Research and Experiment on Semantic-based Retrieval Methods of Idioms

QuQin[1], XunEndong[2], YuDong[3]
Institute of Information Science, Beijing Language and Culture University, Beijing 100083
E-mail:1,quqinjiayou@163.com 2,edxun@126.com 3, yudong_blcu@126.com

Abstract: Semantics-based idioms search is proposed to search the desired idiom when users only know what they want to express. The main missions of this paper are: (1) Selecting idiom sub corpus with size of 5.86G from 282.8G (about 37 million words) weibo and blog text corpus. (2) Starting BCC search engine based on the idiom corpus, extracted keywords by syntax analysis on the query request in the form of natural language, and searches alternative idiom set using the keywords sequence. (3) Training word embedding model to calculate the semantic similarity between the user's request and the alternative idioms, and then sort it from higher to lower so that we can get the idioms which meet user's requests best. The test results indicates that the semantics-based idioms search realizes the function of searching words through semantic and it has a preliminary effect in meeting user's searching requirements.
Keywords: Idioms, Semantic search, Word Embedding, Semantic similarity

1. Introduction

Chinese vocabulary is the sum of all the words and fixed phrases in Chinese [1, 6]. The existing words (words or phrases) inquiry systems can search the use of words, definitions and other related information by morphology (written) and the pronunciation (Pinyin). However, there is often such a situation: the user wants to express a point, but cannot think of the right words. At this time, the existing systems based on word form or pronunciation, or a combination of these two methods are unable to meet this demand, semantic-based system is required!

Keyword-based query can express simple semantic information by keywords set of the input query, but it's just a string match process between the user input and the idioms, it will check out a lot of useless information and in most cases it's difficult for users to express their query requirements simply through a few keywords. Therefore, to achieve semantic query, keyword-based methods are not enough. In response to this situation, taking into account the richness and complexity of Chinese vocabulary, this paper select idiom, which is a language unit that is more abundant in meaning than a word and grammatical function equal to the word, as the study object to explore semantic-based query method.

2. Background

2.1 Existing idiom inquiry systems

All the existing query systems are searching the use, interpretation and other related information of idioms by spelling and pronunciation. For example, Baidu search engine uses keyword matching, preliminarily realize idiom reverse lookup, but it simply returns idioms contain a word or a phrase in user input, rather than idioms semantic related to user query (such as in Figure 2.1.1 "山峰的成语").

Figure 2.1.1 Baidu Result of "山峰的成语"

iCIBA Chinese idiom dictionary a free online dictionary service from Kingsoft. It collects nearly 10000 idioms, but cannot search idioms by idiom definition. WuYou online idiom dictionary includes 41843 idioms. Idioms can be searched through its definition, synonyms or antonyms on WuYou. But when searching by definition is selected, a maximum of four words will be sent to the search engine, besides, the semantic information of user query requests is not taken into consideration.

2.2 BLCU Corpus Center

This paper uses the BCC (Beijing Language and Culture University Corpus Center, BLCU Corpus Center) corpus system. BCC is a multi-domain and multi-language text retrieval system which is built by the Institute of big data and language educational technology of Beijing Language and Culture University. It supports strings and POS combining mode query. Statistical analysis and downloads are also supported for the query results.

BCC corpus supports generalization, fuzzy, multi modal retrieval, such as support for the use of "*" to achieve query generalization, "." can be used for word number generalization. Because of the realization of the generalization of character operation, BCC system can support long distance dependent language model retrieval, that is to say, it can search a complete sentence based on sentence pattern and POS. This is very beneficial to the study of Chinese syntactic structure [4, 5].

2.3 Word embedding

The first step to convert natural language understanding problem into machine learning problem is formalizing these symbolic forms. The most intuitive word representation method in Natural Language Processing (NLP) is One-hot representation which represents every word as a long vector. The dimension of the vector is the vocabulary size, the vast majority of the elements are 0, only one dimension's value is 1, this dimension represents the current word, such as:

"话筒" is represented as [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 …]

"麦克" is represented as [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 …]

But there is a serious problem, that is, "vocabulary gap" phenomenon: any two words are isolated. Therefore, people usually use Distributed Representation in deep learning. In this way, a word is represented by a low dimensional real vector (usually 50 to 100 dimensions), such as [0.792, and 0.177, and 0.107, 0.109 and 0.542...]. This representation method makes similar words have closer distance, for example, the distance between "麦克" and "话筒" will be far less than "麦克" and "天气".

Representation Distributed was first proposed by Hinton in 1986[1]. Its basic idea is map each word into a k - dimensional vector (generally k is the hyper parameter in model), then determine the semantic similarity between different words through their distance (such as cosine similarity, Euclidean distance).

Word2vec [2] is an efficient tool for characterizing words as real valued vectors open source by Google in mid-2013 using Distributed representation.Word2vec uses the idea of deep learning to simplify text

processing into vector operations in K-dimension vector space. Vector similarity can be used to represent semantic similarity. The word vector output by Word2vec can be used to do a lot of NLP related work, such as clustering, find synonyms, speech analysis and so on.

3. Research on semantic based retrieval of Idioms

In the learning and use of idioms, in many cases, users only know what they want to express, but do not know the specific idiom. This demand cannot be satisfied with the existing methods of retrieval. For example, in the exiting idiom search systems users choose to retrieve through idiom interpretation, and input "描写形容品格高尚的人的成语", "人的品质高尚" and "人品质好", they will either failed to retrieve the results, or retrieve three completely different outputs. However, the three inputs express the similar meaning, the search results should be similar. The main reason for this problem is that current retrieval systems just do string matches for user's input when search, and do not consider their semantic relationships neither nor semantic analysis for user's retrieval request. In this paper, we explore semantic based retrieval method for idioms.

Figure 3.1.1 is a semantic based retrieval model for idioms. In this model, users firstly input their search query in the form of natural language. The model will do syntax analysis on the query, obtains the keyword sequence, and then get an alternative set of idioms based on keyword sequence. Finally, according to the semantic similarity of preparing an anthology of scoring and sorting, the retrieval results. Finally, sort semantic similarity between the candidate idiom and the search query from higher to lower, so that we can get the idioms which satisfy users best.

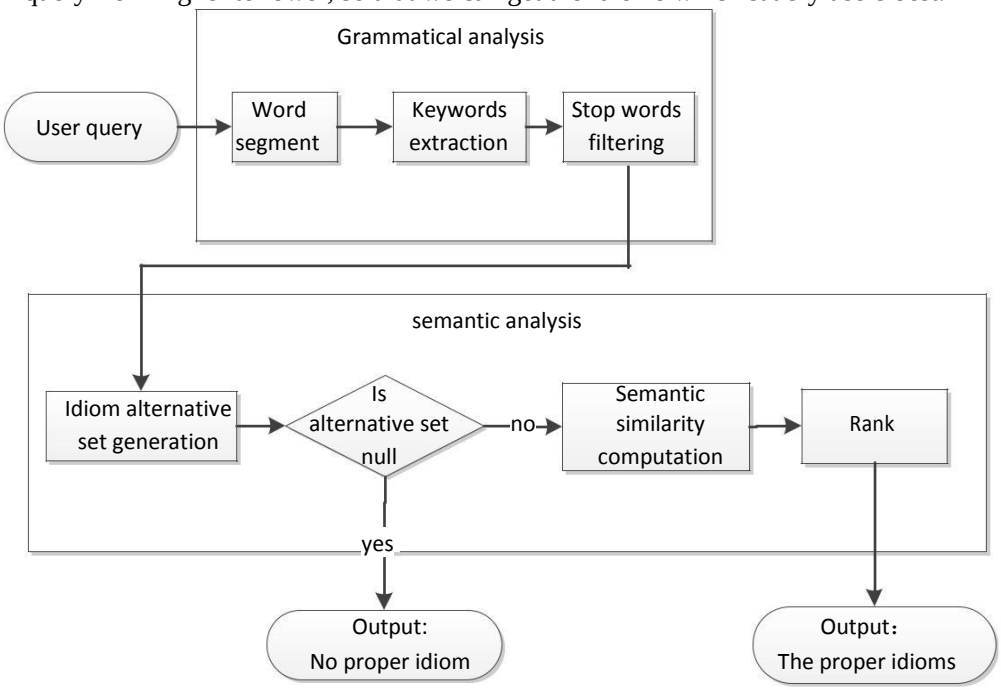

Figure 3.1.1 Semantic based retrieval model for Chinese Idioms

3.1 Keywords extraction

Users input their query request in natural language, to understand the semantic information expressed in these natural language, we should do syntax analysis on the user input in natural language first of all. Word is the smallest meaningful linguistic language component of independent activity. The first step of syntax analysis on user's input in natural language is word segmentation, and then extract the keywords which can express the meaning of the sentence.

In the parsing subsystem, word processing module does segmentation and POS tagging on queries entered by users. Keywords extraction module analyzes the syntactic structure of the whole sentence, extracts keywords based on POS tagging sequence and get the keywords set.

(1) Word segmentation

At present, the Chinese word segmentation and POS tagging algorithms have been well developed, so this paper uses the Chinese language processing package HanLp[1] to do word segmentation POS tagging instead of developing new algorithms. For example, for the sentence "描写形容品格高尚的人的成语", its POS tagging sequence is: 描写/v, 形容/v, 品格/n, 高尚/a, 的/ude1, 人/n, 的/ude1, 词语/n. V represents the verb, n represents the noun, a represents the adjective and ude1 represents the particle.

(2) Keywords extraction

Any sentence is made up of key elements (such as the subject, predicate, object, etc.) and modifying elements (attribute, adverbial, complement, etc.). The key ingredient plays a major role in the sentence and the sentence Modifier plays a secondary role, so extracting keywords only needs to consider the key elements in the sentence. Usually, the subject and object of the sentence are nouns and pronouns, the predicate are verbs and adjectives. Thus, according to linguistic knowledge, nouns, verbs and adjectives can be used as key words. For example, for the segment result "描写/v, 形容/v, 品格/n, 高尚/a, 的/ude1,人/n, 的/ude1, 词语/n ", its keywords set should be: [品格, 词语, 高尚, 描写, 形容]. The keywords extracted this way have a certain ability to express syntactic structure information, and can basically reflect the core idea of the sentence.

3.2 Stop word filtering

In information retrieval, to save storage space and improve search efficiency, people will automatically filter out certain words or word after or before natural language data processing. These words are known as Stop Words.

Previously extracted keywords like "描写" and "形容" etc. They are key components of the sentence .However, for this particular idiom retrieval problem, these words not only has no real meaning, but also will introduce a large number of irrelevant results. They should be filtered out, the similar stop words include: "描写，描述，表示，形容，相关，类似，成语，词语". For example, the key words collection [品格, 词语, 高尚, 描写, 形容] will be [品格,高尚] after filtering out all the stop words.

3.3 Scoring and ranking

To calculate the semantic similarity (vector distance) between query and candidate idioms, this paper makes full use of information like query request, idioms and corresponding idiom interpretation, forming a similarity calculation formula (Equation 1). Wherein, k represents the keywords set extracted from syntax analysis on user request, i denotes the keywords sequence obtained after word segmentation on alternative idiom interpretations, 2 * f is a bonus item represents an alternative idiom contains the word in the keyword sequence of the query request or not, if contains f will be 1, otherwise f will be 0.After scoring is completed, sor ing these scores o descending order, and the highest 20 idioms will fee back to user.

(1)

4. Experiment

4.1 Experiment data

The experiment data for this paper is an idiom sub corpus with size of 5.86G from 282.8G (about 37 million words) weibo and blog text corpus. After word segmentation and POS tagging process, it produces 11.8G text files which will be used for model training and system construction later.

All the idiom terms studied in this paper come from a large-scale Chinese dictionary, the Grand Chinese Idiom Dictionary. Most of the dictionary content comes directly from ancient literature it includes a total of 20366 ancient and modern Chinese idioms, containing pronunciation, interpretation and other information. The average number of these idioms amount to as high as 2284 times in the idiom corpus.

4.2Model training

This paper built a word2vec environment in Linux, and trained a 100-dimensional word vector model with window size of 5 using the 5.86G idiom corpus (as shown in Figure 4.2.1).

---

[1] http://hanlp.linrunsoft.com/

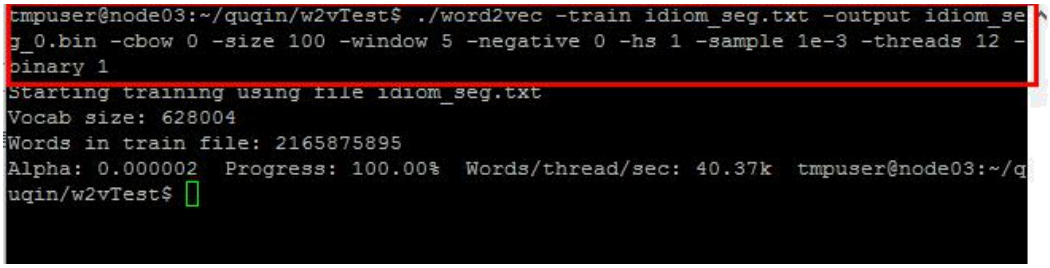

Figure 4.2.1 word2vec model training

4.3 Alternative idiom set generation

BCC query API supports strings and POS combining mode query. It also supports generalization, fuzzy, multi modal retrieval, such as support for the use of "*" to achieve query generalization, indicates that there are 0 or more other characters in the middle, "[]" is used to achieve an "or-category" query. For example, after obtaining the keyword sequence [品格,高尚] ,query request i*[品格 高尚] and [品格,高尚]*i will be sent to BCC. i*[品格 高尚] represents that there's an idiom before "品格" or "高尚", and [品格,高尚]*i represents that there's an idiom after "品格" or "高尚".We can get all idioms co-occurrence with "品格" and "高尚" by searching these two queries in BCC, that is to say, we get the alternative idiom set.

4.4 System construction

The semantic based query system in this paper is implemented using Java and PHP on Windows. Firstly, we search on BCC to get the query alternative idiom set, then use the word2vec model to calculate semantic similarity between user input query and each alternative idiom, sort these similarities in descending order, finally get idioms mostly satisfy user's query requirements. For example, enter the phrase "描写形容品格高尚的人的成语", the final result is as shown in figure 4.4.1.

4.5 Experiment evaluation

This paper aims at enabling users to search the proper idioms when they only know what want to express but don't know the specific idioms. All the test in this paper is carried out based on this demand.

Input "描写形容品格高尚的人的成语", "人的品质高尚" and "描写形容品格高尚的人的成语" to the semantic based retrieval system of this article, the retrieval result is shown in Figure 4.4.1 and 4.5.1:

Figure 4.4.1 Semantic based system search interface

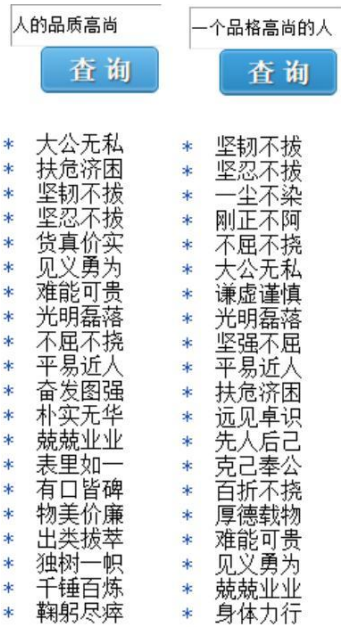

Figure 4.5.1 retrieval result

For the user input query request, this paper also selects several most commonly used existing idiom retrieval system to carry out the same search, through the comparison with the existing retrieval systems to

evaluate and analyze the results of the method proposed in this paper.

Since the three inputs are semantically similar, their retrieval results should also be semantically similar to each other. From the test results can be seen, the retrieval results are basically the same in our system, it can meet user's retrieval needs. Baidu search engine cannot recognize these special query requests, it returns nothing. ICIBA idiom dictionary cannot search by definition, it cannot meet this type of query needs. Wuyou idiom dictionary has results for these three search query, but it only takes the first four words of the search statement to match, if the first four words do not match any idiom, it will take the first two words to match (as shown in Figure 4.5.4). Thus, although there are search results, the three search results are totally different, and they cannot meet user's retrieval need semantically (as shown in Table 4.5.1).

Figure 4.5.4 Search result of "描写形容品格高尚的人的成语" of Wuyou idiom dictionary

Table 4.5.1 experiment results 1

	Retrieval Result			Semantic Analysis	Whether meet user's semantic retrieval requirements
	描写形容品格高尚的人的成语	人的品质高尚	一个品格高尚的人		
Our system	√	√	√	√	√
Baidu Search	×	×	×	×	×
ICIBA Idiom Dict	×	×	×	×	×
WuYou Dictionary	√	√	√	×	×

This article made a similar experiment 60 times, each time input the same query respectively to the four search systems, finally get the success rate which is the ratio of the number of trials satisfying user's retrieval need semantically and the number of all trials. The result is as shown in table 4.5.2.

Table 4.5.2 experiment results 2

	The number of experiments with retrieval results	Whether meet user's semantic retrieval requirements	Success rate （60 times total）
Our system	60	55	91.67%
Baidu Search	20	18	30%
ICIBA Idiom Dict	0	0	0
WuYou Dictionary	58	29	48.30%

From table 4.5.2, we can see that the retrieval success rate of our system is 91.67% which is obviously better than the existing systems. Baidu search engine can identify just a little of the query request, and it has no syntax analysis or parsing, so it can return a search result for just a few query as shown in table 4.5.2. As

shown in table 4.5.2, Wuyou idiom dictionary has result almost for every query, but a lot of its results cannot meet user's search request, and it also retrieves a large number of irrelevant idioms.

5. Conclusion

The purpose of this paper is to assist users to retrieve the proper idioms when they only know what they want to express but have no idea of the specific idiom. The results of above experiments show that the idiom retrieval system based on keywords is not enough to meet the user's demand to retrieve idioms semantically. The fundamental reason lies in, users need semantic based retrieval, but keyword based idiom search just returns idioms by string match without any semantic analysis to user's request.

In this paper, we use the word embedding model based on large data as the core of the idiom retrieval method. We build system to do word segmentation, POS tagging and syntax analysis on user's retrieval request. Besides, we calculate the semantic similarity between alternative idioms and search query which ensure that our system can satisfy users well. The next step will be to further improve the model, such as adding the idiom polarity calculation, etc.

References

[1] Hinton, Geoffrey E. Learning distributed representations of concepts. Proceedings of the eighth annual conference cognitive science society. 1986.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop. ICLR, 2013.

[3] ZangJiaojiao, XunEndong. The Study on Separable Words' Separable Forms of Modern Chinese. CLSW, 2015.

[4] 饶高琦，荀恩东.大数据视角下的语言实证工具：北语汉语语料库语料库系统 BCC.第十一届北京市语言学学会年会（北京）. 2014

[5] Mikolov T,Sutskever I,Chen K,et al.Distributed Representations of Words and Phrases and their Compositionality[C].Proceedings of NIPS,2013.

[6] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. Advances in neural information processing systems, 2009:1081~1088.

[7] Zhang Y, Dubrawski A, Schneider J.Learning the semantic correlation: An alternative way to gain from unlabeled text. Advances in Neural Information Processing Systems, 2008:1945~1952.

[8] 邵艳秋、穗志方、吴云芳．基于词汇语义特征的中文语义角色标注研究[J]．中文信息学报，2009.

[9] Turian Joseph, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), 2010.

[10] Eric Huang, Richard Socher, Christopher Manning and Andrew Ng. Improving word representations via global context and multiple word prototypes. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1,2012.