An improved semantic similarity algorithm based on HowNet and CiLin

Ying Wang[1], *Xiwei* Feng[1,*], *Yue* Zhang[1], *Haiming* Chen[1], and *Lijie* Xing[1]

[1]Liaoning Shihua University, Liaoning Fushun, China

Keywords: Semantic similarity, HowNet, CiLin.

Abstract. This paper explores an improved method for the semantic similarity calculation of words combined with HowNet and CiLin. Firstly, we designing the algorithm based on HowNet's sememe similarity improvement calculation, comprehensively considering the influence of each part of sememe on the overall meaning, and improving the calculation of word similarity based on HowNet by changing the specific calculation method of each part of sememe. At the same time, we adopt different strategies for the different results obtained in the similarity calculation of CiLin. The experimental RG data set proves that the modified Pearson coefficient of the method reaches 0.87.

1 Introduction

Word similarity calculation is one of the basic problems of natural language processing. The current research can be roughly divided into two modes of resource integration[1]: The first one is based on the fusion of semantic description level and statistical large-scale corpus. Second, integration on the ontological level .In the following chapters, after understanding the existing concepts of the HowNet and CiLin, firstly, in the calculation of the similarity of HowNet's sememe, we add the depth of the original sememe and the density of the position of the sememe. Secondly, we recalculate the calculation method of the similarity value in each part of the meaning of the similarity of the knowledge of the network. Then we make different choices in different situations to caculate the word similarity of the CiLin. Finally, we improve the distribution weight of the existing fusion network and the word forest method. According to the experiment at the end of the paper, the performance of the improved method is verified.

2 Research Background

2.1 Introduction to knowledge network

HowNet mainly contains concepts such as sememe, the semantic similarity and so on. Sememe is the atomic concept used to explain the meanings. It is the most basic unit of

* Corresponding author: feng.xw@163.com

HowNet. The meaning can be understood as a concept, which is used to explain words. A word can have multiple meanings. The semantic expression (DEF) is the main body of the meaning term, which is composed of the basic meanings of the combination of the knowledge description symbols and is used to explain the meaning of the meaning term. The basic data classification in HowNet can be as shown in Fig.1:

2.2 CiLin introduction

CiLin is a computable Chinese vocabulary used to realize the division and categorization of Chinese synonyms and similar words. CiLin has been expanded by the Information Retrieval Research Laboratory of Harbin Institute of Technology which has a five-layer tree structure as shown in Fig.2. The first layer is a large class that divided into 12 according to the concept category, coded as A~L. The second layer is a medium class, total of 95, which is encoded by a large class with a lower case letter. The third layer is a small class, which is represented by a medium class code followed by a two-digit decimal code. The fourth layer is the word group classification, which means the paragraph in the small class. The fifth layer is the atomic word group, which represents the lines in the paragraph. CiLin is stored in text. Each atomic group is a line, starting with 8 characters from the big class to the atomic word group, followed by one or more concepts represented by the character.

3 Word semantic similarity calculation

3.1 Based on HowNet's sememe similarity calculation

Liu Qun and Li Sujian[2] considered that the similarity of two words is the possibility that they can be replaced in different contexts without changing the syntactic and semantic structure of the text, and the formula is proposed:

$$sim(p_1, p_2) = \frac{\alpha}{\alpha + dis(p_1, p_2)} \tag{1}$$

In the formula, p1,p2 represents two sememe, dis(p1,p2) represents the distance between two sememe, and the adjustment parameter α represents the path length when the similarity is 0.5.

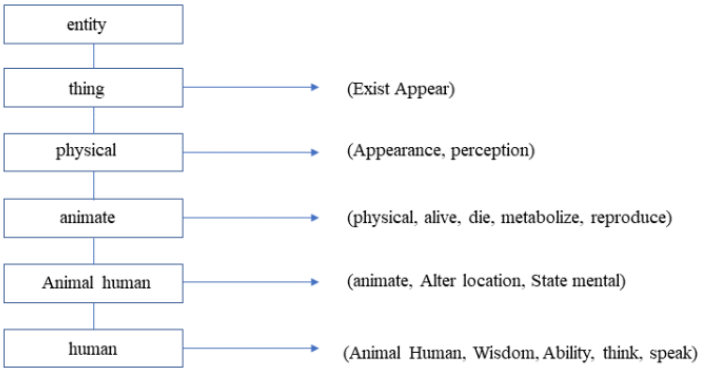

Fig. 1. The basic data classification of HowNet with "entity" as the root point.

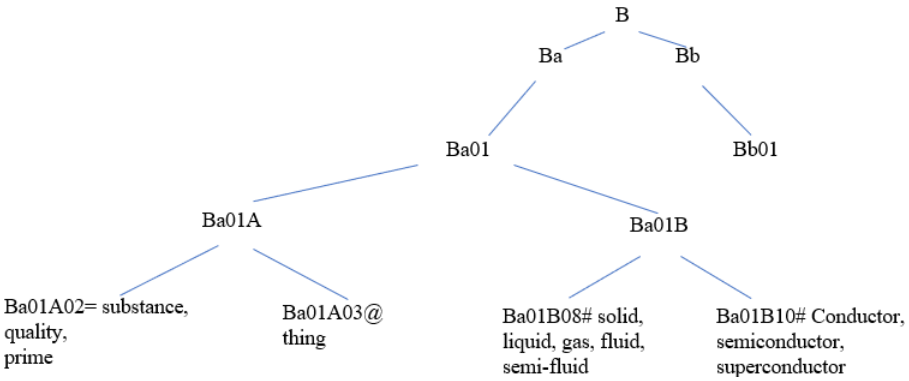

Fig. 2. CiLin topology.

Li Lei[3] and Zhu Xinhua[4] improved the edge weight formula follows:

$$weight(i_{p,q}) = c_1 \times \frac{depth - 1 - k_p}{depth - 1} \times (1 + \sin(\theta \times k_p \times \pi / 180)) + c_2 \times (1 - \frac{\log f(p)}{\log \max}) \quad (2)$$

In the formula, i(p,q) represents the path between the original nodes p and q where p is the current node and q is the parent of p;θ is the adjustment parameter that defined as 4 in here; Max represents the total number of all the original nodes of the sememe tree; c1 is 0.7 and c2 is 0.3.

According to the edge weight formula obtained by Equation (2), the distance formula for calculating the original node p1 and p2 is obtained. Where G represents the common parent of sememe. We substitute the formula (1) to find the similarity between p1 and p2.

$$dis(p_1, p_2) = \sum_{p=p_2, p \neq G}^{p_2} weight(i_{p,q}) \quad (3)$$

3.2 Calculation and improvement of words similarity in HowNet

For the two words W1 and W2, it is assumed that W1 has m meanings: S11, S12, S13, ..., S1m, W2 have n meanings: S21, S22, S23, ..., S2m.The similarity between two words is attributed to the similarity of two meanings, and the maximum value of each concept is the similarity of words W1 and W2.

$$sim(W_1, W_2) = \max_{i=1...m, j=1...n} sim(S_{1i}, S_{2j}) \quad (4)$$

According to Liu Qun[2], we derive the semantic similarity between the meanings.

$$\mathrm{sim}(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} sim_j(S_1, S_2) \quad (5)$$

In this paper, the specific steps of the improved calculation method for the four parts of sememe:

a) Set the similar vacations of the two sememe to -1 and traverse the S2 original.
b) Match the specified part of the two primitives (For example, the specified part is

the symbolic sememe part).

c) If the specified parts are identical, the similarity is 1; If the specified parts are different, the partial similarity is calculated by the formula proposed in 3.1; If one of the two meanings is a specific word, the similarity is directly assigned γ. Take the maximum value in this step

d) Repeat (b)(c) until all the original parts of the specified part match. If the maximum value is still -1 at this time, the similarity will be assigned to this part.

e) Determine the length of the specified part of the sememe S1 and S2. If S2 is longer than S1, we multiply the excess unmatched part of S2.

According to the meaning of each part of the sememe semantics for word semantics can be defined as β1+β2+β3+β4=1,and β1>=β2>=β3>=β4, β1 is 0.5，β2 is 0.2，β3 is 0.17，β4 is 0.13，α is 1.6，γ is 0.2，δ is 0.2 in this paper.

3.3 Calculation and improvement of the similarity of CiLin

In the exploration of CiLin, this paper refers to Peng Qi[12] based on the content similarity calculation of information content and the definition of Seco[5] in WordNet.

$$IC(C) = 1 - \frac{\log(hypo(C)+1)}{\log(\max nodes)} \tag{6}$$

where hypo(C) represents the number of lower nodes for concept C and maxnode represents the total number of nodes.

Firstly, this paper improves the similarity calculation of CiLin :

$$\mathrm{dis}(C_1, C_2) = IC(LCS(C_1, C_2)) - \frac{1}{2}(IC(C_1) + IC(C_2)) + 1 \tag{7}$$

From equation (7), the difference between the two concepts or synonyms that are identical is at least 0, and the similarity is 1. When the two concepts are leaf nodes and the nearest public parent is the root node, the concept of the two ontology edges is the most different. According to formula (6), the number of lower nodes of the root node is the total number of nodes, the information content of the root node is about 1. The number of lower nodes of the leaf node is 0, so the information content of the leaf nodes is also 1, and the similarity is 0. When the average similarity is greater than a smaller constant α, it means that each node has a certain correlation, but the correlation is not high. Propose the following formula for similarity:

$$sim(C_1, C_2) = \frac{\sum dis(C_1, C_2) - \max dis}{len - 1} \tag{8}$$

In the formula, maxdis represents the maximum value in dis(C1,C2), and len represents the number of groups included in the word. This article α takes 0.2. When the maximum similarity is greater than the constant β, indicating that the similarity of a pair of nodes reaches a higher level, it can directly defined as the maximum similarity. If none of the above is true, the minimum similarity is defined as the word similarity. Finally, the formula for the similarity calculation is discussed as below.

$$sim(W_1, W_2) = \begin{cases} \dfrac{\sum dis(C_1, C_2) - \max dis}{len - 1} \\ dis(C_1, C_2) \\ \min dis \\ \max dis \end{cases} \tag{9}$$

where W1, W2 represent two words, C1, C2 represent nodes.

3.4 Comprehensive HowNet and CiLin word similarity calculation

We consider the word similarity between HowNet and CiLin and calculate HowNet similarity s1 and CiLin similarity s2 for two words W1 and W2. Two similarities are assigned to HowNet similarity weights λ1 and CiLin similarity weights λ2. Comprehensive similarity calculation formula is like :

$$s = \lambda_1 s_1 + \lambda_2 s_2 \tag{10}$$

According to the inclusion of two words, it can be divided into the following situations:

(1) HowNet similarity is used when both words are only included in HowNet, λ1=1，λ2 =0.

(2) CiLin similarity is used when both words are only included in CiLin, λ1=0，λ2 =1.

(3) If a word is only included in CiLin and another word is only included in HowNet, look for synonyms in CiLin. If there is no synonym, we record the similarity of CiLin as 0.2. If there has synonym, we calculate the synonym in HowNet.

(4) If a word is only included in HowNet and another word is included in HowNet and CiLin, we look for synonyms in CiLin. If there is no synonym, the similarity is determined by HowNet similarity. If there has synonym, we look for the synonym and the value with the highest similarity serves as the CiLin similarity.

(5) If a word is only included in CiLin and another word is included in the common inclusion, we look for synonyms of the words contained only in CiLin. If there is no synonym, the similarity depends on the similarity of CiLin. If there has synonym, we find the synonym and medium similarity maximum as HowNet similarity.

(6) If the words are included neither in HowNet nor in CiLin, the similarity cannot be calculated.

4 Experimental results and analysis

4.1 Determination of the value of λ1 and λ2 weighting factors

This paper uses the internationally popular 30 pairs of word data sets published by Miller & Charles (MC)[6] and the word data sets published by Rubenstenin & Goodenough (RG)[7] as test cases. When the words are included in HowNet and CiLin, the similarity weights λ1 and λ2 set 4 different weights, and 30 pairs of words in the MC30 data set are used as the research object, and the Pearson values are as shown in Table 1.

From the Table 1, we can see that when both words are included in HowNet and CiLin, the similarity of CiLin is 0.9 and the similarity of HowNet is 0.1, the Pearson value reaches the highest. When a word is only included in the CiLin or HowNet and the other word is both included, we set 4 different weight factor combinations to calculate the word similarity by

comparing with the artificial test value of RG65. We choice the best weight that is more in line with the actual situation. The experimental results are shown in Table 2.

In Table 2, when λ1 and λ2 are (0.7, 0.3) and (0.6, 0.4), the similarity between the words "pillow" and "pillow" differs greatly from the value of manual evaluation. When λ1 and λ2 are (0.4, 0.6) and (0.5,0.5), although they differ from the manual evaluation value by 0.004, for other groups of words such as "mountain" and "slope", the word similarity of (0.4, 0.6) reaches 0.558, and (0.5, 0.5) is much different from the manual test value. Therefore, when words are not included in CiLin and HowNet, λ1= 0.4 and λ2= 0.6 can obtained higher similarity.

Table 1. Comparison of Pearson coefficients of different weighting factors.

CiLin similarity weight	HowNet similarity weight	Pearson value
0.9	0.1	0.914
0.8	0.2	0.904
0.7	0.3	0.886
0.6	0.4	0.858

Table 2. Comparison of word similarity between different weighting factors.

First word	Second word	Manual test value	（0.7，0.3）	（0.6，0.4）	（0.5，0.5）	（0.4，0.6）
Berm	seaside	0.2425	0.259	0.252	0.338	0.377
Mountain	forest	0.37	0.099	0.084	0.07	0.056
Graveyard	Grave	0.4225	0.142	0.083	0.238	0.285
Mountain	Slope	0.8225	0.527	0.537	0.548	0.558
pillow	pillow	0.96	0.938	0.947	0.956	0.964
Graveyard	Cemetery	0.97	1	1	1	1

4.2 Comparative experiment

Through experiments, we can see that the comparison between Table 3 and Table 4 shows that the Pearson coefficient of word similarity in the fusion of HowNet and CiLin is better than other experiments. The merged method can more reflect the difference between words and the obtained word similarity is more scientific.

Table 3. Pearson correlation coefficient of RG artificial value between different methods.

Similarity method	Word dictionary	RG Pearson coefficient
Liu Qun[2]	HowNet	0.699
Seco[5]	HowNet	0.738
Tian Jiule[8]	CiLin	0.53
Chen Hongchao[9]	CiLin	0.85
This paper	HowNet and CiLin	0.87

Table 4. Comparison of RG word pair set calculation results.

word	word	HowNet	CiLin	This paper	RG manual test value
水果	火炉	0	0.32	0.289	0.0125
署名	海滨	0.523	0	0.052	0.015
汽车	巫师	0	0	0	0.0275
高地	火炉	0	0	0	0.035
大笑	器械	0	0	0	0.045
庇护所	水果	0	0	0	0.0475
庇护所	和尚	0.481	0	0.048	0.0975
墓地	精神病院	0.529	0	0.053	0.105
男孩子	公鸡	0	0	0	0.11
垫子	宝物	0	0.275	0.248	0.1125
庇护所	墓地	0.465	0.136	0.169	0.1975
大笑	小伙子	0	0	0	0.22
男孩	圣人	0.6	0	0.06	0.24
汽车	垫子	0	0.314	0.283	0.2425
护堤	海滨	无	0.142	0.378	0.2425
海滨	航行	0	0	0	0.305
鸟	树林	0	0.296	0.266	0.31
火炉	器械	0.554	0.326	0.348	0.3425
鹤	公鸡	1	0.703	0.733	0.3525
山岗	树林	无	0.142	0.057	0.37
墓地	坟堆	无	0	0.286	0.4225
玻璃	珠宝	0.289	0.29	0.29	0.445
魔术师	圣贤	0.579	0.3	0.328	0.455
圣人	巫师	0.579	0.356	0.378	0.615
圣贤	圣人	1	1	1	0.6525
山岗	斜坡	无	0.496	0.558	0.8225
绳索	绳子	1	1	1	0.8525
玻璃	杯子	0.289	0.282	0.283	0.8625
大笑	微笑	0	1	0.9	0.865
农奴	奴隶	0.6	1	0.96	0.865
署名	签名	1	1	1	0.8975
森林	树林	1	1	1	0.9125
雄鸡	公鸡	1	1	1	0.92
靠枕	枕头	无	0.912	0.965	0.96
墓地	墓园	无	1	1	0.97

5 Conclusion

The improved semantic similarity method for HowNet and CiLin proposed in this paper. We combine CiLin and HowNet to make full use of the information content of words in different knowledge bases, complement each other's missing points in the knowledge base, correct the corresponding rough points and improve the limitations of knowledge description language. However, the similarity of some words in the experiment is still not ideal. The similarity of HowNet can still be improved and the fusion of the two can more considered so that the similarity of the similarity has higher reliability.

This research was financially Supported by the Natural Science Foundation of Liaoning Province of China(20180550130).

References

1. Mei Lijun, Zhou Qiang, Qi Lu, Chen Zuyu. Research on Information Fusion of HowNet and CiLin. In *Chinese Journal of Information*. 2005.19(01)pp:63-70.

2. Liu Qun, Li Sujian. Vocabulary semantic similarity calculation based on HowNet. In *The 3rd Chinese Vocabulary Semantics Seminar*. 2002. pp:59-76
3. Li Lei, Yang Lihua. Improved algorithm for semantic similarity of words based on HowNet. In *Computer technology and development* . 2019.29(04)pp:42-46.
4. Zhu Xinhua, Ma Runcong, Sun Liu, Chen Hongchao. The semantic similarity calculation of words based on HowNet and CiLin. In *Chinese Journal of Information.* 2016. 30(04)pp:29-36
5. Seco N，Veale T，Hayes J． An intrinsic information content metric for semantic similarity in WordNet. In *Proc of Eureopean Conference on Artificial Intelligence*．2004.pp:1089-1090．
6. Miller G A，Charles W G． Contextual correlates of semantic similarity. In *Language Cognition & Neuroscience* . 1991.6(1)pp:1-28．
7. Rubenstein H，Goodenough J B． Contextual correlates of synonymy.In *Communications of the ACM*.8(10) 1965. pp:627-633．
8. Tian Jiule, Zhao Wei. 2010. Method for calculating similarity of words based on CiLin. In *Journal of Jilin University (Information Science Edition)*,28(06)pp:602-608.
9. Chen Hongchao, Li Fei, Zhu Xinhua, Ma Runcong. Similarity Calculation of CiLin Based on Path and Depth. In *Chinese Journal of Information*. 2016.30(05)pp:80-88.