Research and Experiment on Semantic-based Retrieval Methods of Idioms

QuQin1, XunEndong2, YuDong3

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 …]

但是这种表示存在一个严重的问题，就是“词汇鸿沟”现象：任意两个词之间都是孤立的。因此，Deep Learning中一般用到的是Distributed Representation表示的一种低维实数向量（常为50到100维），如 [ 0.792, −0.177, −0.107, 0.109, −0.542, …]。这种表示让相似的词在距离上更近，例如，“麦克”和“话筒”的距离会远远小于“麦克”和“天气”。

But there is a serious problem, that is, "vocabulary gap" phenomenon: any two words are isolated. Therefore, deep learning in general use is distributed representation 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 allows similar words in the distance closer, for example, "Mike" and "microphone" distance will be far less than the "Mike" and "the weather" ".

Distributed representation 最早由Hinton在1986年提出[1]。其基本思想是通过训练将每个词映射成 K维实数向量（K一般为模型中的超参数），通过词之间的距离（比如 cosine相似度、欧氏距离等）来判断它们之间的语义相似度。

Word2vec[2]是Google在2013年年中开源的一款将词表征为实数值向量的高效工具，使用的便是使用的是 Distributed representation 的词向量表示方式，其利用深度学习的思想，把对文本内容的处理简化为 K 维向量空间中的向量运算，向量空间上的相似度可以用来表示文本语义上的相似度。Word2vec输出的词向量可以被用来做很多 NLP 相关的工作，比如聚类、找同义词、词性分析等等。

3. 基于语义的成语检索研究

在成语的学习和使用过程中，很多情况下，用户仅知道想要表达的意思，而不知道具体成语。这一需求现有的成语检索方式根本无法满足。例如，在现有的成语检索系统中用户选择通过成语释义来检索成语，分别输入“描写形容品格高尚的人的成语”、“人的品质高尚”和“人品质好”进行检索，要么检索不到结果，要么检索出来的结果完全不同，然而这三句话表达的意思是相同的，检索结果也应该相似。出现这一检索结果的原因主要是，现有成语检索系统对用户输入的查询请求进行字串匹配，并没有考虑其中的语义关系，没有对成语以及用户的检索请求进行语义分析。本文针对这一问题探索了基于语义的成语检索方法。 图 3.1.1 是基于语义的成语检索模型。用户首先以自然语言的形式提出查询请求，该型模对用户的查询请求进行语法分析，得到关键词序列，再根据关键词序列检索得到备选成绩集，最后根据语义相似度对备选集进行打分排序，得到检索结果。



图 3.1.1 基于语义的成语检索模型

3.1关键词提取

用户在进行成语检索时以自然语言形式输入查询语句，要对这些自然语言所表达的语义信息进行理解首先要对用户输入的自然语言进行语法分析。词是最小的能够独立活动的有意义的语言成分。语法分析首先就是要将用户输入的自然语言进行分词处理，然后再从中提取出能表达句子意义的关键词词组。

在语法分析子系统中，分词处理模块对用户输入的查询请求进行分词和词性标注，关键词提取模块对整个句子的句法结构进行分析，并结合分词处理后的词性标注提取出关键词，得到相应的关键词集合。

1. **分词处理**

目前的中文分词和词性标注的算法得到了很好的发展，因此本文没有再开发新的算法，而是使用了上海林原信息科技有限公司的汉语言处理包HanLp[[1]](#footnote-1)的分词处理系统进行分词处理和词性标注。例如，“描写形容品格高尚的人的成语”词性标注后结果为：描写/v, 形容/v, 品格/n, 高尚/a, 的/ude1, 人/n, 的/ude1, 词语/n。其中，v表示动词，n表示名词，a表示形容词，ude1表示助词。

1. **关键词提取**

任何句子都是由关键成分（主、谓、宾等）和修饰成分（定、状、补等）构成。关键成分对句子起主要作用，修饰成分对句子起次要作用，提取关键词只需要考虑句中的关键成分。通常情况下，句子的主语和宾语多为名词和代词，作谓语的多位动词和形容词。因此，根据语言学知识可将名词、动词、形容词作为关键词。如对分词结果：“描写/v, 形容/v, 品格/n, 高尚/a, 的/ude1, 人/n, 的/ude1, 词语/n”进行关键词提取，得到一个相应的关键词集合：[品格, 词语, 高尚, 描写, 形容]。这样提取的关键词具有一定的句法结构信息表达能力，基本上能反映出句子的核心思想。

3.2停用词过滤

在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为Stop Words（停用词）。

前文提取的关键词，描写、形容等是句子的关键成分，然而对于成语检索这一特定问题，这些词不仅没有实际意义，还会引进大量无关结果，应该过滤掉，本文中的类似停用词包括：描写，描述，表示，形容，相关，类似，成语，词语。比如关键词集合[品格, 词语, 高尚, 描写, 形容]过滤停用词后应该是[品格,高尚]。

3.3打分排序

为了计算查询请求和备选成语之间的语义相似度distance，本文充分利用了查询请求、成语以及成语对应的释义等信息，形成了相似度计算公式（公式1所示）。其中，k表示对用户请求进行语法分析得到的关键词集合，i表示备选成语释义分词处理后得到的关键词序列，2\*f是加分项，f表示备选成语是否包含查询请求关键词序列中的字，包含取值为1，反之，取值为0。打分完成后，根据得分由大到小排序，并取得分最高的20个备选成语反馈给用户。

(1)

4.实验

4.1实验数据

从281.8G微博、博客语料中筛选得到5.86G包含成语的子语料。经过分词、词性标注处理后得到11.8G文本文件，用于模型训练和系统建设。

成语词条方面，本系统中成语来源于《中国成语大辞典》，是一部大型语文辞书，内容大都直接取材于历代文献，共收录古今汉语成语 20366 条，包含成语读音释义等信息。而这些成语词条在本文所选语料中平均出现次数高达2284次。

4.2模型训练

本文在Linux环境下搭建了word2vec环境，利用5.86G成语语料训练窗口大小为5的100维词向量模型（如图4.2.1所示）。

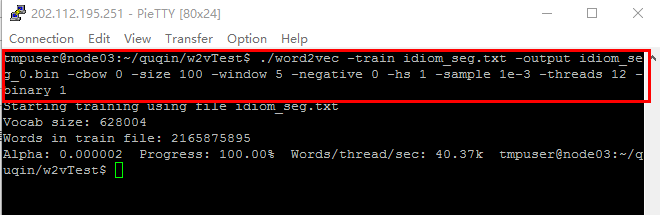


图4.2.1 word2vec模型训练

4.3成语备选集生成

BCC的查询API支持字串和词性组合的模式查询，也支持泛化、模糊、多模态检索，如“\*”号实现查询泛化，表示中间存在0个或多个其他字符，“[]”实现了括号内字或词的组合查询。例如，得到关键词序列[品格,高尚]后，可向BCC发起查询请求，i\*[品格 高尚]和[品格,高尚]\*i，两个请求分别表示品格或高尚前接成语和品格或高尚后接成语，这样就能在大规模语料中查询所有与“品格”和“高尚”共现的成语，得到备选集。

4.4系统建设

本文在Windows环境下基于php和Java，实现了基于语义的成语查询系统。首先，BCC检索得到用户查询备选集，然后用word2vec模型计算查询语句和各个备选成语的语义相似度，由大到小排序，得到满足用户查询要求的成语。例如，输入“描写形容品格高尚的人的成语”，最终检索结果如图4.4.1所示。

4.5实验评估方法

本文的目的是使用户在只表达意图而不知道具体成语的情况下能检索出满足需求的成语。本文的试验都是针对用户的这一需求进行的。 向本文检索系统分别输入“描写形容品格高尚的人的成语”、“人的品质高尚”和“一个品格高尚的人”。检索结果如图 4.4.1和4.5.1所示：

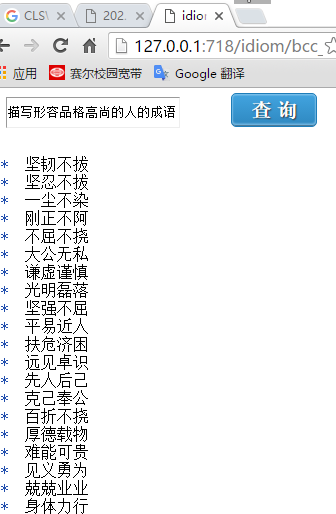
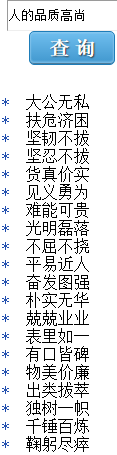
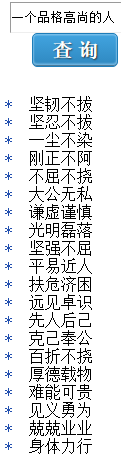
  

图4.4.1 系统检索界面 图4.5.1 本文系统检索结果

对于用户输入的检索语句，本文还选取了现有的成语检索系统进行了同样的检索，通过与现有检索系统所检索出结果的比较来评测和分析本文方法。

由于三个输入在语义上是相近的，所以其检索结果也应该是语义相近的，从试验结果可以看出，本文系统检索出的结果基本相同，从语义上分析能较好满足用户的检索需求。百度搜索不能识别这些特殊查询请求，没返回成语查询结果。ICIBA 成语词典，没有根据解释检索成语的功能，无法满足用户的这类检索需求。无忧成语词典，在输入以上三个检索语句后都有检索结果，但是它只取检索语句的前四个字进行匹配，如果前四个字没有匹配结果就取前两个字去匹配（如图 4.5.4），因此，虽然都有检索结果，但三个检索结果差异太大，并且从语义上分析其结果也不能满足用户的检索需求（具体如表4.5.1所示）。



图4.5.4 无忧在线成语词典“描写形容品格高尚的人的成语”检索结果

表4.5.1 试验结果 1



本文做了60次类似的试验，每一次都分别向四个系统输入相同的检索请求进行检索，最后得到的检索成功率是能从语义上满足用户检索需求的试验次数与总的试验次数的比率，结果如表4.5.2。由于篇幅有限在此不再一一展示所有的检索结果图。

表4.5.2 试验结果 2



从表 4.5.2 的试验结果可以看出，本文系统从语义上的检索成功率为 91.67%，效果明显的优于现有的成语检索系统。百度搜索能识别的成语查询请求较少、且没有语法分析，所以在表 4.5.2 中它检索出的能满足用户需求的成语数量都较少。表 4.5.2 显示，无忧在线成语词典几乎每次检索都有检索结果，但是它的检索结果很多都不能满足用户的检索请求，并且还会检索出大量无关的检索结果。

5.结语

本文系统的目的是，使用户在只有表达意图而不知道具体成语的情况下能检索出比较满意的结果。以上的试验结果表明，基于关键词的成语检索系统在满足用户通过语义检索成语这个需求上是远远不够的。其根本原因在于，用户的检索需求是语义上的检索，而基于关键词匹配的成语检索仅仅只是对字串的完全匹配而没有对用户的检索请求进行语义分析。本文使用基于大数据的词嵌入模型为核心的成语检索方法，搭建系统对用户的检索请求进行了分词、词性标注和语法分析，并对备选集进行了语义相似度分析，能较好的满足用户的这类检索需求。下一步将对模型进行进一步改善，如加入成语极性计算等。

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.

1. http://hanlp.linrunsoft.com/ [↑](#footnote-ref-1)