word semantic similarity measures. With the progress of the globalization, information emerged from a variety of languages to the Internet. Under this circumstance, cross-lingual applications such as machine translation, cross-lingual information retrieval, cross-lingual text categorization and clustering etc., become more and more attractive. Consequently, cross-lingual word semantic similarity measure becomes a meaningful research topic.

Although there are some statistical and hybrid measures which have good overall performance (Li *et al.*, 2003; Xia *et al.*, 2011; Hassan and Mihalcea, 2009), the knowledge-based measures, to us, is still an important method. Even in a hybrid measure, the knowledge-based part is still a key part. Furthermore, if a large scale ontological knowledge base is available, the knowledge-based measures can be further improved if the knowledge base is comprehensively investigated and the similarity measures are elaborately designed.

In this work, we “re-examine” merely knowledge based word similarity measure and a cross-lingual semantic similarity measure is proposed. As we know, concepts are language-independent and it is natural that each concept can be defined by several aspects, i.e., attributes. For example, the concept of the animal *dog* can be roughly defined by a *kind-of* attribute with value *livestock* and a *modifier* attribute with value *domesticated*. If all the words in this world are defined in this way, it is possible to compute the semantic similarity of any pair of words in spite of their

* The authors would like to thank Prof. Z.D. Dong for his valuable advices.

language. This is the basic idea of this work. In this work, this idea is implemented by utilizing a bi-lingual ontology, HowNet (Dong and Dong *et al.*, 2006), to locate attributes and concepts.

Although the proposed method can deal with English and Chinese solo-lingual word pairs, this paper focuses on cross-lingual ones. To evaluate our method, a series of experiments were conducted on a cross-lingual benchmark data sets with human ratings. A conclusion could be safely drawn from experimental results that the proposed measure is promising when compared with previous solo-lingual and cross-lingual word similarity measure.

The rest of this paper is organized as follows. Section 2 discusses the related works. The methodology of our method is presented in Section 3. Brief introduction of HowNet and its attribute network are introduced in Section 4. The implementation of the method is presented in Section 5. We evaluate the method in Section 6 before concluding this paper in Section 7.

2 RELATED WORKS

Previous researches of word semantic similarity measure mainly focus on monolingual settings. Early research efforts have been devoted to design the knowledge-based measures, in which word synonym set in thesaurus plays an elementary role to word similarity calculation. In particular, WordNet (Fellbaum, 1998) has been widely adopted in word similarity measures for English due largely to the hierarchically organized synsets for English words. Researchers investigate taxonomic structure or ontological framework, and attempt to calculate the word similarity by counting conceptual distance. For example, (Yang *et al.*, 2005; Alvarez *et al.*, 2007).

Some research efforts have been made to develop the corpus-based measures (Jarmasz, 2003; Terra *et al.*, 2003; Bollegala *et al.*, 2007; Alvarez *et al.*, 2007) or hybrid measures (Li *et al.*, 2003; Strube *et al.*, 2006). The basic idea is to measure dependence between words by using statistics extracted from corpora. As a result, the corpus-based measures are found less effective than the knowledge-based ones in most word similarity measuring evaluations due to coverage problem.

In recent years, HowNet has been investigated by more and more researchers in several applications including word semantic similarity measuring. For example, (Liu and Li, 2005; Lin *et al.*, 2009; Fan and Chen, 2009). These works focused on mono-lingual context and were not evaluated on data sets with human ratings. Newly works began to apply HowNet on English data set (Dai *et al.*, 2008). Xia *et al.* (2011) combine HowNet and parallel corpus to measure cross-lingual word similarity. But their algorithms ignore the relations between sememes, which is proven to be very informative in this paper.

3 BASIC IDEA

Human beings judge similarity and dissimilarity between objects by their attributes. For example, when concept *automobile* and *bicycle* are concerned, people think they are similar because both of them have the *is-a* attributes with value *vehicle*. At the same time, people think they are different because an automobile is *automatic* while a bicycle is *manual*.

There are numerous concepts in this world which are represented by different languages. To define concepts in an efficient and standard manner and make the definitions easy to be handled by computers, we can use a set of fundamental and language independent concepts to define other concepts. We call them meta-concepts. Furthermore, the real world is complex and there are abstract concepts as well as concrete concepts. To define all the concepts properly, meta-concepts must be organized into hierarchical trees.

Because concepts are defined by meta-concepts, meta-concept similarity is extremely important to concept similarity. For a given concept pair $\{c_1, c_2\}$, we propose that the similarity between them $s(c_1, c_2)$ is a function of their feature sets as follows:

$$s(c_1, c_2) = f(fs_1, fs_2) \quad (1)$$

where f_{s_1} and f_{s_2} is the feature set of concept c_1 and c_2 respectively.

Meta-concept similarity is the basis of similarity between feature (i.e., attribute) sets. Distance and depth information of meta-concepts in meta-concept hierarchies are fundamental evidences for computing meta-concept similarities. Given two meta-concepts in a hierarchy, there must exist the shortest path between them, which connects their hypernyms. We call their common hypernym on the shortest path the *subsumer* of them and the length of the shortest path the *distance* between them. Intuitively, meta-concept similarity is a monotonically decreasing function of the distance a monotonically increasing function with respect to the depth of subsumers.

Besides, when computing meta-concept similarity, the impact of various relations should be considered. Actually, besides hypernym-hyponym relation which forms tree hierarchies, there are many other kinds of relations between concepts.

Now, we propose that the similarity between meta-concept mc_1 and mc_2 is a function of the distance l , depth h , and relatedness r , as follows:

$$s(mc_1, mc_2) = f(f_l(l), f_h(h), f_r(r)) \quad (2)$$

where $f_l(\cdot)$, $f_h(\cdot)$ and $f_r(\cdot)$ is the contribution of distance, depth and relation respectively. $f(\cdot)$ is the framework that combine the three parts into final measure.

It worth noting that values of both the distance and depth may cover a large range, while the interval of similarity should be finite with extremes of exactly the same or not similar at all. If we assign exactly the same with a value 1 and no similarity as 0, then the interval of similarity is [0, 1].

4 ATTRIBUTE AND RELATION SYSTEM IN HOWNET

To fulfill the task of this paper, a well structured knowledge base in which concepts are defined by feature(s) is demanded. In this section we briefly introduce the attribute and relation system of HowNet, which meets our request.

As a knowledge base for natural language processing, HowNet provides plenty of taxonomically semantic knowledge as well as real-world knowledge. As a knowledge system that describes relations between concepts, HowNet attempts to construct a net structure of its knowledge base from the inter-concept relations and inter-attribute relations. This is the fundamental distinction between HowNet and other tree-structured lexical databases. The philosophy of HowNet entails its unique structure.

The meta-concept in HowNet is called *sememe*, which is the smallest semantic unit that can't be reduced further and used to define concepts. Through combining sememes selected from a finite sememe set, HowNet can describe infinite concepts. There are 2219 basic sememes in the current version of HowNet which are organized into five taxonomies and four subsidiary taxonomies.

As the sememes are concerned, in addition to being fundamental description units, there are complicated relations among them. The relations include hypernym-hyponym, antonym, converse, whole-part, material-product, etc. These relations are given in two ways. One way is the tree structure which represents hypernym-hyponym in taxonomies. The other way is the definition of sememes themselves. Although sememes are the meta-concepts used to define other concepts, the sememes are again made up by sememes. For example¹, the frame of sememe *bird* is: $\{animal:materialOf=\{edible\}, \{eat:patient=\{\sim\}\}, \{fly:agent=\{\sim\}\}\}$, and the frame of sememe *food* is: $\{edible:\{cook:PatientProduct=\{\sim\}\}\}$. Both of these two frames have a sememe *edible*. By this way, HowNet gives a strong relation between sememe *bird* and *food*.

¹ The current version of HowNet is bilingual, i.e., English and Chinese. Sememes are identified by English sememes and Chinese sememes at the same time. Both English version and Chinese version of sememe set is competent to define words of any language. We only give the English version of the sememes.

From the relations among sememes we can see that the actual structure of HowNet is indeed a net rather than merely a tree. Although there exist tree hierarchies in it, HowNet is totally different to a traditional thesaurus. In a traditional thesaurus such as WordNet, concept is the minimum unit for word sense description and each concept is a node in a hierarchy. But in HowNet, sememes are nodes in hierarchies and each concept is defined by sememes.

HowNet uses *Knowledge Database Markup Language* (KDML, in short) to construct the semantic expression of concepts and sememes. KDML consists of a set of grammar rules and keywords. The detailed introduction of KDML is beyond this work. We give several definition examples in Table 1.

Word	Definition
<i>cock</i>	{bird:modifier={domesticated}{male}}
<i>fruit</i>	{fruit}
<i>asylum</i>	{InstitutePlace:domain={medical},{doctor:content={disease:CoEvent ={mad}},location={~}}
<i>grin</i>	{CausePartMove:PatientPart={part:PartPosition={mouth},whole= {AnimalHuman}}
<i>jewel</i>	{material:MaterialOf={treasure}}
<i>stove</i>	{tool:{WarmUp:instrument={~}},{burn:location={~}}

Table 1: The examples of definition expression of concepts in HowNet

From the examples we can conclude several rules that are necessary for similarity computing.

- The first sememe in any definition expression gives the most essential sense, i.e., *is-a* attribute of corresponding concept.
- A concept has one or more features. Each feature is defined by an attribute-value pair. The attribute symbols are in the front of the equal sign and values go after the equal sign.
- The sign ~ is a special and frequently used meta-concept which means the defined concept itself.

The current version of HowNet includes 100,168 English words and 96,370 Chinese words. Its large coverage makes it competent to act as a knowledge base in many NLP researches.

5 IMPLEMENTATION

From section 4 we know that the nature of HowNet matches the methodology proposed in section 3 very well. This section gives the detailed implementation of the methodology.

5.1 Sememe Similarity

As already discussed, the sememe is the basic unit of word definition and sememe similarity is extremely important to word similarity. We will use function (2) to compute sememe similarity. In (2), there are three information resources, i.e., distance, depth and relation. The performance of different strategies for formula (2) will be experimentally investigated in section 6.3.

5.2 Concept Similarity

Because concepts are defined by frames, concept similarity can be calculated from their definitions. From section 4 we know that a frame is constructed with sub-frames in a nested manner. The first sub-frame of a frame, which is always a simple sememe without attribute label, points out the essential feature of the corresponding concept. It specifies the *is-a* attribute of each concept. Meanwhile, the other sub-frames describe other necessary attributes.

To simplify the computation and make full use of the semantic information provided by frames, we break a frame into a set of *features*. Sememes in all levels are reserved in this feature set.

Each feature is a $\{attribute-label, value\}$ pair. We name the feature from the first sub-frame as *primary feature* and the corresponding sememe as *primary sememe*. And other features are called *secondary feature* and corresponding sememes are called *secondary sememe*. The attribute-label of a secondary feature reserves the attribute-labels of upper features. For example, the feature set of *grin* is $\{\{CausePartMove\}, \{PatientPart, part\}, \{PatientPart-PartPosition, mouth\}, \{PatientPart-whole, AnimalHuman\}\}$.

Similarity of feature pairs is measured by:

$$s_{ft}(ft_1, ft_2) = k * s(s_1, s_2) \quad (3)$$

where $s(\cdot)$ is the similarity of the corresponding sememe pair, and k is a scaling factor decided by their attribute labels. If their attribute labels are the same, k is set to 1.0, or it is set as a decimal with value from 0.0 to 1.0, which reflects the strength of their relations.

Now, frame similarity is converted to similarity between feature sets. When computing the similarity between two given feature sets, following principles are considered.

- Primary features are more informative than secondary features.
- Attribute-label *modifier* and *null* are equal to other attribute-labels.
- For two features both have value of the sign \sim , the similarity between them is 1 if their attribute-label is the same. Otherwise the similarity is a small value, which is experientially set to 0.1 in this work.
- Similarity between any feature set and empty set is 0.

To satisfy the above principles, we use the following formula to compute concept similarity.

$$SimC(C_1, C_2) = \begin{cases} s_{set}(fs_1, fs_2) & \text{if } s_{ft}(pft_1, pft_2) = 0 \\ s_{ft}(pft_1, pft_2) \cdot s_{set}(fs_1, fs_2) & \text{if } s_{ft}(pft_1, pft_2) > 0 \end{cases} \quad (4)$$

where fs_1, fs_2 is the feature set of concept C_1 and C_2 , pft_1 and pft_2 is the primary feature of concept C_1 and C_2 .

At last we compute the similarity of two feature sets $s_{set}(\cdot)$. The measure of feature-set similarity must satisfies two basic requirements.

- (1) Similarity between two exactly same sets should be 1.
- (2) Given two sets, both have n elements and there are z identical elements. Their similarity should be z/n .

Based on the understanding of above principles, we use following steps to compute feature-set similarity.

- (1) Find the feature-pair with the highest similarity from two feature sets. Accumulate their similarity value to sim_{sum} .
- (2) Delete the two features from their original feature sets.
- (3) Repeat 1 and 2, until any one feature set is empty.
- (4) The final similarity is: $2 * sim_{sum} / (m + n)$, where sim_{sum} is the accumulated feature-pair similarity, m and n are the size of two feature sets respectively.

5.3 Word Semantic Similarity

Word semantic similarity is decided by concepts carried by them. For the case of polysemy, we select the maximum concept similarity as word similarity, which matches the psychology of human beings. Let $C_1 = \{C_{1i}\}(i = 1, \dots, m)$ and $C_2 = \{C_{2j}\}(j = 1, \dots, n)$ denote concept sets of word W_1 and W_2 , respectively. The similarity between word W_1 and W_2 is:

$$SimW(W_1, W_2) = \arg \max_{i,j} \{SimC(C_{1i}, C_{2j})\} \quad (5)$$

6 Evaluation and Discussion

6.1 Data Set

To evaluate an algorithm for word semantic similarity, we should investigate its performance against human common sense. A commonly used word-pair set is constructed by Rubenstein *et al.* (1965) for evaluating word semantic similarity measures (denoted as RG65). Twenty five years later, Miller *et al.* (1991) selected 30 pairs(denoted as MC30) from RG65 data set to form another data set. Earlier researchers use the 28 word pairs (denoted as MC28) of the Miller-Charles set due to the absence of two word pairs from WordNet.

The data sets mentioned above are monolingual which can't be directly used in this work. Up to now, there is a critical lack of a benchmark set for the evaluation of cross-lingual similarity measures. To evaluate the performance of our algorithm under cross-lingual circumstances, we constructed a English-Chinese bilingual data set (denoted as EC62) from RG65. For each word pair, we form 2 pairs by translating the first and second word into Chinese respectively. There are 2*62 word pairs because the word *oracle* is not presented in the current version of HowNet. To make EC62 reliable, Native speakers of Chinese who are also highly proficient in English, were asked to translate the words in the two data sets. They were asked to disambiguate the words and avoid multi-word expressions, slang or culturally-biased terms in their translations.

6.2 Tasks

As discussed above, useful information for measuring word similarity includes the distance between sememes, depth of consumers, and relations between concepts. We believe that, to get a good similarity measure, all these information should be taken into account through a reliable scheme. This section experimentally investigates the contribution of different information to deduce a good combination strategy of these information.

Researchers evaluated the performance of word similarity measures on benchmark data sets. The best way of evaluate similarity measure on a data set is to perform training-testing framework. But the size of RG65 is limited, to train and test algorithms on fixed subsets of it will causes overfitting and bias. In this work, a cross-validation scheme is used to get the performance. We randomly divide EC62 into training set and testing set. The size of training sets is 84 and the size of testing sets is 40. For each strategy, its parameters are tuned on training sets, and it is tested on testing sets. Similar to most works, we adopt the correlation coefficient² as evaluation criteria. Each strategy will be evaluated 10 times and the average correlation on testing set is recorded as its performance.

6.3 Experimental Results

From section 5.2 we know that sememe similarity is the key part of word similarity measure. In this section, we experimentally investigate the performance of sememe similarity strategies. The parameter k of feature similarity function (3) is fixed at 0.8 through out the experiments according to our previous experiences.

Strategy 1. Sememe similarity measure is linear and exclusively decided by the shortest distance between sememes. A segmental linear function is investigated and sememe similarity is defined as

$$S_1(s_1, s_2) = \max\{0, 1 - \frac{l}{L}\} \quad (6)$$

where L is a parameter which makes sememe similarity to be 0 when l is larger than L . This formula ensures the sememe similarity value falls into $[0, 1]$. The correlation coefficient with optimal L between this strategy and human judgments of Rubenstein-Goodenough's is 0.8594.

² Let x and y the variables with covariance σ_{xy} and standard deviations σ_x and σ_y . The correlation coefficient between x and y is $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$.

This value proves that the shortest path length is an effective information source for word similarity measure.

Strategy 2. Sememe similarity measure is linear and exclusively based on the depth of the subsumer of given sememe pair. A segmental linear functions is used in this strategy.

$$S_2(s_1, s_2) = \min\left\{1, \frac{h}{H}\right\} \quad (7)$$

where H is a parameter which makes sememe similarity being 1 when h is larger than H . This formula ensures that sememe similarity value falls into $[0, 1]$. This strategy is plausible because it reflects the truth that higher subsumer results in more abstract meaning the two sememes share and vice versa. The correlation coefficient with optimal H between the results of this strategy and human ratings is 0.8443. This value proves that the depth of subsumers is an effective information source for word similarity measure.

Strategy 3. Sememe similarity is merely decided by the relationship between given sememe pair. We can simply model the contribution of relations between two sememes as a segmental linear function:

$$S_3(s_1, s_2) = \min\left\{1, \frac{D_r}{R}\right\} \quad (8)$$

where D_r is the depth of the deepest relation-sememe, R is a parameter that makes similarity to be 1 when D_r is larger than R . D_r is set to zero if there is no relation-sememe. The correlation coefficient with optimal R between the results of this strategy and human ratings is 0.8343. This value proves that the relation between sememes should be considered when designing word similarity measure.

Strategy 4. Strategy 1 is nonlinearly combined with Strategy 2 through multiplication. This strategy considers both the shortest path and depth of subsumer.

$$S_4(s_1, s_2) = S_1(s_1, s_2) \cdot S_2(s_1, s_2) \quad (9)$$

The average correlation coefficient between the results of this strategy and human ratings is 0.8713, greater than that of Strategy 1, 2 and 3. This experiment illustrates that a simple combination of the shortest path length and the depth of subsumers using multiplication can significantly increase the accuracy of the similarity measure.

Strategy 5. Similar to Strategy 4, sememe similarity measure is a function of the length of the shortest path and the depth of subsumer. But exponential functions are considered in this strategy. the sememe similarity is defined as

$$S_5(s_1, s_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \quad (10)$$

where $\alpha \geq 0$ and $\beta \geq 0$ are parameters scaling the contribution of the shortest path and the depth of subsumer, respectively. The average correlation coefficient between this strategy and human judgments of Rubenstein-Goodenough's is 0.8861, which is greater than that of Strategy 4.

Strategy 6. To consider the shortest path, depth and relation at a same time, Strategy 3 is linearly combined with Strategy 4.

$$S_6(s_1, s_2) = \lambda \cdot S_3(s_1, s_2) + (1 - \lambda) \cdot S_4(s_1, s_2) \quad (11)$$

where λ is the parameter which leverages the weight of S_3 and S_4 . The average correlation coefficient with optimal λ between this strategy and human judgments of Rubenstein-Goodenough's is 0.8372. This value is less than Strategy 4, so this combination of relation contribution with length and depth cannot produce a better similarity measure.

Strategy 7. Strategy 3 is linearly combined with Strategy 5.

$$S_7(s_1, s_2) = \lambda \cdot S_3(s_1, s_2) + (1 - \lambda) \cdot S_5(s_1, s_2) \quad (12)$$

where λ is a parameter which leverages the contribution of S_3 and S_5 . The average correlation coefficient between this strategy and human judgments of Rubenstein-Goodenough's is 0.8499, which does not bring a better correlation than using S_5 alone.

Strategy 8. Strategy 3 is used as a modifier of Strategy 4.

$$S_8(s_1, s_2) = \min\{1, S_4(s_1, s_2) \cdot (1 + \lambda \cdot S_3)\} \quad (13)$$

where λ is a parameter which leverages the impact of S_3 on S_4 . When λ is 0, this strategy is equal to S_4 . The average correlation coefficient between this strategy and human judgments of Rubenstein-Goodenough's is 0.8945. This measure performs nearly at a level of human replication, where correlation between individuals is 0.9015 (Resnik, 1999). The strongest correlation against human similarity judgments is obtained at $\lambda=2$.

Strategy 9. Strategy 3 is used as a modifier of Strategy 5.

$$S_9(s_1, s_2) = \min\{1, S_5(s_1, s_2) \cdot (1 + \lambda \cdot S_3)\} \quad (14)$$

where λ is a parameter which leverages the impact of S_3 on S_5 . When λ is 0, this strategy is equal to S_5 . The average correlation coefficient between this strategy and human judgments of Rubenstein-Goodenough's is 0.8667, less than that of S_5 .

Strategy 10. Relation between sememes is used to improve Strategy 4. An exponential non-linear function is involved in this strategy.

$$S_{10}(s_1, s_2) = \min\{1, S_4(s_1, s_2) \cdot \exp(\lambda \cdot D_r)\} \quad (15)$$

where D_r is the depth of the deepest relation-sememe, λ is a parameter which adjusts the contribution of relation. When λ is set to 0, this strategy is equal to S_4 . The average correlation coefficient between this strategy and human ratings is 0.8469, less than that of S_4 .

Strategy 11. Relation between sememes is used to improve Strategy 5. The same exponential nonlinear function in Strategy 10 is involved.

$$S_{11}(s_1, s_2) = \min\{1, S_5(s_1, s_2) \cdot \exp(\lambda \cdot D_r)\} \quad (16)$$

This strategy is equal to S_5 when λ is set to 0. The average correlation coefficient between this strategy and human judgments of Rubenstein-Goodenough's is 0.9035, which is greater than Strategy 8 and a little bit greater than individual average. The strongest correlation against human similarity judgments is obtained at $\lambda=0.22$.

6.4 DISCUSSION

Referring to results in Section 6.3, we can reach several observations. From the results of S_1 , S_2 and S_3 , we can see that the shortest path, depth of subsumer, and relation between concepts all contributes a lot to word similarity. When compared with previous research results, we can see that with any single factor, the performance is comparable to early work. It is particular worth noting that the results of Strategy 3 indicate the importance of relations in computing word similarity. Strategy 4 and Strategy 5 combines the shortest path length and the subsumer depth in different ways. They have better performance than Strategy 1 and 2 which uses single factor. This proves that combining effective information sources with a proper strategy will produce higher performance. Strategy 6 and Strategy 7 explores three kinds of information at the same time in different ways. But both of them produce lower performance than Strategy 4 and Strategy 5.

This explains that relationship can not be linearly combined with distance and depth to improve algorithm's performance. From Strategy 8 to Strategy 11, relationship is used as a modifier of Strategy 4 or Strategy 5. When we use a linear function of relation to modify Strategy 4 in which distance and depth are used in linear manner (Strategy 8), or when we use a nonlinear function of relation to modify Strategy 5 in which distance and depth are used in nonlinear manner (Strategy 11), we can get higher performance. Otherwise, lower performance will be reached (Strategy 9 and Strategy 10). As Strategy 4 and Strategy 5 are concerned, we can see that Strategy 5 is better than Strategy 4. It also can be seen that Strategy 11 is better than Strategy 8. This proves that nonlinear exponential function works better than linear function on this problem.

As a conclusion, we propose an algorithm for word semantic similarity measuring using Strategy 11 to compute sememe similarity. According to the experimental results, optimal parameters are $\alpha=0.3$, $\beta=0.4$, $\lambda=0.22$. However, since our method combines three kinds of information in a intuitive and experimental way, we believe the performance could be further improved if better combination strategies are found.

The state-of-the-art monolingual algorithms (Alvarez *et al.*, 2007) (Yang *et al.*, 2005) get correlation values around 0.90 against Rubenstein-Goodenough's human ratings, which is very close to the human performance. On the English-Chinese data set EC62, our algorithm reaches an average correlation of 0.9035, which is comparable to monolingual measures of similarity, higher than (Xia *et al.*, 2011)'s result. This grade of performance encourages the use of word similarity measuring algorithms in various cross-lingual applications.

It is worth noting that the proposed method can be competent to English or Chinese solo-lingual word similarity measure, with fairly better performance than cross-lingual context. Due to the lacking of space, we won't give out the detailed result here.

7 CONCLUSION AND FUTURE WORK

This paper presents an attribute and relation based measure to calculate semantic similarity of English-Chinese word pairs. The key feature of this work which differs from the previous ones is that our algorithm explores attributes and semantic relations that connect sememes and concepts. Experiments on English-Chinese cross-lingual data sets show that the performance of the new measure is close (i.e. correlation coefficient value of 0.9035) to human judgments as well as comparable to monolingual word similarity measures. It can thus be concluded that attribute and semantic relation contribute a great deal to word similarity measuring and HowNet is applicable and effective in the task for English-Chinese cross-lingual scenarios.

The inspiring results stimulate a few future works. For instance, we will explore the feasibility to integrate our measure in real NLP applications such as word sense disambiguation, cross-lingual information retrieval, and machine translation.

References

- Alvarez, M.A., and S. Lim. 2007. A Graph Modeling of Semantic Similarity between Words. *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pp.355-362. IEEE Computer Society.
- Bollegala, D., Y. Matsuo, M. Ishizuka. 2007. Measuring semantic similarity between words using web search engines. *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pp.757-766. ACM.
- Dai, L., B. Liu, Y. Xia and S. Wu. 2008. Measuring Semantic Similarity between Words Using HowNet. *ICCSIT '08: Proc. of the International Conference on Computer Science and Information Technology*, pp.601-605. IEEE.

- Dong, Z., and Q. Dong. 2006. *HowNet And the Computation of Meaning*. World Scientific Publishing Co., Inc..
- Fan, X., and J. Chen. 2009. Similarity Computation of Low-frequency Chinese Words. *FSKD '09. Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, pp.524-528. IEEE.
- Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- Hassan, S., and R. Mihalcea. 2009. Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge. *Proc. EMNLP'09*, pp.1192-1201.
- Jarmasz, M., S. Szpakowicz. 2003. Roget's Thesaurus and Semantic Similarity. *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pp.212-219.
- Li, Y., Bandar, A. Zuhair and D. McLean. 2003. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Trans. on Knowl. and Data Eng.*, pp.871-882. IEEE Educational Activities Department.
- Lin, P., Y. Zhang and T. Zhang. 2009. Research on HowNet-based Chinese Word Lexical Semantic Similarity Measurement. *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, pp.282-287.
- Liu, Q., and S. Li. 2005. Word Semantic Similarity Computation based on HowNe. *The 3rd Chinese lexical and semantic proseminar*, pp.17-30.
- Miller, G.A., and W.G. Charles, 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, pp.1-28. Psychology Press.
- Resnik, P. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, pp.95-130.
- Rubenstein, H. and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Commun*, pp.627-633. ACM Press.
- Strube, M., and S. Ponzetto. 2006. WikiRelate! computing semantic relatedness using wikipedia. *AAAI'06: proceedings of the 21st national conference on Artificial intelligence*, pp.1419-1424. AAAI Press.
- Terra, E. and C. Clarke. 2003. Frequency estimates for statistical word similarity measures. *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp.165-172. Association for Computational Linguistics.
- Yang, D., and D. Powers. 2005. Measuring semantic similarity in the taxonomy of WordNet. *ACSC '05: Proceedings of the Twenty-eighth Australasian conference on Computer Science*, pp.315-322. Australian Computer Society, Inc..
- Xia, Y., T. Zhao, J. Yao, and P. Jin. 2011. Measuring Chinese-English Cross-Lingual Word Similarity with HowNet and Parallel Corpus. *Proc. CICLing 2011, Part II*, pp.221-233. LNCS.