Research on Key Technologies of Knowledge Graph Construction Based on Natural Language Processing

Guilei Wang¹*, Yue Tao², Haixu Ma³, Tong Bao⁴, Jingmiao Yang⁵

¹-⁵School of Computer Science and Technology, Changchun University Of Science and Technology, Changchun, Jilin, 130022, China
*Corresponding author’s e-mail: 1318057273@qq.com

Abstract: As we all know, building a domain knowledge graph from a large amount of text requires a very large amount of work, including entity recognition, entity disambiguation, relationship extraction, and event extraction, etc. It is difficult to build a very comprehensive domain knowledge graph from scratch. Fortunately, with the rapid progress of natural language processing technology, we can use a large number of natural language processing tools to help us build a domain knowledge graph. This article mainly studies the extraction of domain terms in the process of constructing the knowledge graph. The natural language processing techniques used are mainly new word discovery, word segmentation, and keyword extraction. This paper improves the existing imperfect natural language processing technologies and applies them to the process of constructing the domain knowledge graph in order to construct the domain knowledge graph accurately and efficiently.

1. Introduction

As the application of knowledge graph in various aspects becomes more and more mature, how to construct a high-quality and wide-covered domain knowledge graph becomes more and more important. A lot of the original data of the knowledge graph comes from unstructured text data, and extracting the professional domain terms we need from the massive text data is the first job of knowledge graph construction, because the core concept of the domain is a subset of the domain term, so we can get the domain term first, and then get the core concept of the domain from the domain term.

With the advancement of natural language processing technology, we can use a large number of natural language processing tools to help us extract domain terms. The main work of this paper is as follows: First, in order to avoid the problem that the existing word segmentation system cannot correctly segment new words, we use the mutual information between points and adjacency entropy to discover new words. The dictionary tree is used to store the words and count the word frequency. Then add new words discovered by the new word discovery algorithm to the user-defined dictionary, which improves the word segmentation system; then we use the keyword extraction algorithm to extract the keywords of the article, and then extract the domain terms.

2. Reasons for using new word discovery technology and current research status of new word discovery

Many current word segmentation algorithms use word segmentation based on string matching. It matches a Chinese character string to be analyzed with entries in a "fully-large" machine dictionary according to a certain strategy. If a string is found in the dictionary, the match is successful (a word is recognized). The biggest drawback of this word segmentation method is that it cannot find new words,
and some new words are often the vocabulary of the domain itself. If these specialized vocabulary are not found by the word segmentation system, it’s not good for us to build a professional domain knowledge graph. How to improve it? Here we use the improved new word discovery algorithm to discover new words in the corpus, and then put the discovered new words into the user-defined dictionary of the word segmentation algorithm, which will increase the accuracy of the word segmentation. The current researches on new word discovery algorithms include: Chen Fei et al. used statistical features to determine the boundaries of new words, then added CRF models while synthesizing these features to find new words on Sogou's large-scale corpus [1]. Zhou Shuangshuang et al.[2] proposed a new method of Weibo new word discovery that combines rules and statistics, extracted new word formation rules of Weibo through classification and induction, reconstructed the NC-value objective function by improving statistics, and trained the CRF model to identify new words.

Our improvement measures mainly measure the degree of solidification of text fragments by calculating mutual information between points, measure the degree of freedom of text fragments by calculating left and right entropy, and use a dictionary tree to store words and count word frequencies.

In addition, since the word frequency counted from the document alone cannot reflect the rarity of a word in the entire language, that is, the frequency of a single word cannot be accurately reflected, therefore, using the word frequency table that comes with jieba as an external data source plays some auxiliary roles.

3. Key technology

3.1. Pointwise Mutual Information - Degree of solidification

Formula for pointwise mutual information

\[
PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}
\]  

Where \( p(x, y) \) represents the probability of two words appearing together, and \( p(x) \) and \( p(y) \) represent the probability of each word appearing.

From the formula above, we can conclude that the greater the mutual information between points, the more often these two words appear together, meaning that the greater the solidification of the two words, the greater the possibility of forming a new word.

3.2. Left and Right Entropy - Degree of freedom

The concept of entropy is divided into left entropy and right entropy. Left entropy refers to the entropy of the left boundary of a multi-word expression, and right entropy refers to the entropy of its right boundary. The specific calculation formula is as follows:

\[
E_L(W) = -\sum_{a \in A} P(aW | W) \times \log_2 P(aW | W)
\]
\[
E_R(W) = -\sum_{b \in B} P(Wb | W) \times \log_2 P(Wb | W)
\]

What is the role of left and right entropy?

Another criterion for a text fragment to be a word is that the left and right words of the word should be rich enough. That is, if a piece of text can form a word, it should appear in a rich context, that is, with rich left and right neighbors. The contextual entropy [3] can measure the uncertainty of the left and right neighboring words of a text segment. The greater the uncertainty, the richer the adjacent words, the higher the probability of the text segment forming a word. And we need to consider the left and right adjacent entropy, If the left adjacency entropy of a text fragment is large and the right adjacency entropy is small, or its left adjacency is small and its right adjacency is large, then this text fragment is also unlikely to be a word. Therefore, we define the degree of free use of a text segment as the smaller of its left-neighbor information entropy and right-neighbor information entropy. The smaller of the left-neighbor information entropy and the right-neighbor information entropy counts as a score, this score
plus the PMI’s score, also consider the word probability, and finally get a score. The higher the score, the more likely this text snippet becomes a new word, we will talk about this later in the algorithm.

3.3. Trie
Trie, also known as dictionary tree, word search tree, it is a multi-tree structure used for fast retrieval. The Trie tree has the advantage of minimizing useless string comparisons and higher query efficiency. So we also want to speed up the algorithm by introducing a dictionary tree.

4. Improved new word discovery algorithm
Based on the above theory, based on the original new word discovery algorithm, we use a dictionary tree to store words and count word frequencies, and propose a new word discovery algorithm:

(1) Use the existing word segmentation system to do rough segmentation of the text
(2) Use 3-gram to build nodes and use trie tree to store word segmentation
(3) Calculate PMI using trie tree
(4) Calculate left and right entropy using trie tree
(5) Calculate the score, score = PMI + min (left entropy, right entropy), then score * word probability, and finally get a score. Among them, min (left entropy, right entropy), the theoretical basis of this step is that the smaller value of left and right entropy represents the degree of free use of a text segment.
(6) Sort by the score, and take out the first N new words (N can be set by yourself)

Some algorithms will use this rule to delete a part of the candidate words: the suffix of the newly appeared word, in the prefix of the old word, delete the newly appeared word according to this rule. This is also bad for us to build a domain knowledge graph. For example, the situation like [forest_area, random_forest]: we all know that forest area and random forest are completely two concepts in two fields. According to the above rules, the word random forest will be deleted during a new word recognition program run. And will delete a lot of new words according to the above rules, so we have made improvements to remove the filtering rules so that both new and old words in the above situation can be retained.

5. Overall process
To build a domain knowledge graph, you need to use many domain terms. In theory, important terms in the field need to meet two basic conditions:

1. Terms appear relatively frequently in domain-related documents.
2. Terms appear more frequently in domain-related documents than in common documents.

From the above we can see that the domain terms are very similar to the keywords in the document, so we can use keyword extraction technology to help us reduce the search scope of important terms[4].

Here we use the TextRank algorithm for keyword extraction.

TextRank algorithm is a graph-based ranking algorithm for text. This algorithm first performs word segmentation on the provided sentences. Keywords extraction using TextRank also requires word segmentation. We can apply the above improved word segmentation algorithm to TextRank keyword extraction.

The overall process of this article is like this:
6. Results and analysis

6.1. Example:
Suppose there are several words: random forest, Dutch Navy. Before using our proposed new word discovery algorithm, the jieba word segmentation program did not recognize these new words and would split them into two words. After using our proposed new word discovery algorithm, the word segmentation program can correctly identify the above words, and they will be segmented from the sentence as a word.

6.2. Evaluation method: The evaluation is performed by the following criteria: accuracy rate, recall rate, F value

\[ P = \frac{AN}{N} \times 100\% \]  \hspace{1cm} (4)

\[ R = \frac{AN}{M} \times 100\% \]  \hspace{1cm} (5)

\[ F = \frac{2PR}{P+R} \times 100\% \]  \hspace{1cm} (6)

Among them, AN represents the number of new words correctly recognized by the algorithm. N represents the total number of new words recognized. M represents the number of correct new words in the corpus[5].

We selected 10,000 manually labeled data, extracted new words from the corpus and performed manual verification. A total of 660 new words were used as the standard new vocabulary.

6.3. Comparative experiments
Two experiments were performed this time. The new word discovery algorithm used in experiment one is proposed in reference [5]. Experiment two is the new word discovery algorithm proposed in this paper.

| Table 1. Comparison of experimental results of new word discovery algorithms |
|---------------------------------|-------|-------|-------|
|                                | P     | R     | F     |
| experiment one                 | 83.66%| 80.65%| 82.13%|
| experiment two                 | 86.12%| 81.56%| 83.78%|
7. Conclusion

From the experimental results, we can see that the new word discovery algorithm in this paper can better recognize new words in the text. Then the new words can be added to the user-defined dictionary to get better word segmentation results (improve the accuracy of word segmentation). Then we use the keyword extraction algorithm to extract domain terms for the construction of the knowledge graph. Of course, this is far from enough for the construction of knowledge graphs. We also need to perform entity disambiguation and relationship extraction. In the follow-up work, we will continue to use natural language processing tools to build and improve the domain knowledge graph, and lay the foundation for subsequent tasks using the knowledge graph to make recommendations.

References

[1] Chen Fei, Liu Yiqun, Wei Chao, Zhang Yunliang, Zhang Min, Ma Shaoping. New Word Discovery in Open Field Based on Conditional Random Field Method [J]. Journal of Software, 2013, 24 (05): 1051-1060.

[2] Zhou Shuangshuang, Xu Jinan, Chen Yufeng, Zhang Yujie. A new method of microblog discovery based on rules and statistics [J]. Journal of Computer Applications, 2017, 37 (04): 1044-1050.

[3] Huang Jinfu, Powers D. Chinese word segmentation based on contextual entropy [J]. Diseases of the Colon & Rectum, 2000, 4(6): 402.

[4] Yang Yuji, Xu Bin, Hu Jiawei, Tong Meihan, Zhang Peng, Zheng Li. An accurate and efficient method for constructing a domain knowledge map [J]. Journal of Software, 2018, 29 (10): 2931-2947.

[5] Huang Wenming, Yang Liu Qingqing, Ren Chong. Discovery of new words in the field combining amount of information and deep learning [J]. Computer Engineering and Design, 2019, 40 (07): 1903-1907 + 1914.