

Research on Calculation Method of Semantic Correlation of Chinese Text

Yan-Fang Li¹, Shu-Tao SUN², Shuang FENG³ AND Qi WANG⁴

^{1, 2, 3, 4}Communication University of China, Dingfuzhuang NO. 1, Chaoyang District, Beijing, China

Abstract. The semantic correlation is a very important research direction in Natural Language Processing, and the semantic relatedness is different from the semantic similarity. At present, the semantic correlation algorithm, which is based on the similarity of the semantic meaning, and it can't achieve the desired results in a certain extent. In this paper, by using the large-scale corpus and How Net words' concepts to dig out the hidden semantic relations between the words, finally, according to the semantic relations of the words in the text, the text correlation algorithm is proposed. Experimental results show that the use of How Net calculation of the text relevance which use less time than the use of large-scale corpus and Tongyici Cilin calculation of the text relevance, and in terms of accuracy of the calculation results also have to upgrade.

1 Introduction

With the popularity of the Internet, there are a variety of resources to grow in an explosive way every day, and their share of the degree is also getting higher and higher. The information is convenient for people's life, but how to deal with and use it, which is the top priority. According to reliable statistics, 70.2% of the information is in the form of text, and how to present the specific field of text information to the user, which is the problem that needs to be solved currently.

At present the theory and technology of existing natural language processing are mostly in English as the research object, because English is hypotaxis (merplotactic) language and sentence structure requirements of word form which change in accordance with the rules, and pay attention to the syntactic level. But Chinese pronunciation, shape, word formation, word forms and semantic aspects are quite different, these significant differences let western the current relatively mature theory and technology can't be directly in Chinese information processing.

At present, there are two methods for calculation the words similarity, which are based on the semantic dictionary and statistical method. In this paper, the word similarity algorithm is improved to calculate the correlation of the words, and then calculate the correlation of the text.

2 Basic Knowledge

In this chapter, some preprocessing are required before calculating the correlation of the Chinese text. In this experiment, it is very important to calculate the correlation of the words, because this result can directly affect the calculation results of the correlation of the text. Therefore, the first step is to text segmentation, and then is the data pre processing. This chapter is the introduction of these technologies.

2.1 HowNet

How Net use concepts to describe object, the concept is in Chinese and English. HowNet is the basic content of commonsense knowledge, which reveals the relationship between concepts and the concept of attribute [1]. It uses Knowledge Dictionary Mark-up Language KDML, which is a special knowledge dictionary. The word is expressed as a number of "concepts", the use of "concept" to describe the semantic meaning of the words [2]. So before calculation the correlation of the words, the first is to read glossary. dat, because the file stores the words and its concept. To “安慰”(comfort), there are three concepts, as follows:

安慰: V AtEase|安心

安慰: N emotion|情感, AtEase|安心

安慰: V soothe|安慰

There are comma in word's concept, and the function of the comma is used to separate word's concepts. The words is included in the HowNet which can be divided into the

content words and the function words. In this paper, when we use HowNet to calculate the semantic relatedness, we will always use the content words. And the HowNet primitive can be divided into the first primitive, other primitive, relational primitive, relational symbol primitive. In this paper, we mainly focus on the relationship between the words to compute the semantic relatedness of the words.

2.2 Tongyici Cilin

In this experiment, the use of the dictionary is the Harbin Institute of Technology Information Retrieval Lab TongyiciCilin extended edition [3]. It is edited by Harbin Industrial University Information Retrieval Laboratory which use many words related resources, and put a lot of manpower and material resources to complete the Chinese words table. However, at present, the full version of the thesaurus does not share, only the dictionary file can be used, and the dictionary file records some of the words. Such as: computer Bo01A27=, computer Bo01A27=, microcomputer Bo01A27=, microcomputer Bo01A27=. The thesaurus dictionary mainly use a collection of some characters to represent the meaning of the words. If the meaning of the two words is the same, the two words will have a set of the same character set. If the words is an ambivalent word, there will be many groups of character set to represent the different meanings of the words.

2.3 The Data Preprocessing

The first step is the data preprocessing when we calculate the correlation of the text. The data preprocessing includes word segmentation, feature extraction and de - stop word. The purpose of feature extraction is to reduce the dimension of data and remove noise. In this experiment, we use the feature extraction and that is TFIDF. TF-IDF(term frequency-inverse document frequency) is a commonly used weighting technique, which is mainly used for data mining and information retrieval. TF-IDF is a statistical method that uses to assess the importance of a word [4]. The importance of the words is proportional to the presentation times in the document, but it is inversely proportional to the frequency of the word in corpus. In this experiment, we need to use a large number of text sets to calculate the TFIDF of all words in a text, the formula of TFIDF is:

$$TFIDF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} * \log \frac{|D|}{\{j:t_i \in d_i\}} \quad (1)$$

The $n_{i,j}$ expresses the number of times that a word appears in the text, the $\sum_k n_{k,i}$ expresses the total number of the words in the text, $|D|$ expresses the total number of text in corpus. $\{j:t_i \in d_i\}$ expresses the number of text that appears in the text of a certain word. The formula shows the high frequency of the word in a particular file, and the low file frequency in the entire document collection, can

produce a high weight of TF-IDF. Therefore, TF-IDF tends to filter out the common words and keep the important words.

The data preprocessing also includes word segmentation. in this experiment, the word segmentation method is ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System) Chinese word segmentation system, and it is a Chinese word segmentation system which is developed by a number of researchers in the Institute of computing technology of the Chinese Academy of Sciences after years of research. Its main functions include Chinese word segmentation, part of speech tagging, microblogging word, word recognition, named entity recognition, keyword extraction. At the same time it also supports the user dictionary, support traditional Chinese, supports GB2312, GBK, UTF8 and other coding formats. ICTCLAS can be widely used, because it has a good segmentation effect, and the word segmentation accuracy is up to 98.45%.

3The Calculation Method of Word Correlation Based On HowNet

Using HowNet calculates the correlation of the two words, mostly according to the relationship between the word similarity and word correlation. The semantic similarity of the words is that the words can be replaced by each other in different texts, which means the words' meaning is the same [5]. In this paper, we should compute word correlation, which is to calculate the word similarity and the semantic relationship according the relationship of the words, and then the two data are weighted average as the final result.

According to HowNet word, similarity calculation formula is as follows [6]:

$$sim(p_1, p_2) = \begin{cases} \frac{\delta}{(\delta+d)}, & if(pos = 1) \\ 1, & if(pos \neq 1, p_1 = p_2) \\ 0, & if(pos \neq 1, p_1 \neq p_2) \end{cases} \quad (2)$$

p_1 and p_2 represent the word concept, d represents the path length of p_1 and p_2 in the original hierarchy, δ is a regulating parameter, pos represents the position of concept in the concept description, $pos=1$ represents the word concept is the first primitive. In HowNet the words' concepts are primarily divided into three types, respectively the first primitive, other primitive, the relational primitive and the relational symbol primitive as a type. In the calculation of the similarity of the words, the formula will get a lot of results, then make the maximum value as the final result. In this experiment, we mainly compute the correlation of the words based on HowNet concept relation, and table 1 is a summary of the symbol about the relational symbol primitives.

Table 1. Summary of the symbol.

Symbol	Definition
#	This symbol represents the related concept.
%	This symbol represents the overall concept.
\$	This symbol represents the object.
*	This symbol represents the main body of action.
&	This symbol represents the host.
@	This symbol represents the time or place.
?	This symbol represents the material.

Symbol in Table 1 describes the symbol's meaning in relational symbol primitives. In this experiment, the researchers finds that there are eight relations in the word concepts, namely: universal correlation, Finished product - the whole-property of the host relationship, original meaning/ Finished goods/ host-first primitive, original meaning/ Finished goods/ host-dynamic role object, Agent-object-time/ place relationship, Agent/object/time-first primitive, Agent/object/time-dynamic role object, First primitive-dynamic role object. And universal correlation can be divided into eight kinds, namely: two words have the same related concept; a word's related concept is other word's first primitive; a word's related concept is other word's other primitive; a word's related concept is other word's overall concept; a word's r related concept is other word's dynamic role object; a word's related concept is other word's Finished goods; a word's related concept is other word's event semantics; a word's related concept is other word's host. According to these eight relationships, the weight of each relationship is r_i , and $\sum_{i=1}^8 r_i = 1$, then the formula for calculating the correlation of the word is as follows:

$$asso(p_1, p_2) = \begin{cases} 1, & \text{if } (p_1 = p_2) \\ 0, & \text{if } (p_1 \neq p_2) \end{cases} \quad (3)$$

p_1 and p_2 represent the word concept, The calculation method for the correlation of Chinese words is as follows:

$$asso(s_1, s_2) = \sum_{i=1}^8 r_i \times asso(p_1, p_2) \quad (4)$$

The higher the similarity of the two words, the higher they are related. So when we calculate the correlation degree of the words, we should also take into account the similarity of the words. According to the similarity of the words and the correlation of the words, the formula of calculating the correlation of the words is obtained, as follows:

$$Sim(s_1, s_2) = \alpha_1 \times sim(s_1, s_2) + \alpha_2 \times asso(s_1, s_2) \quad (5)$$

And $\alpha_1 + \alpha_2 = 1$, α_1 and α_2 is adjustable parameters, which are obtained by a lot of experiments and experience to assign. In this experiment, α_1 is 0. 4, and α_2 is 0. 6.

4The Calculation Method of Word Correlation Based On Cilin and Corpus

In this paper, we use the Cilin, which is mainly used to calculate the synonyms. In this chapter, the understanding of the relevance of words, that is, the probability of statistics of two words is in an article. Therefore, we need to crawl to get a large of texts, of course, as the text is marked by a specific area. For this experiment, the use of the corpus is a type of sports news text. However, in this experiment, we exclude the ambiguous words, when a word has a number of meanings, we will not calculate their synonyms which appearing in the text. But in this experiment, when a word has a number of synonyms, we calculate thesynonym which appearing in the text in corpus. Mainly because a word if there are a number of different meanings, on behalf of the different contexts in which the meaning will be different, so if you get rid of such words which can increase the accuracy of the calculation.

The calculation formula is as follows:

$$wordCount = AB / (A + B - AB) \quad (6)$$

According to the following figure, you can clearly understand that the use of corpus to calculate the relevance of words.

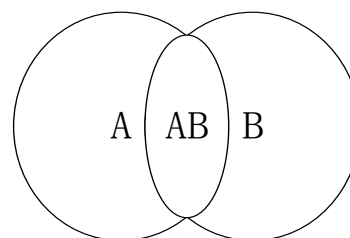


Figure 1. Algorithm principle of this chapter.

The AB represents the number of these two words and their synonyms appear at the same time in the text in corpus, A represents the number of A and its synonyms appear in corpus, and B represents the number of B and its synonyms appear in a corpus. But in statistical terms, the number of text, if the words or any of its synonyms has appeared once, don't repeat counting.

5 Method for Calculating Text Correlation Based On Word Correlation

Above the use of HowNet, Cilin and corpus compute the word correlation, and this chapter is according to the word correlation calculation the textual relevance. Because Chinese sentence pattern structure is complex and each person has different ways, and according to the relationship

between sentences to compute the text relevance which is very complex and that can be very difficult nowadays. In this experiment, we only calculate the relevance of the words in the text, and then calculate the correlation of the text.

After the data preprocessing, the remaining words of permutation and combination, calculate the correlation of each group of the words and the calculation results are stored in a matrix, which the rows of the matrix is the words of a text, and the columns is the words of other text. Then traverse the matrix, find the maximum value, and the value stores in an array, then remove all the values of the row and column of the value, and then continue to traverse until the matrix is empty. Finally, the maximum value is found by the statistics of the array. Then the weighted average of these values is obtained, and the result represents the correlation of the two texts, and the value is obtained by the relationship between the words of the two texts.

6 Comparison of Experimental Results

The purpose of this experiment that it is to use the method mentioned in this paper to calculate text correlation, then compare the two methods which has the high accuracy, and it also needs to consider the efficiency of running the program, such as the use of corpus, the time of reading the text. The figure 2 is the use of the two methods to calculate the results screenshot.

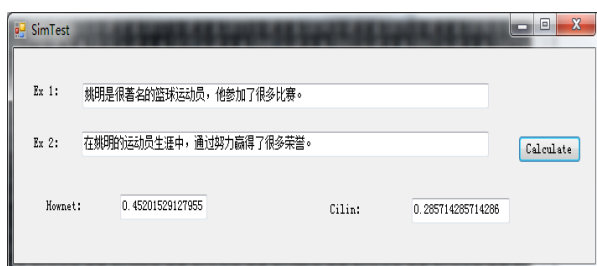


Figure 2. Algorithm Principle of This Chapter.

From the experimental result, it can be inferred, the result which is obtained by using of HowNet, which is often greater than the result that is obtained by using of Cilin and corpus. Textual relevance principle by using of HowNet mainly takes into account the relationship between the words' concepts, it weakness is without taking into account the context of the text, and some words have no relationship, but they are fixed collocation. The Cilin and corpus, which are the relationship between the words and expressions, are found through extensive texts. So using HowNet, is according the word itself meaning to calculate text correlation, and the use of Cilin and corpus is through inductive statistical method to get the relationship between the words, and then calculate the text correlation. But if we use HowNet to calculate the text relevance, there will be a defect, such as: For the two meaning of the text is

completely different, if the words of the two text is very similar, then it can get a high correlation between them.

Acknowledgement

The paper is supported by the National key Science & Technology Pillar Program of China (2015BAK05B03).

References

1. Dong Z, Dong Q. HowNet – a hybrid language and knowledge resource[C]. International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003:820-824.
2. Guan Y, Wang X L, Kong X Y, et al. Quantifying semantic similarity of Chinese words from HowNet[C]. International Conference on Machine Learning and Cybernetics, 2002. Proceedings. 2002 :234-239 vol. 1.
3. L Liang C, Shao Y, Zhao J. Construction of a Chinese Semantic Dictionary by Integrating Two Heterogeneous Dictionaries: TongYiCi Cilin and HowNet[C]. Ieee/wic/acm International Joint Conferences on Web Intelligence. IEEE Computer Society, 2013:203-207.
4. Zhang W, Yoshida T, Tang X. TFIDF, LSI and multi-word in information retrieval and text categorization[C]. IEEE International Conference on Systems. IEEE, 2008:108-113.
5. Li T, Yang X, Hong Q, et al. A hybrid method for syntactic and semantic structure disambiguation for Chinese[J]. 2001, 2:847-852 vol. 2.
6. Dai L, Liu B, Xia Y, et al. Measuring Semantic Similarity between Words Using HowNet[C]. International Conference on Computer Science and Information Technology. 2008:601-605.