International Journal of Intelligence Science, 2016, 6, 17-28
http://www.scirp.org/journal/ijis
ISSN Online: 2163-0356
ISSN Print: 2163-0283

The Research of Chinese Words Semantic Similarity Calculation with Multi-Information

Rihong Wang, Chenglong Wang, Ying Xu, Xingmei Cui

Computer Engineering Institute, Qingdao University of Technology, Qingdao, China
Email: rihongw@126.com, sdwfwcl891025@126.com, xuying198702@163.com, 736412036@qq.com

How to cite this paper: Wang, R.H., Wang, C.L., Xu, Y. and Cui, X.M. (2016) The Research of Chinese Words Semantic Similarity Calculation with Multi-Information. *International Journal of Intelligence Science*, **6**, 17-28.
http://dx.doi.org/10.4236/ijis.2016.63003

Received: May 10, 2016
Accepted: July 22, 2016
Published: July 25, 2016

Abstract

Text similarity has a relatively wide range of applications in many fields, such as intelligent information retrieval, question answering system, text rechecking, machine translation, and so on. The text similarity computing based on the meaning has been used more widely in the similarity computing of the words and phrase. Using the knowledge structure of the <HowNet> and its method of knowledge description, taking into account the other factor and weight that influenced similarity, making full use of depth and density of the Concept-Sememe tree, an improved method of Chinese word similarity calculation based on semantic distance was provided in this paper. Finally the effectiveness of this method was verified by the simulation results.

Keywords

HowNet, Similarity, Chinese Words Similarity, Multi-Information

1. Introduction

With the rapid development of social information, the requirement coming from the needs that people deal with a lot of information by computer becomes more and more, especially in text information processing. Based on the data analysis, handling large amounts of textual information, conducting scientific research, business management and business decisions have become a hot topic in computer science. The process that uses computer to convert, transfer, store and analyze words and language belongs to category of natural language processing. It is a subject associated to linguistics, computer science, mathematics, information theory, and so on. The text similarity calculation is basic but important in the natural language processing. It has an important application in many fields, such as: information retrieval, question answering system, and text rechecking [1] [2].

Similarity is a complex concept. It has been discussed in semantics, philosophy and information theory. Now, there is no general method that defines similarity, because it involves language, statement structure and some other factors.

The Vector Space Model (VSM) presented by Salton is used more and it has good results [3]. The Latent Semantic Indexing (LSI) proposed by Deerwester uses Singular Value Decomposition (SVD) to convert word frequency matrix to singular matrix, calculates cosine's similarity by standard inner product and then compares the text similarity according to the calculate results [4]. There are many methods that use ontology to study text similarity. This method uses the existing ontology model such as the WordNet proposed by Christine (1998) to help deal with text processing [5]. The Concept Forest proposed by James Z. Wang and William Taylor (2007) is the method of calculating text similarity based on the ontology [6]. Bo Jin and Yanjun Shi (2005) have put forward the text similarity calculation based on the semantic comprehension, using the knowledge structure and the grammar of the language describing the knowledge of the <HowNet> [7]. Sun Shuang (2006) proposed the text clustering algorithm (TCUSS) based on the semantic similarity [8]. The TCUSS does clustering analysis based on the picture, avoiding the limit to the cluster shape caused by the algorithm. Ma Junhong proposed a text similarity calculation method, which is from the sentence, paragraph to the text stage. The calculation process is more complicated [9]. Tan Xueqing, Zhang Lei *et al*. designed a text classification algorithm based on clustering density [10]. Du Kun, Liu huailiang and Wang bangjin researched the Chinese text clustering by combining the weight of feature item in the text to construct semantic weighting factor of text similarity [11].

Word is the most basic semantic and grammatical unit of Chinese. The semantic similarity calculation of Chinese word is the basic of calculating sentences similarity. The similarity of word is a subjective concept. There is no definite objective standard that can measure. Because the relationship between words is complex, their similarity or difference is hard to measure with a simple value.

The method of calculating words semantic similarity can be divided into two types: one is dependent on the conceptual structure semantic dictionary method for hierarchical organization; the other is large-scale corpus-based statistical methods [1] [2] [12] [13].

The method based on the relationship of concepts semantic dictionary is mainly according to hyponymy of concepts structure relationship and synonymy to calculate.

The method based on the large-scale corpus statistics uses the words hyponymy information probability distribution to calculate words semantic similarity, such as Brown's method based on the average mutual information [12], and Lillian Lee's method based on the entropy [13]. This type of method is built on the basis when two words have some semantic similarity to some extent and when and only when they appeared in the same hyponymy. That is their context should be same when their meaning are close.

The quantitative methods based on the statistics can measure the semantic similarity

between words precisely and effectively. But, this method depends too much on the corpus train used, and it needs too much calculation and its methods are also too hard. Besides, it also can be influenced by the data sparse and data noise so that it can lead to absolute mistake. The similarity calculation mentioned in this article means how close the text expresses in their semantic distance.

2. The Words Similarity Calculation Research Based on <HowNet>

<HowNet> is a common commonsense knowledge database whose object described are the concepts that Chinese and English words expressed, and its basic content are the concepts, the relationship between concepts, the relationship between properties of concepts. There are two most basic concepts in <HowNet> called "Concepts" and "Sememe". The concepts are used to describe what the words mean, and, the sememe is the minimum unit of concepts.

<HowNet> take use of 1500 types of sememe and take them as the most basic unit todescribe concept. The relationship between sememe is also rather complex rather than mutual independence. There are eight relationships between sememe described in <HowNet>: hyponymy, synonymy, semanticrelation, antonyms, part-whole, property-host, material-product, event-role. The structure composed of the relationship between sememe if a complex mesh structure, not a simple tree structure. But the most important relationship in sememe relationship is hyponymy. Based on the hyponymy, all the basic sememe form a sememe level system (as Figure 1). This sememe level system is a tree structure, and also is the basis of our semantic similarity calculation.

In <HowNet>, it is not that every concept correspond to a node in the tree concept level system, but take use of a series of sememe, and use some knowledge describe language to describe a concept. The sememe form a tree sememe level structure taking use of hyponymy. It tries to find a method that can calculate the similarity between two semantic expressions.

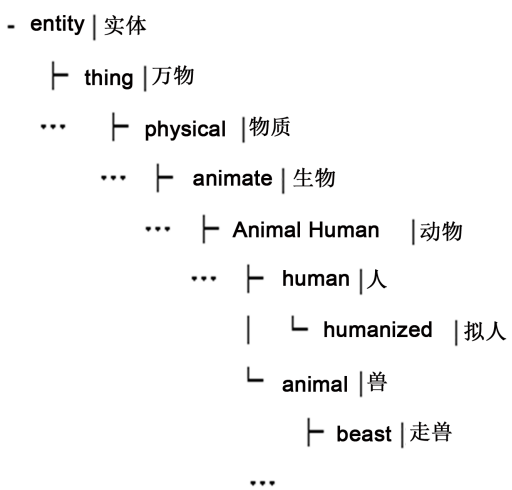

Figure 1. Schematic diagram of sememe tree structure.

A simple method taking use of <HowNet> to calculate semantic similarity use the first independent sememe belonging to word sememe expression, saying that word similarity equal to the first independent sememe similarity. The advantage of this method is that it calculate simple, but it doesn't use the other information in HowNet semantic expression.

For two Chinese words W_1 and W_2, assumed that W_1 has n synset (concept): $P_{11}, P_{12}, \cdots, P_{1n}$, W_2 has m synset (concept): $P_{21}, P_{22}, \cdots, P_{2m}$, here we set that the similarity between W_1 and W_2 is the maximum among the every synset's similarity, the formula follows:

$$Sim(W_1, W_2) = \max_{i=1,\cdots,n,\, j=1,\cdots,m} Sim(S_{1i}, S_{2j}) \tag{1}$$

Doing this, we can convert the problem that calculating two words' similarity into the problem that two concepts' similarity.

Since that every concept are presented in sememe (some specific in specific words), the sememe similarity calculating becomes the basic of calculating concepts' similarity.

In consideration of all sememe formed a tree sememe level system based on the hyponymy, we adopt the method using semantic distance to calculate the sememe similarity. Assumed that the path length between two sememe in the level system is d, then the distance of the two sememe is [14]:

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha}. \tag{2}$$

Among this, P_1, P_2 are two sememe, d is the distance of P_1, P_2 in the level system. It is a positive integer. α is an adjustable parameter.

In view of that specific words occupied small percentage in the semantic representation of <HowNet>, to simplify, we set the following roles:

A. The similarity between specific words and sememe is regarded as a small constant (γ).

B. The similarity between specific words and specific words is 1 if the two specific words are same, otherwise 0.

In actual text we think it like this that function word and notional word can't replace each other, so the similarity between function word concept and notional word is 0.

Function word concept always use "relation sememe" or "syntax sememe" to express, so function word concept's similarity just need to calculate its similarity of corresponding relation sememe or syntax sememe.

The basic thought we calculate notional word concept similarity is: integral similarity need to be built base on the portion similarity. Resolve a complex entirety into several or more parts, and calculate parts' similarity to get the integral similarity.

Assumed that two entireties A and B can be resolved into the following several parts: A is resolved into $A_1, A_2, \cdots, A_n$, B is resolved into $B_1, B_2, \cdots, B_m$, then there are $m \times n$ types of relationships between these parts. Every part in the entirety plays different role, and it makes sense only when compare the parts that have the same effect. So, when compare the two entireties similarity, the relationship between the two entireties' every

part should be build first, and then compare these parts.

The similarity between any sememe or specific words and null is defined a small constant (δ);

The entirety's similarity can be got by weighted averaging parts' similarity.

About notional word concept semantic express, we can divide it into four parts:

a) The first independent sememe description: we consider the two concepts' similarity about this part as $Sim_1(S_1, S_2)$.

b) The other independent sememe description: all the other independent sememe or specific word in the semantic express except the first independent sememe, we consider these two concepts' similarity about this part as $Sim_2(S_1, S_2)$.

c) Relation sememe express: all the relation sememe express in semantic express, we consider the two concepts' similarity about this part as $Sim_3(S_1, S_2)$.

d) Symbol sememe express: symbol sememe in the semantic express, we consider the two concepts' similarity about this part as $Sim_4(S_1, S_2)$.

So, the entirety's similarity about two concepts' semantic express can be written as:

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i Sim_i(S_1, S_2). \tag{3}$$

$\beta_i (1 \le i \le 4)$ is an adjustable parameter, and: $\sum_{i=1}^{4} \beta_i = 1$, $\beta_1 \ge \beta_2 \ge \beta_3 \ge \beta_4$. The latter shows that the effect that Sim1 to Sim 4 have on the entirety similarity calculation if descending. Because the first independent sememe description express is a leading characteristic of a concept, its weight should be defined larger, generally ≥0.5.

If Sim1 is very smaller and Sim 4 is very big, it will lead to the entirety similarity big. So, the Formula (3) can be modified:

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(S_1, S_2). \tag{4}$$

The meaning of Formula (4) consists in that the similarity of main parts can restrict the similarity of minor parts. On the other words, the similarity of main parts is lower and the effect that the similarity of the minor parts have on the entirety similarity will reduce.

3. The Similarity Calculation of the Words Fusing Multiple Information

The method that calculate words similarity based on the concept semantic distance usually use the same synonymy dictionary in which all the words is organized in one or a few tree level structure, using the distance of two nodes in the tree as a measure of semantic distance of the two concepts. WangBin [15] (1999) use this method and use the <Chinese thesaurus> to calculate the similarity of Chinese words, Agirre & Rigau [16] also take the depth of the concept level tree and area destine into consideration in addition to the distance of nodes in the words similarity calculation based on the WorldNet.

The thought that think the other factors is used in the words similarity based on the <HowNet>. All the sememe are divided into nine sememe tree on account of hyponymy and one tree represent a sememe class, each node in the tree only belonging to one sememe tree. So, the path of the two sememe in the same sememe tree exists one and only one and this path distance is the measure of the semantic distance between sememe: the semantic distance of the two sememe is not in the same sememe tree meanwhile is defined a larger constant. Using the calculate formula gave by Reference [14].

$$Sim(p_1, p_2) = \frac{\alpha}{\alpha + d} \tag{5}$$

It is only thought about the distance of two sememe by using the Formula (5) to calculate the sememe similarity, and the other information the sememe tree offered are not thought. On the condition that the two sememe is in the same distance, the semantic distance will be smaller if the depth of the sememe tree it belonging to is bigger. And, the similarity among sememe is also related to semantic coincidence degree, the bigger the coincidence degree, the bigger the similarity. Of course, the degree of refinement also influence the sememe similarity calculation.

So, the factors that affects the sememe similarity can be thought about from the aspects following: path distance, node depth, node density and so on.

a) Node distance. The distance of two sememe referred to the shortest distanceconnecting the two nodes. The similarity is lower when the distance of two sememe is bigger, on the contrary, it will be bigger when the distance of two sememe is smaller. A simple corresponding relationship can be built between the distance of sememe and similarity. Especially when the distance of two sememe I 0, their similarity is 1, and when the distance is infinite, their similarity is thought as a smaller constant.

b) The depth of the sememe tree. When the two sememe is in the situation that their path length is same, the semantic distance of sememe should be smaller if the depth of the sememe tree if belonging to.

c) Node depth. Node depth refers to the number of sides that the shortest path contains.As in the sememe level tree, each level is the detailing of the former level sememe. On that the distance of sememe is same, when the sum of the depth of two nodes is bigger, the sememe similarity is bigger; when the difference of the depth of two nodes if smaller, the sememe similarity is bigger.

d) Node density. Node density is defined the density of nodes belonging to the ancestorthat two nodes is nearest to. The density of different sememe in sememe level system is different, some sememe may own a sew nodes, and some may own hundreds of nodes. Generally, when the sememe distance is same and the son nodes density is bigger showing that detailing sememe concept is more specific, their similarity is smaller; otherwise bigger.

e) Adjustable parameter. Semantic similarity is a strong subjective concept, the different the sememe, the different the similarity.

When the paper talk about how the depth and density of sememe tree influence the

similarity, only the situation that two sememe is in the same tree is considered. Because the depth of two sememe in the different tree is different and don't have common nodes. For this case, the Formula (5) will be adopted, α, d will take the default value [9].

The Formula (5) calculate the similarity only base on the distance of sememe, and base on the formula, the depth of the sememe tree should be considered first, on the condition that the distance is same, the bigger the depth of the sememe tree the two sememe belonging to, the bigger the similarity. So, this parameter α can be replaced by h the depth of the sememe tree and this can show the influence that the depth of sememe tree have on the result. From this, Formula (5) can be expressed like this:

$$Sim_1(p_1, p_2) = \frac{h}{h+d} \tag{6}$$

Here, h is the depth of the sememe tree, d is the path distance of the two sememe.

The effect that the depth the sememe in sememe tree have on similarity is called level coefficient. Here introduce the Least Common Node (LCN), showing that the node nearest to common node that the two sememe in sememe level tree. L is the level it in the level tree. In order to show the effect that the LCN have on similarity, the parameter l/h is introduced, h is the depth of the sememe tree, level parameter being used to realized it that the higher the depth of sememe, the larger the similarity.

Through the analysis above, the Formula (5) is translated into:

$$Sim_1(p_1, p_2) = \frac{\alpha}{\alpha+d} = \frac{h}{h+d} \times \frac{l}{h} = \frac{l}{h+d} \tag{7}$$

In the formula, l is the level that the LCN in.

Then the effect that the density of LCN have on sememe similarity will be discussed. The density of the sememe node is defined $\text{density}(\text{LCN}) = \frac{n}{m}$, the density (LCN) represents the density of the ancestor node LCN of two sememe A, B. The number of common ancestor node of sememe A, B is represented in n, m represents the whole number of nodes of the sememe tree the sememe belonging to. So, the formula that density influence similarity is:

$$Sim_2(p_1, p_2) = \frac{1}{\text{density}(\text{LCN})}.$$

Absolutely, $0 < \text{density}(\text{LCN}) < 1$, $\frac{1}{\text{density}(\text{LCN})}$ must larger than 1, which violates similarity values [0,1], so $Sim_2(p_1, p_2) = \frac{1}{\text{density}(\text{LCN})} = \frac{m}{n}$ is modified as follows:

$$Sim_2(p_1, p_2) = \frac{m}{n+k} \tag{8}$$

In the formula, k is an adjustable parameter, $k > m - n$, k values 2 m in the paper, it means that when the number of son nodes of LCN is 0, $Sim(p_1, p_2)$ values 0.5.

All the above taken into consideration, combining Formula (7) and Formula (8), fusing the depth and density of level tree, we give the formula as follow:

$$Sim(p_1, p_2) = Sim_1(p_1, p_2) \times A + Sim_2(p_1, p_2) \times B \tag{9}$$

Among these, $Sim_1(p_1, p_2)$ is the similarity that sememe distance and depth calculate, $Sim_2(p_1, p_2)$ is the similarity through the density. A is the degree that the $Sim_1(p_1, p_2)$ influence the similarity calculation, means weight, B is the degree that the $Sim_1(p_1, p_2)$ influence the similarity, $A + B = 1$.

There are many factors influencing literal similarity, such as path distance, depth of a node, the node density, as well as an adjustable parameter. The algorithm given here is consideration of these aspects and, thus it than any single factor calculation method is more adaptable.

In the formula calculating the sememe similarity, the two weight A, B means the importance the two formulas is in the whole similarity calculation, and the size of the weight will influence the decision result directly. The method of deciding weight can adopt the following methods: experts estimate, frequency statistics, fuzzy comprehensive evaluation and so on. In this paper, the weight A was set 0.75, and was set 0.25.

4. Conclusions

The method of word similarity calculation above is adopted in sentence similarity calculation.

The method based on the semantic dictionary <HowNet> is applied in this paper. First, the semantic similarities between words are calculated, and then get the sentences semantic similarity based on the words similarity.

Assume that two sentences A and B; after preprocessing, the words that A contains are $A_1, A_2, \cdots, A_M$; the words that B contains are $B_1, B_2, \cdots, B_N$; the similarity between $A_i\ (1 \le i \le m)$ and $B_j\ (1 \le j \le n)$ is $S(A_i, B_j)$; then the array can be got:

$$M(A, B) = \begin{bmatrix} S(A_1, B_1), S(A_1, B_2), \cdots, S(A_1, B_n) \\ S(A_2, B_1), S(A_2, B_2), \cdots, S(A_2, B_n) \\ \vdots \\ S(A_m, B_1), S(A_m, B_2), \cdots, S(A_m, B_n) \end{bmatrix}.$$

Through this array, the semantic similarity $S(A, B)$ between sentences A and B can be calculated:

$$S(A, B) = \frac{\sum_{i=1}^{m} \max\left(S(A_i, B_1), S(A_i, B_2), \cdots, S(A_i, B_n)\right)}{m}. \tag{10}$$

$S(A_i, B_j)$ is the semantic similarity between the No. i word in sentence A and the No. j word in sentence B.

The semantic information of the words is thought in the calculating sentences similarity. It is better than the method based on the vector free model when processing less same words and the two sentences have close meaning, but it is restricted by the quality of the semantic dictionary. It was adopted at the subjective topic decided in this paper. When deciding the similarity between the answer the examinee provided and the stan-

dard answer, it just talks about the similarity of semantic and it won't think about syntactic structure too much.

The calculation of semantic similarity between the examinee's answer and the standard answer should give full consideration to the influence of their comparison order. The right sequence should be: calculating the first sentence of the standard answer and the examinee's answer, getting a group of similarity, and then finding a maximum being the similarity of the first sentence, and deleting the sentence from the examinee's answer; then, calculating the semantic similarity between the second sentence of the standard answer and the examinee's answer, getting the maximum being the similarity of the second sentence, and then deleting the sentence. The rest can be done in the same manner: calculating the similarity between each sentence of standard answer and examinee's answer; adding the similarity of each sentence of standard answer and then dividing the number of the sentences of standard answer. And the similarity between examinee's answer and standard answer can be got.

Experiment A:

a) Standard answer: Memory is the storage device of computer system, using for store program and data.

After dividing the words: Memory/is/the storage/device/of computer/system, using for/store/program/and data.

b) Examinee 1 answer: Memory is used to store data.

After dividing the words: Memory/is/used/to store/data.

c) Examinee 2 answer: Memory is used to store program and data.

After dividing the words: Memory/is/used/to store/program/and data.

Figure 2 and Figure 3 show the grade calculation of different examinees about the same topic. Here think that the grade of each sentence is same, giving the total points of the topic. The point of this topic is the similarity calculated and the grade of total

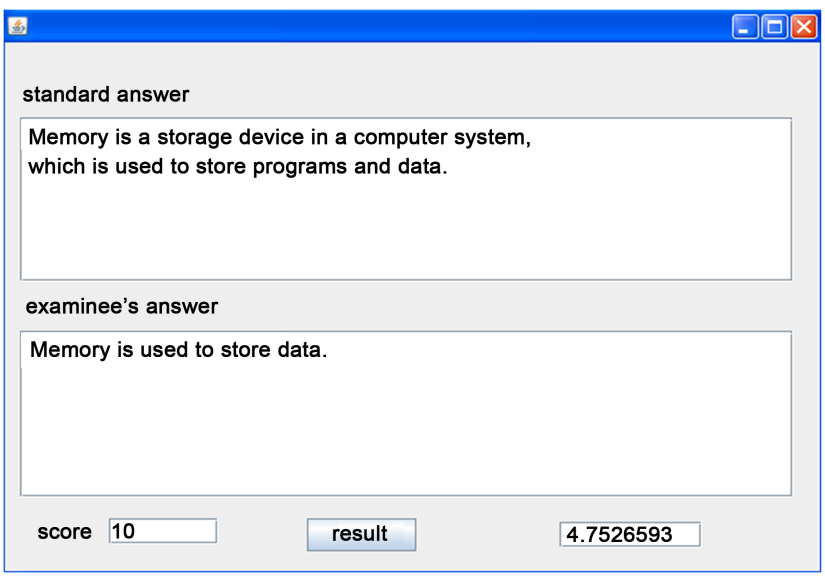

Figure 2. The grade of examinee 1 calculation.

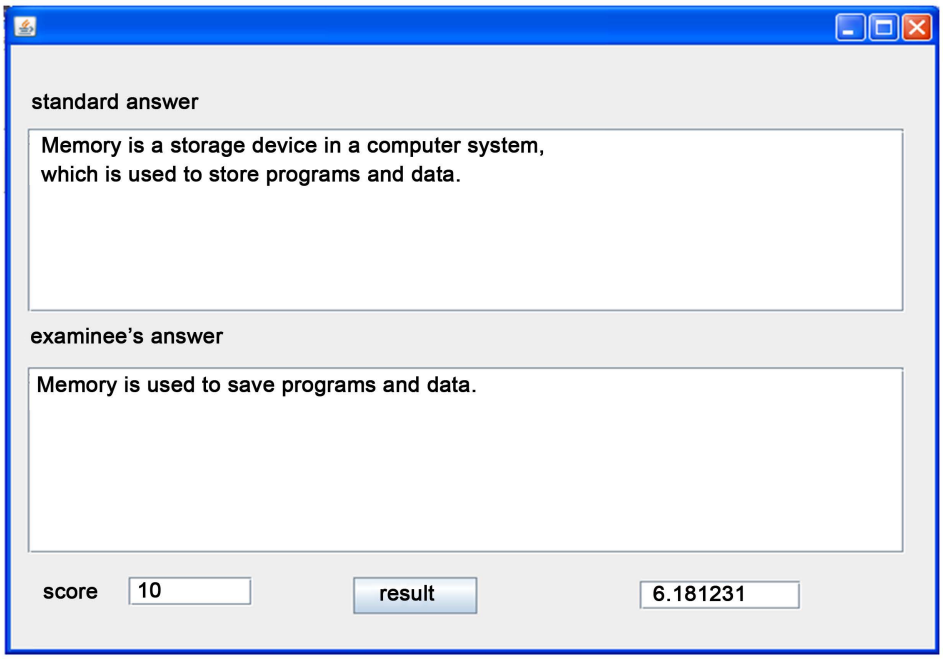

Figure 3. The grade of examinee 2 calculation.

points. Account of subjective judgment, the answer of examinee 2 is fuller, so the grade is higher than examinee 1. In addition, the experiment in this paper shows the precise grade. The value can be processed according to the physical truth.

Experiment B: selecting the test questions from the software engineering test paper, using the method of independent reading to get a grade, and comparing with the practical score of the examinee.

Topic: How many stages and times can software life be divided into? What is the target of each stage?

The answer: Three times: software definition, software development, operation maintenance. The three times can be divided into 8 stages: problem definition, feasibility research, requirement analysis, overall design, detailed design, coding unit test, comprehensive test and software maintenance.

20 test papers were selected in the experiment. The scores of automatic correcting and teacher correcting are shown as **Table 1**. In order to show the differences between the automatic correcting and the teacher correcting, the experiment keeps a valid decimal number calculating the grade and when practical use can be rounded or rounded numbers.

The error value is the teacher correcting minus automatic correcting.

In order to observe the differences between automatic correcting and teacher correcting, the calculation result is shown in line chart, as shown in **Figure 4**.

The line chart shows that the result of the automatic correcting is consistent with teacher correcting, and the result is also close. The effectiveness of this method is verified by the experiment results.

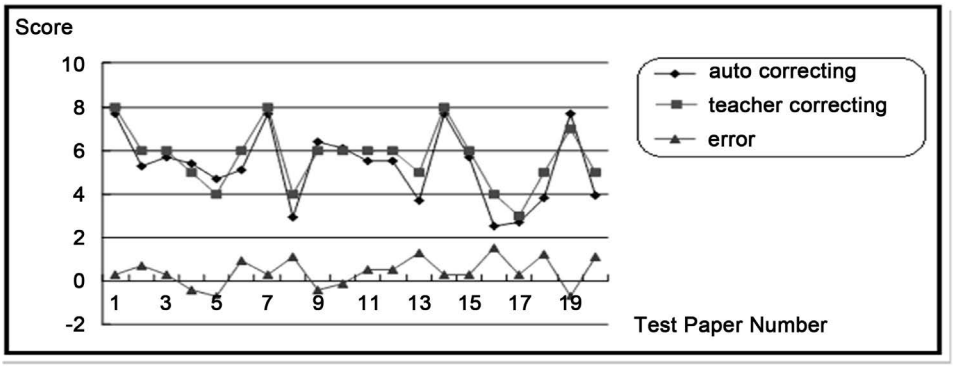

Figure 4. The comparison of item between automatic correcting and practical score.

Table 1. The comparison table between automatic correcting and teacher correcting.

Item (8 points)			
Automatic correcting	Teacher correcting	Error	Pencentage
7.7	8	0.3	96.25%
5.3	6	0.7	91.25%
5.7	6	0.3	96.25%
5.4	5	−0.4	95.00%
4.7	4	−0.7	91.25%
5.1	6	0.9	88.75%
7.7	8	0.3	96.25%
2.9	4	1.1	86.25%
6.4	6	−0.4	95.00%
6.1	6	−0.1	98.75%
5.5	6	0.5	93.75%
5.5	6	0.5	93.75%
3.7	5	1.3	83.75%
7.7	8	0.3	96.25%
5.7	6	0.3	96.25%
2.5	4	1.5	81.25%
2.7	3	0.3	96.25%
3.8	5	1.2	85.00%
7.7	7	−0.7	91.25%
3.9	5	1.1	73.75%

Acknowledgements

The authors wish to thank Dr. Jingsheng Zhao and Dr. Wei Zhou.

References

[1] Banea, C., Hassan, S., Mohler, M., *et al*. (2012) A Superivsed Synergistic Approach to Semantic Text Similairity. *Proceedings of the* 1*st Joint Conference on Lexical and Computational Semantics*, 635-642.

[2] Glinos, D. (2012) Chunk-Based Determination of Semantic Text Similarity. *Proceedings of the* 1*st Joint Conference on Lexical and Computational Semantics*, 547-551.

[3] Salton, G., Wong, A. and Yang, C.S. (1975) A Vector Space Model for Automatic Indexing. *Communications of the ACM*, **18**, 613-620. https:/doi.org/10.1145/361219.361220

[4] Deerwester, S.C., Dumais, S.T., Landauer, T.K., *et al*. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, **41**, 391-407. https:/doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[5] Fellbaum, C. (1998) WordNet. An Electronic Lexical Database. MIT Press, MA.

[6] Wang, J.Z. and Taylor, W. (2007) Concept Forest: A New Ontology-Assisted Text Document Similarity Measurement Method. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. https:/doi.org/10.1109/WI.2007.11

[7] Jin, B., Shi, Y.J. and Teng, H.F. (2005) Similarit Algorithm of Text Based on Semantic Understanding. *Journal of Dalian University of Technology*, **2**, 139-145.

[8] Sun, S. (2006) A New Method for Text Clustering Based on Semantic Similarity. Nanjing University of Aeronautics and Astronautics.

[9] Ma, J.H. (2013) A Staged and Integrated Semantic Similarity Algorithm of Text. *New Technology of Library and Information Service*, **29**, 20-26.

[10] Tan, X.Q., Zhang, L., Zhou, T. and Luo, L. (2016) A Text Classification Algorithm Based on the Desity Clustering. *Research on Library Science*, **13**, 74-83.

[11] Du, K., Liu, H.L. and Wang, B.J. (2016) Research on Chinese Text Clustering Method Based on Semantic Relevacy. *Information Studies*: *Theory & Application*, **39**, 129-133.

[12] Brown, P.F., Pietra, S.A.D., Pietra, V.J.D. and Mercer, R.L. (2002) Word-Sense Disambiguation Using Statistical Methods. http://www.researchgate.net/publication/2472878_Word-Sense_Disambiguation_Using_Statistical_Methods

[13] Lee, L. (1997) Similarity-Based Approaches to Natural Language Processing. Ph.D. Thesis, Harvard University, Cambridge.

[14] Liu, Q. and Li, S.J. (2002) Word Similarity Computing Based on How-Net. *Computational Linguistics and Chinese Language Processing*, **7**, 59-76.

[15] Wang, B. (1999) The Research of Chinese-English Bilingual Corpora Alignment. Ph.D. Thesis, Institute of Computing Technology, Chinese Academy of Science.

[16] Agirre, E. and Rigau, G. (1995) A Proposal for Word Sense Disambiguation Using Conceptual Distance. *International Conference* "*Recent Advances in Natural Language Processing*", Tzigov Chark, Bulgaria.