Word Semantic Similarity Based on CiLin and Word2vec

Yushang Mao
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
maoyushang2018@ia.ac.cn

Guixuan Zhang
Beijing Digital Content Research Center &AICFVE
Beijing, China
guixuan.zhang@ia.ac.cn

Shuwu Zhang
Beijing Digital Content Research Center
Beijing, China
shuwu.zhang@ia.ac.cn

Abstract—This paper presents a method to calculate the semantic similarity with TongyiciCiLin and Word2vec. In the part of CiLin, the semantic similarity of words is calculated by using the distance of words as the main factor, the number of branches and the distance between branches as the fine-tuning parameters. In the part of Word2vec, this paper constructs a special Corpus based on movie review, and uses Word2vec model to calculate the semantic similarity of Chinese words. Then, the final semantic similarity is calculated by using the dynamic weighting strategy to fuse CiLin and Word2vec. The method makes full use of the semantic information of words in the knowledge base and Corpus. The experimental results show that the algorithm has better accuracy and more robust to domain sensitivity.

Keywords-semantic similarity; CiLin; Word2vec; semantic distance; word embedding

I. INTRODUCTION

Word similarity is used to calculate or compare the similarities between two words. Word similarity is widely used in natural language processing, intelligent retrieval, text clustering, text categorization, automatic question answering, word sense disambiguation and machine translation. At present, domestic and foreign research strategies for word similarity calculation are mainly divided into two categories: one is the similarity calculation method based on knowledge, which is mainly based on the subordinate relationship and homology between concepts in such linguistic resources relationship to calculate the similarity of words. The knowledge-based method is simple and effective, does not require corpus training, and is relatively intuitive and easy to understand. However, the results obtained by this method are greatly affected by human subjective consciousness, and sometimes they cannot accurately reflect objective facts. The second is a corpus-based method, which mainly uses the probability distribution of context information as a reference basis for lexical semantic similarity. Corpus-based methods are more objective and comprehensively to reflect the similarities and differences in terms of syntax, semantics, and pragmatics. However, this method is more dependent on the corpus used for training. It has large calculation complexity and is greatly disturbed by data sparsity and data noise.

This article proposes a word similarity method based on the combination of knowledge and corpus. A word is represented through CiLin and Word2vec, and then the dynamic word weighting strategy is used to calculate the final word semantic similarity. This method has achieved good results through verification.

II. RELATED WORK

A. Knowledge-based Semantic-similarity Methods

In recent years, with the emergence and continuous improvement of ontology knowledge models, research on the similarity of Chinese words has begun to flourish. Q. Liu et al. [1] proposed a method based on Hownet to calculate the semantic similarity of words, which adopted the weighted average of partial similarity to represent the overall similarity when calculating the similarity between the semantic expressions of the two concepts, and obtained the similarity of the two meanings according to the upper and lower relationship semantic distance and conversion method. L. H. Lv et al. [2] used CiLin to comprehensively consider the density information and path information of words in the dictionary, and simulated the calculation function to calculate the similarity. Z. J. Zhan et al. [3] put forward a new method of word similarity measurement based on BaiduBaike. By analyzing the information of BaiduBaike terms, the term similarity is comprehensively analyzed from the aspect of explaining the content, also the formula for calculating the term similarity is defined. And the whole similarity can be obtained by calculating the similarity between the parts.

B. Corpus-based Semantic-similarity Methods

Corpus is the basic resource of linguistic research. The semantic similarity calculation model based on corpus generally used a large-scale corpus such as news and novels to calculate semantic relevance by statistical methods. In the corpus-based semantic similarity calculation method, the idea of the bag-of-words model proposed by Salton played a very important role in later research [4]. When representing the sample data, the bag-of-words model simplifies the model through the following assumptions: a text or

document can be regarded as a bag of words, the grammar and word order relationships in the text or document are not considered, and each word is independent. In 2013, Google proposed the Word2vec model [5-7] based on the continuous bag-of-words and skip-gram, which attracted wide attention. However, the main shortcoming of the continuous bag of words model is that cannot solve the problem of polysemy. E. H. Huang et al. [8] combined documents and sentences, and applied the model to word sense disambiguation, achieving better performance than the continuous bag of words model. F. Tian et al. [9] trained a set of context-sensitive semantic models according to the different ideas of the context of words with different meanings, and showed better results in the experiment of word similarity.

C. *Hybrid Methods*

Considering to absorb the advantages of the two methods, some researches have begun to propose the hybrid computation of the two methods, that is, the feature of knowledge is also taken into account when the Corpus is used to calculate similarity. R. Mihalcea et al. [10] use a mixture of corpus-based and knowledge-based method, the error rate of the traditional corpus-based approach in the interpretation experiments of the Microsoft Research is 13% lower than that of the traditional one. J. Y. Zhai et al. [11] proposed a hybrid semantic correlation calculation method that comprehensively considers the meaning of words, the weight of words in sentences, and the structure of sentences. The similarity calculation for sentences has performed better than the calculation using TF-IDF algorithm.

III. PROPOSED METHOD

A. *The Part of CiLin*

CiLin is a computable Chinese lexicon compiled by Mei Jiaju in 1983. Its design goal is to achieve the division and classification of Chinese synonyms and similar words. After expanded by Harbin Institute of Technology Information Retrieval Laboratory, there are currently more than 70,000 words. These words are divided into 12 major categories, 94 middle categories, and 1,428 subcategories. The twelve categories are: the first to third categories are mostly nouns; the fourth category are numerals and quantifiers; the fifth category is mostly adjectives; the sixth to tenth categories are mostly verbs; the eleventh categories are mostly function words; the twelfth category are difficult to be treated as the words in the above categories. The ordering of major and middle categories follows the principle which is from concrete concept to abstract concept. The subcategories are further divided into two levels of word group and atomic word group. In this way, the expanded version of CiLin has a five-layer tree structure.

There are only three cases for the eighth bit code, where "=" means "equal" and "synonymous"; "#" stands for "not equal" and "similar", which are related words; "@" stands for "self-closing" and "independent", that is, there are neither synonyms nor related words in the dictionary. After the current seven-bit code is determined, the eighth bit is fixed, either "=", "#", or "@". The CiLin used in this article comes from CiLin 1.0, an extended version of CiLin developed by the Harbin Institute of Technology Information Retrieval Research Office. The coding of entries is shown in Table I.

TABLE I. CODING STRUCTURE OF WORDS IN CILIN

Code bit	1	2	3	4	5	6	7	8
example	A	a	0	1	B	0	2	=/#/@
Nature	Major	Middle	Small		Word Group	Atomic Word Group		
Rank	First Layer	Second Layer	Third Layer		Fourth Layer	Fifth Layer		

The structure of CiLin is a five-layer tree structure, as shown in Fig. 1. In particular, the connection path of the two words in the word tree is the main factor affecting the similarity of words. The first layer of CiLin is a large category, and the distance between words that do not belong to the same large category is treated as 18, and the four edges connecting the upper and lower layers are given a weight Wi ($1 \le i \le 4$) from the bottom to the top, and $0.5 \leqslant W1 \leqslant W2 \leqslant W3 \leqslant W4 \leqslant 5$, $W1+W2+W3+W4 \leqslant 10$. In this paper, the weights of these four types of edges are assigned value of 0.5, 1, 2.5 and 2.5. Since the word encoding is on the fifth-level leaf node, the word encoding distance **d** can be taken as 1, 3, 8, 13 , 18.

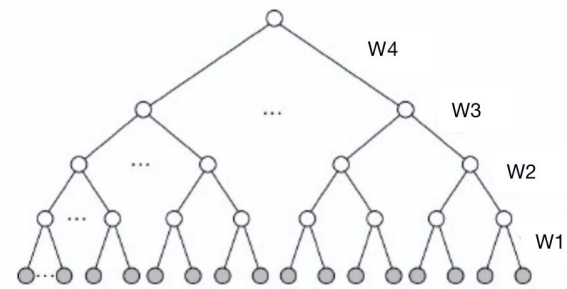

Figure 1. The Five-layer Tree Structure of CiLin.

There are also two secondary factors that affect the similarity of words: the total number of nodes at the branch layer **n**, and the separation distance **k** of the branches where the two words are located. The total number of nodes at the branch layer **n** reflects the density of the common parent node and word similarity is proportional to it. At the same level, CiLin classifies and arranges words in a certain semantic order, so the branch interval **k** is inversely proportional to the similarity. X. H. Zhu et al. [12] proposed a formula for calculating the similarity of words in the synonymous vocabulary with the word distance **d** as the main influencing factor, the number of branch nodes **n** and the branch interval **k** as the adjustment parameters.

$$\mathrm{sim}(C_1, C_2) = (1.05 - 0.05 dis(C_1, C_2))\sqrt{e^{\frac{-k}{2n}}} \qquad (1)$$

In formula (1), dis (C_1, C_2) is the distance function of the word codes C_1 and C_2 in the tree structure, which is equal to the sum of the weights of the sides of the connection path of the word pair, which can take the values 2*W1, 2*(W1+W2),

2*(W1+W2+W3), 2*(W1+W2+W3+W4). In this paper, this formula is used to calculate the similarity. When two words follow the same "=" in the code, the similarity is processed as 1; when the same "#" is coded, the similarity is processed as 0.5. When two words are not in a large category, the distance between words is treated as 18. When a word corresponds to multiple codes, the similarity of all code combinations is calculated, and the maximum similarity is taken as the similarity of the words.

B. *The Part of Word2vec*

In 2013, Google proposed the Word2vec model. Word2vec can efficiently train words in the corpus into vector forms through efficient training on millions of dictionaries and hundreds of millions of data sets. Each word generates a unique word vector. Word2vec simplifies the processing of text content into a vector operation method in a vector space. This way allows all vocabulary to have the ability to perform mathematical operations. Each vocabulary has a unique vector representation and a unique coordinate position in space, which can introduce the concept of "distance" between words. By calculating the size of the cosine angle or the spatial distance through two vectors, you can determine the spatial extent of the two vocabularies. This distance can be used to express the semantic similarity of the two vocabularies, that is, it can be transformed into a distance comparison between two vectors when two words similarity are compared. This article uses the cosine angle for the calculation method of the semantic similarity of two word vectors, which is defined by:

$$\cos(\theta) = \frac{\vec{A} \bullet \vec{B}}{\|\vec{A}\| \bullet \|\vec{B}\|} = \frac{\sum_{i=1}^{n} (\vec{A}_i \times \vec{B}_i)}{\sqrt{\sum_{i=1}^{n} (\vec{A}_i)^2} \times \sqrt{\sum_{i=1}^{n} (\vec{B}_i)^2}} \tag{2}$$

where **A** and **B** are two word vectors. The larger **cos(θ)** after the cosine angle calculation, the greater similarity between the two words. The corpus selected in this article is more than 500,000 long movie reviews, and the size is 4.64GB.

C. *Fusion Algorithm of CiLin and Word2vec*

The idea of considering the similarity calculation of CiLin and word2vec is: For any two words **A** and **B**, use CiLin and Word2vec to calculate the two similarities of the words, and record them as $\mathbf{S_1}$ and $\mathbf{S_2}$. Each similarity is respectively given weights $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$, which also satisfies $\boldsymbol{\lambda_1} + \boldsymbol{\lambda_2} = \mathbf{1}$, and then the semantic similarity of words in the comprehensive CiLin and Word2vec is calculated according to equation (3):

$$\mathrm{S} = \lambda_1 S_1 + \lambda_2 \mathrm{S}_2 \tag{3}$$

Due to the particularity of the vocabulary in the field of film, it is more about considering the context information of the vocabulary, and adopts the following dynamic weighting calculation strategy:

- When $\mathbf{S_1}=0$ and $\mathbf{S_2}\neq 0$, $\boldsymbol{\lambda_1}=0$ and $\boldsymbol{\lambda_2}=1$.
- When $\mathbf{S_1}\neq 0$ and $\mathbf{S_2}=0$, $\boldsymbol{\lambda_1}=1$ and $\boldsymbol{\lambda_2}=0$.
- When $\mathbf{S_1}\neq 0$ and $\mathbf{S_2}\neq 0$, $\boldsymbol{\lambda_1}=0.3$ and $\boldsymbol{\lambda_2}=0.7$.

For the case where $\mathbf{S_1}$ and $\mathbf{S_2}$ are equal to 0 at the same time, it is not temporarily considered.

IV. Result and Analysis

At present, there is no unified standard for the research of Chinese word similarity, so it is difficult to design the experiment and select the data of Chinese word for similarity calculation. In this paper, wordsimilarity-353 pairs in English are translated into Chinese as the experimental standard data set. The data set includes 353 pairs of English words whose similarity values are obtained by artificial judgment and can be used to collect, train or test methods for computers to perform semantic similarity. A randomly selected set of standard data word pairs is shown in Table 2.

TABLE II. Some of Experimental Data

English	Standard value	Chinese
journey-voyage	0.929	旅程-航程
money-cash	0.908	货币-现金
computer-software	0.85	计算机-软件
network-hardware	0.831	网络-硬件
nature-environment	0.831	自然-环境
psychology-Freud	0.821	心理学-弗洛伊德
news-report	0.816	新闻-报告
war-troops	0.813	战争-部队
bank-money	0.812	银行-货币
stock-market	0.808	股票-市场
energy-secretary	0.181	能源-秘书
stock-phone	0.162	股票-手机

TABLE III. Some of Experimental Results

English	Chinese	Standard value	Based on Hownet	Based on Baidubaike	Based on CiLin and Word2vec
journey-voyage	旅程-航程	0.929	0.04	0.63	0.85
money-cash	货币-现金	0.908	1	0.57	0.90
computer-software	计算机-软件	0.85	0.44	0.78	0.86
network-hardware	网络-硬件	0.831	0.29	0.54	0.83
nature-environment	自然-环境	0.831	0.05	0.62	0.85
psychology-Freud	心理学-弗洛伊德	0.821	0	0.61	0.78
news-report	新闻-报告	0.816	0.62	0.65	0.73
war-troops	战争-部队	0.813	0.15	0.75	0.86
bank-money	银行-货币	0.812	0.11	0.72	0.82
stock-market	股票-市场	0.808	0.11	0.45	0.65
energy-secretary	能源-秘书	0.181	0.10	0.08	0.12
stock-phone	股票-手机	0.162	0.26	0.11	0.53

The methods based on Hownet in the article [1] and based on Baidubaike in the article [3] are compared with the proposed method. The experimental results are listed in Table III. The results based on Hownet are better for some terms with obvious conceptual relationships, such as “货币-现金”. However, the Hownet based method cannot handle the words which do not appear in Hownet, so it fails to calculate the similarity for new and uncommon words, such as the pair “股票-市场”. The method based on Baidubaike cannot reflect the contextual relation of words well. In contrast, the proposed approach can calculate the word similarity more precise for most cases. It demonstrates the effectiveness of the proposed fusion method.

V. CONCLUSION

This paper presents a method to calculate the semantic similarity between CiLin and Word2vec. Different from the traditional method based on semantic dictionary and large-scale Corpus, this paper calculates the final semantic similarity of words by considering the dynamic weighting strategy of CiLin and Word2vec. The experimental results show that the new method produces better results than the existing methods. Based on the existing research on word similarity, this paper will further analyze the similarity features between words and sentences, and consider the structural information of sentences to improve the effect of semantic similarity.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (2018YFB1403900) and the Science and Technology Program of Beijing (Z201100001820002). It was also the research achievement of the Key Laboratory of Digital Rights Services, which is one of the National Science and Standardization Key Labs for Press and Publication Industry.

REFERENCES

[1] Q. Liu, S. Li, Word similarity computing based on Hownet [J]. Taipei: The Third Chinese Vocabulary Semantics Seminar, 2002.

[2] L. H. Lv, W. Liang, S. Ran, A method for measuring word similarity based on CiLin [J]. Modern Computer, 2013, pp.3-6.

[3] Z. J. Zhan, L. Liang, X. Yang, Word similarity measurement based on BaiduBaike [J]. Computer Science, vol. 40, 2013, pp.199-202.

[4] G. Salton, Developments in automatic text retrieval [J]. Science, American Association for the Advancement of Science, vol. 253, 1991, pp. 974-980.

[5] M. Tomas, K. Chen, G. Corrado, Efficient estimation of word representations in vector space[J]. Computer Ence, 2013.

[6] M. Tomas, I. Sutskever, K. Chen, Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, pp.3111-3119.

[7] M. Tomas, Y. W-T, G. Zweig, Linguistic regularities in continuous space word representations [J]. HLT-NAACL, 2013, pp.746-751.

[8] E. H. Huang, R. Sochor, C. D. Manning, Improving word representations via global context and multiple word prototypes [J] Proceedings of ACL. Jeju, Korea:ACL Press, 2012, pp.873-882.

[9] F. Tian, H. Dai, J. Bian, A probabilistic model for learning multi-prototype word embeddings [J]. Choose, 2014, pp.151-160.

[10] R. Mihalcea, C. Corley, C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity [J]. AAAI, 2006, pp.775-780.

[11] J. Y. Zhai, A hybrid measurement for sentence similarity based on semantic [J]. Science Technology and Engineering, vol. 14, 2014, pp.81-85.

[12] X. H. Zhu, R. C. Ma, Word semantic similarity computation based on HowNet and CiLin [J]. Journal of Chinese Information Processing, 2016.