A conceptual similarity and correlation discrimination method based on HowNet

Yunnian Ding[1*], *Yangli* Jia[1], and *Zhenling* Zhang[1]

[1]School of Computer Science and Technology, Liaocheng University, Liaocheng, China

Keywords: Similarity, Relevance, Natural language processing, HowNet.

Abstract. The similarity and correlation analysis of word concepts has a wide range of applications in natural language processing, and has important research significance in information retrieval, text classification, data mining, and other application fields. This paper analyzes and summarizes the information of sememes relationship through the definition of words in HowNet and proposes a method to distinguish the similarity and correlation of words. Firstly, using a combination of the part of speech and sememes to distinguish the similarity and correlation between words concept. Secondly, the similarity and correlation calculation results between vocabulary concepts are used to further optimize the judgment results. Finally, the similarity and correlation distinction and discrimination between vocabulary concepts are realized. The experimental results show that the method reduces the complexity of the algorithm and greatly improves the work efficiency. The semantic similarity and correlation judgment results are more in line with the human intuitive experience and improve the accuracy of computer understanding of natural language. which provides an important theoretical basis for the development of natural language.

1 Introduction

Natural Language Processing (NLP) has always been a research hotspot in the field of information retrieval and artificial intelligence. How to make computers like human beings to think independently and understand natural language is one of the more difficult problems at present. The computation of semantic similarity and correlation has important research significance for computer understanding of natural language. Most researchers' current research work is mainly divided into two categories: corpus-based statistical methods [1], knowledge base-based methods [2] and a combination of the two methods, which have achieved good results.

Zhao Yingqiu et al [3] summed up the eight implicit semantic relations between concepts based on the relationship between sememes in HowNet to calculate the semantic relevance of words and introduced similarity to calculate the correlation between words. The calculation results are more in line with human intuitive feelings; Wang Yi [4] and

* Corresponding author: dingyunnian3298@163.com

others made 12 semantic relationships between explaining sememes as the horizontal relationship between sememes to calculate the relevance of the sememes, greatly improving the calculation accuracy of semantic relevance between words.

Although they have done a lot of research on the similarity and correlation between words and achieve good research results, they did not distinguish the similarities and correlations between words. In natural language processing, the similarity between vocabulary refers to the substitutability of two words in a certain context. The correlation between words means that in syntactic analysis, two words in a phrase structure can form the possibility of modified relationship, subject-predicate relations, and the same-point relationship. Similarity and correlation are two concepts that are both interconnected and distinct. Similarity is a special correlation, and relevance implies similar. Therefore, when calculating the semantic similarity of words, the two concepts are very easy to be confused [5]. The vocabulary with high similarity calculated by current research work may have a high degree of relevance, such as "doctor-nurse", but the highly relevant words are not necessarily similar, for example: "sun-solar". Therefore, the distinction between semantic similarity and correlation has become a problem to be solved in the current academic world.

This paper analyzes and summarizes the existing methods of calculating similarity and relevance between words and finds that the first basic sememe in HowNet reflects the most important features of vocabulary and has an important role for the distinction between similarity and correlation. This paper introduces the part of speech and the weak sememe to judge and distinguish the correlation and similarity of the words, which greatly reduces the complexity of the algorithm and improves the work efficiency. At the same time, this paper distinguishing the similarity and correlation by comparing sememes, then further optimize the results for words without same sememe using the similarity and correlation between vocabulary concepts. Finally, whether the similarities and correlations between words are determined. But this does not mean that similar words are irrelevant or related words are not similar, but that the distinction results more tend to be similar or related, and even some words have two properties at the same time.

2 Introduction to HowNet

HowNet [9] is a commonsense knowledge that uses the concepts represented by Chinese and English words as description objects to reveal the relationship both concepts and the attributes of concepts. its main purpose is to reflect the commonality and individuality between concepts, proposed and created by the famous machine translator Dong Zhendong.

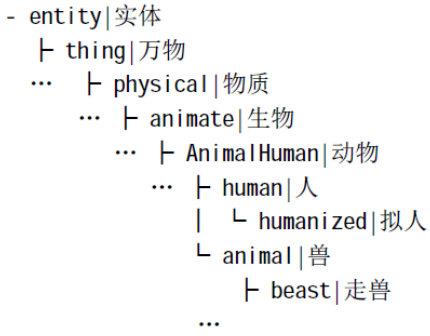

Fig. 1. Sememes tree hierarchy.

In HowNet, a word has one or more concepts, each of which is represented by a record. each word is composed of one or more homonyms. Each of the homonyms is composed of one or more sememes, which is the smallest semantic unit of words. There are 2196

sememe in HowNet. All the words are represented by these sememes, and each sememe has its interpretation. There are 10 types in sememe, each type can form a separate sememes tree hierarchy, as shown in Figure 1. Through the vertical semantic relationship in the sememes hierarchy, the semantic similarity between sememes can be calculated.

HowNet is a semantic system, not just a semantic dictionary. It focuses on the commonalities and individualities between concepts, such as: "teacher" and "student", "people" is their commonality, "teacher" is the executor of "teaching", and "student" is the victim of "teaching", which reflects the individuality of the both. HowNet not only reflects that the conceptual information of words is composed of sememes, but also there are various complex relationships between sememes. The hierarchical structure of sememes reflects the upper and lower relationship of sememes, which is called vertical relationship. In addition to the upper and lower relationship, synonymous relationship, antisense relationship and derogatory relationship, there are 12 horizontal semantic relations between sememes. The calculation of the relationship between sememes can be performed by 12 horizontal semantic relations between explaining sememes.

3 Related work

Words are described by sememes. sememe is the smallest unit that describes concept information. Therefore, the similarity and correlation between word concepts are determined by similarity and relevance between sememes, so the calculation of similarity and relevance between sememes are the most basic and important tasks in calculating word similarity and relevance.

3.1 The correlation calculation between sememes

In HowNet, the correlation between sememes can be calculated through the horizontal semantic relationship of sememes. There are 12 kinds of horizontal semantic relations in sememes, which are represented by the following symbols, namely { *, @, ？, ！, ~, $, ￥, %, ^, &, +, null }, giving them different weights according to experience, namely {0.7,0.6,0.7,0.4,0.75,0.9,0.7,0.5,-1,0.8, 0.9, 0.5} [6].

Li Shengqi [6] and others used the relationship between sememes to calculate the correlation degree of sememes, as shown in formula (1).

$$Rele(p_1, p_2) = \sum_{p_i \in exp(p_1)} w_i * \frac{Sim(p_i, p_2)}{n} + \sum_{p_j \in exp(p_2)} w_j * \frac{Sim(p_1, p_j)}{m} \quad (1)$$

where $p_i \in \exp(p_1)$ and $p_j \in \exp(p_2)$ represent the interpreted sememes of the sememes p_1 and p_2, $Sim(p_i, p_2)$ and $Sim(p_1, p_j)$ represent the similarity between p_2 and p_i interpreted sememes in p_1 and the similarity between $p_1$and p_j interpreted sememes in p_2 respectively. w_i and w_j represent the weights corresponding to each of them, and m and n represent the number of p_1 and p_2 explanatory sememes respectively.

Wang Yi et al [4] improved it and used the relationship between sememes to calculate the degree of correlation between sememes, as shown in formula (4).

$$Rele(p_1, p_2) = max_{p_i \in exp(p_1), p_j \in exp(p_2)}(max(w_h) * Sim(p_i * p_j) \quad (2)$$

In which, $1 \le i \le m, 1 \le j \le n, 1 \le h \le 12$, $\mathrm{Rele}(p_1, p_2)$ represents the degree of association between the sememes p_1 and p_2, and serves as the correlation between the two vocabulary concepts. $\exp(p_1)$ and $\exp(p_2)$ represent two sets of interpreted sememes of p_1, p_2. $Sim(p_i * p_j)$ denotes the similarity between p_1, p_2 interpreted

sememes. h denotes the number of relations between the interpreted sememes, and w_h denotes the weight of the h_{th} relationship between the sememes.

3.2 The similarity and correlation calculation between concepts

The correlation between concepts is determined by the similarity and relevance. The greater the similarity between the two concepts, the greater their correlation. Similarly, the greater the relevance between concepts, the larger the correlation. But the greater the relevance of the two concepts, their similarity is not necessarily large. Therefore, if the concepts of the two words are similar, they are likely to be relevant as well.

Literature [4] introduces the similarity between concepts and the case influence factor to calculate the concept correlation. The concept correlation calculation formula is as follows:

$$Associ(C_1, C_2) = \beta_1 * sim(C_1, C_2) + \beta_2 * Rele(C_1, C_2) + \beta_3 * Effe(C_1, C_2) \quad (3)$$

$\mathrm{Associ}(C_1, C_2)$ represents the correlation between two lexical concepts; $sim(C_1, C_2)$ represents the similarity between two lexical concepts, which can be calculated according to the literature [2]. $Rele(C_1, C_2)$ represents the degree of association between two concepts. $Effe(C_1, C_2)$ represents the relevance of the instant impact factors between concepts. Among them, $\beta_1, \beta_2, \beta_3$ represent their respective weights, $\beta_1 + \beta_2 + \beta_3 = 1$.

4 An algorithm distinguishing similarity and correlation between concepts

The definition and description information of the vocabulary concept is represented by sememe information. Through the analysis of sememe data, this paper finds that many sememes in HowNet have very weak ability to distinguish between similarity and correlation, such as the weak sememes “entity, material, things, positions”. This semantic information has no real meaning. If these words are applied to the calculation of conceptual similarity and relevance, it will reduce the accuracy of the calculation results. Therefore, this paper considers the use of these weak sememes to distinguish the similarity and relevance of words, which will interfere with the judgment of its similarity and correlation. Even if two weak sememes appear in two unrelated words at the same time, two unrelated words will be related or similar, so when distinguishing the lexical similarity and correlation, these weak sememes need to be removed.

This paper thinks that part-of-speech information has a direct influence on the judgment of the similarity and correlation of words. Two similar words must have the same part of speech. However, the related words can be the same or different. That also is, two words with different parts of speech must not be similar, such as "doctor" and "cure." Through this idea, this paper can reduce the computational complexity of the algorithm by considering the part of speech information.

The literature [7] believes that the first basic sememe reflects the most important characteristics of the concept. Through this idea, this paper uses the first basic sememe as an important factor for the similarity determination. The two concepts that are identical to the first basic sememe are similar. If the two sememes are the same and both are the first basic sememe, this paper considers that the concepts of the two words are similar if the two sememes are the same but not at the same time as the first sememe, this article considers two concepts to be related.

However, not all concepts have the same sememe. For the two concepts with different sememes, we need to calculate the similarity and relevance of the vocabulary concept. First, the similarity of sememes is determined by the vertical relationship of sememes, then

calculating the similarity between concepts. When the similarity between the two concepts is very large, this paper considers the two concepts to be similar.

Based on the above ideas, this paper analyzes and summarizes the definition of word concept, combining the concept similarity with relevance calculation method to summarize the similarity and correlation of concepts, and draws the flow shown in Figure 2.

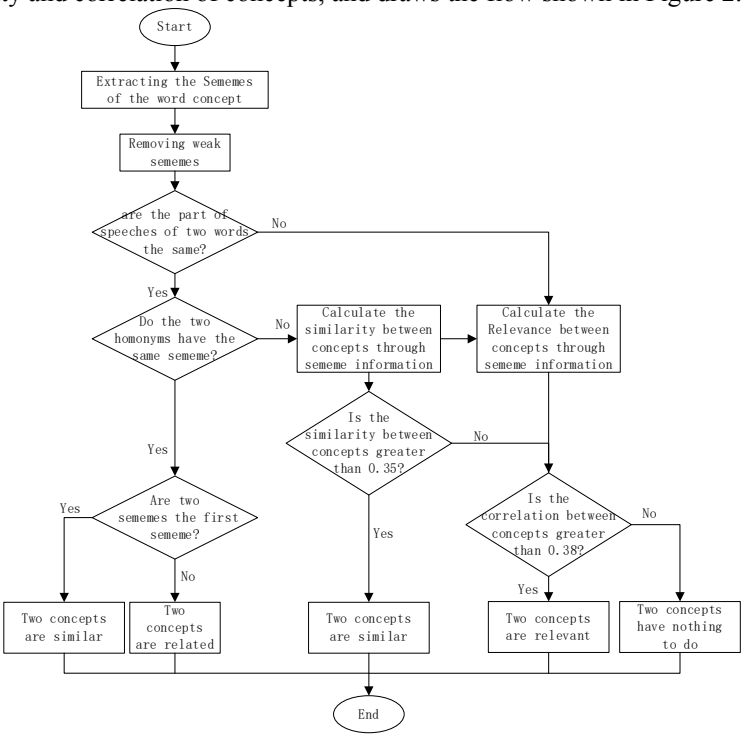

Fig. 2. Discrimination flow chart for word similarity and correlation.

According to the discriminant idea of conceptual similarity and correlation above, this paper designs the following algorithm:

Input: two words w_1 and w_2

Output: Whether the two words have similarities and correlations

method:

1) Extract the semenes information $p_{11}, p_{12}, \dots, p_{1m}$ and $p_{21}, p_{22}, \dots, p_{2n}$ represent the concepts C_1 and C_2 of the two words w_1 and w_2 in HowNet, and remove the weak sememes to obtain $p_{11}, p_{12}, \dots, p_{1i}$ and $p_{21}, p_{22}, \dots, p_{2j}$, where $1 \le i \le m, 1 \le j \le n$.

2) Extract the two concepts of part-of-speech information P_1 and P_2 from the information representations of the two words $C_1$and C_2 in HowNet to determine whether P_1 and P_2 are the same.

3) If P_1 and P_2 are different, directly calculate the conceptual relevance of the two words $\mathrm{Associ}(C_1, C_2)$ to determine whether the two words are related.

4) If P_1 and P_2 are the same, it is necessary to judge whether the two homonyms have the same sememe, that is whether p_x, p_y are the same, where $1 \le i \le m, 1 \le j \le n$. If $p_x = p_y$, then need to judge whether the two concepts are the first basic sememe? If so, the two words are similar, otherwise, the two words are related. If $p_x \ne p_y$, the calculation of the concept similarity and relevance of the words is needed, and the

similarity and correlation of the words are distinguished and judged by the calculation results. The distinction between word similarity and relevance is shown in equation (4).

$$two\ words\ are \begin{cases} similar, (pos = 1, p_1 = p_2\ or\ C_1 \sim C_2) \\ related, (pos \neq 1, p_1 = p_2\ or\ C_1 \propto C_2) \\ unrelated, \quad other \end{cases} \tag{4}$$

where P_1 and P_2 are the two sememes of the concept, $pos = 1$ means that both of the sememes are the first basic meaning, $p_1 = p_2$ means that the two sememes are the same, and C_1 and C_2 represent the concepts corresponding to the two words. $C_1 \sim C_2$ indicate that the two concepts are largely similar, and $C_1 \propto C_2$ indicates that the two concepts are largely related.

5 Experimental results and analysis

In this paper, we first use formula (3) to judge the similarity and correlation between two words. If there is no identical sememe in the conceptual description of the vocabulary, then we need to use the values of similarity and correlation of the vocabulary concepts. In the calculation of concept similarity and relevance, this paper uses the formula (2) to calculate the degree of sememes relevance. The similarity calculation between concepts uses the method in [8], combined with formula (3) to calculate the correlation between word concepts. In this paper, the similarity and correlation of a large of word pairs are marked, and the data of similarity and relevance are analyzed. Finally, 0.35 and 0.38 are determined as the similar or related boundaries between the judged words. In this paper, some typical words are used as experimental data, and the experimental results are shown in Table 1.

Table 1. Experimental results of similarity and correlation between concepts.

word1	word2	WordSimilarity-353（human）	similarity	relevance	similar	related	unrelated
teacher	school	—	—	—		√	
doctor	nurse	0.700	—	—	√	√	
eat	bread	—	—	0.557		√	
eat	paper	—	0.083	0.115			√
bread	chocolate	—	—	—	√		
bread	butter	0.619	—	—		√	
bread	paper	—	0.247	0.295			√
movie	star	0.738	—	—		√	
law	lawyer	0.838	—	—		√	
football	tennis	0.663	—	—		√	
cut	knife	—	—	0.559		√	
cut	apple	—	—	0.180			√
cucumber	potato	0.592	—	—		√	
professor	doctor	0.662	—	—	√	√	

To prove the validity of the experimental results, WordSimlarity-353 was selected as the evaluation set. WordSimlarity-353 was manually evaluated by 13~16 experts on the relevance of 353 English word pairs, Finally, the average value was used as the relevance of word pairs. This article uses the English evaluation results as its corresponding Chinese evaluation data and converts the correlation of the manual evaluation into the value between [0, 1], making the experimental results look more intuitive.

From the above experimental results, we can see:

The use of part of speech and the first basic sememe as a criterion for the distinction between similarity and correlation applies to most words, and the similarity and correlation can be distinguished without calculating the similarity and relevance between words. For example, doctors - hospitals, car - cars, milk - cows, sun - solar energy, etc.

(1) Taking “teacher” and “school” as examples, in the field of teaching, teachers and schools are closely related. “Teachers are teaching in schools” is the most basic common sense; their definitions in HowNet are as follows:

"Teacher": {human: HostOf = {Occupation}, domain = {education}, {teach: agent = {~}}}

"School": {InstitutePlace: domain = {education}, {study: location = {~}}, {teach: location = {~}}}

Firstly, in the information representation in HowNet, the part of speech of two concepts are nouns, so the two words may be similar or related. However, as can be seen from the definition of the two concepts above, there is a weak sememe in the "teacher", so we need to remove it. When we compare the sememes, concept similarity and correlation calculation, we don't need to consider the weak sememes, which greatly improves the efficiency of the algorithm. Meanwhile, the first basic sememe of the "teacher” is " human" and the first basic sememe of the school is " InstitutePlace ". The first basic sememe is different, so the two concepts do not have similarities. But there are two same sememes "education" and "teaching" in other sememes, so the two words are related. Experiments show that this method can distinguish concept similarity and relevance well, and it is more in line with human intuitive feeling.

(2) Taking "doctor" and "nurse" as examples, we can say that the two concepts are similar because "doctor" and "nurse" are both people, this is their commonality; of course, we can also say that they are related because they are all members of the medical field. In the definition of two vocabulary concepts in HowNet, , their first basic sememe is "human", so the two concepts are similar, but other sememes also have the same sememe information "medical", Therefore, the two concepts are also related; in WordSimilarity-353, the correlation of the two words is 0.700, which also verifies the validity of the experimental results.

(3) Taking “eat” and “bread” as examples, “bread” is used to “eat”. There is no doubt that the concepts of these two words are very relevant. However, in the information representation of two concepts in HowNet, their word-of-speech is different, so the two words do not have similarities. Only need to calculate their relevance, the correlation of the two words is 0.557 by calculation. Their relevance is very high, so the two concepts are related, but the "bread" and "newspaper" in the sememes description do not have the same sememe and the semantic similarity and correlation between the concepts are very low, so two words are neither related nor similar. Experiments show that the calculation of conceptual similarity and correlation is used to solve the situation of different sememes information, which makes the judgment result more rational.

6 Conclusions

The method of distinguishing lexical similarity and relevance using the methods of part-of-speech, de-weakness and sememe comparison is applies most words. And there is no need to calculate the similarity and correlation for each group of words, just calculate some special words, which can reduce the time complexity of the operation and improve the efficiency of the algorithm. The similarity and correlation of vocabulary concepts are used to further optimize the discrimination of similarity and correlation, which makes the experimental results more reasonable.

So far, the distinction between similarity and relevance has no clear boundary. Their classification criteria depend to a large extent on an application area or an article, such as the frequency of co-occurrence of two words in an article is very high, the correlation between them is very large, which requires a method based on corpus statistics. If the statistical-based method is introduced into the distinction between similarity and relevance, the accuracy of similarity and relevance distinction will be greatly improved. Therefore, the introduction of corpus into the method of lexical similarity and correlation distinction is currently a problem to be solved.

References

1. Zhi-Chao S, Xiao-Peng T. Semantic Similarity Computing Method Based on Wikipedia[J]. Computer Engineering, 2011, 37(7):193-195.
2. Fan M, Zhang Y, Li J. Word similarity computation based on HowNet[C]// International Conference on Fuzzy Systems & Knowledge Discovery. IEEE, 2016.
3. Zhao Yingqiu, Luo Jun,Zhang Junyan. Word semantic relevancy computation based on HowNet[J]. Information technology. 2010(3):90-93.
4. Wang Yi, Wang Xiaolin. Algorithm for Words' Semantic Relevancy Based on Modeified Algorithm for Sememes' Relevancy[J]. Journal of the China Society for Scientific Information. 2012, 31(12): 1271 -1275.
5. Wang Ruiqin,Yang Xiaoming, Lou Jungang. Research of Lexical Relatedness Measurement[J]. Journal of the China Society for Scientific Information. 2016, 35(4):389-404.
6. Li Shengqi, Tian Qianyan, Tang cheng. Disambiguating Method for Computing Relevancy Based on HowNet Semantic Knowledge[J]. Journal of the China Society for Scientific Information. 2009, 28(5):706-711.
7. Qun LIU, Sujian LI. Word Similarity Computing Based on HowN-net[J]. Computational Linguistics and Chinese Language Processing, 2002.
8. Wang Xiaolin, Wang Yi. Improved word similarity algorithm based on HowNet[J]. Journal of Computer Applications, 2011, 31(11):3075-3090.
9. Chen Zhiqun, Gao Fei, Zeng Zhijun. Word Relatedness Measure Based on Chinese Wikipedia[J]. Journal of the China Society for Scientific Information, 2012, 31(12):1265-1270.
10. Taieb M A H, Aouicha M B, Hamadou A B. Computing semantic relatedness using Wikipedia features[J]. Knowledge-Based Systems, 2013, 50(50):260-278.
11. Dong Zhendong, Dong Qiang. Introduction to HowNet[2007-04-03].http://www.keenage.com.
12. Shi Junbing. Calculating Method of Word Similarity Based on the HowNet[J]. Journal of Taiyuan University (Natural Science), 2017(1):69-72.
13. Zeng Shuqin, Wu Yangyang. The model of words relation computing based on the HowNett[J]. Microcomputer & Applications, 2012, 31(8):77-80.
14. LIN Li, XUE Fang, REN Zhongsheng. Modified word similarity computation approach based on HowNet [J]. Journal of Computer Applications, 2009(1):217-220.
15. Li Sujian. Research of Relevancy between Sentences Based on Semantic Computation[J]. Computer Engineering and Applications, 2002, 38(7): 75-76.