Research on Domain Term Dictionary Construction
Based on Chinese Wikipedia

Yu-wen ZHANG¹,*, Bao-an LI¹, Xue-qiang LV¹, Ning SUN² and Jing-Jing TIAN²

¹Beijing Key Laboratory of Internet Culture and Digital Dissemination Research Beijing Information Science & Technology University Beijing 100101, China
²SAMR Defective Product Administrative Center China Institute of Standardization Beijing 100191, China

*Corresponding author

Keywords: Automatic domain-specific term extraction, Domain thesaurus construction, Wiki, New word discovery.

Abstract. Domain terms are words or phrases that represent concepts or relationships in a specific domain. It can represent the characteristics of corresponding domains. The automatic construction of domain-specific dictionary is an important task in natural language processing, which can be adopted in domain-specific ontology construction, vertical search, text classification, information retrieval, question answering system etc. In this paper, we propose a novel method for constructing domain term dictionary based on Chinese Wikipedia web resource and deep learning technology. We for first time explore to word representation by Word2vec model integrating Wikipedia link structure. Then we use word clustering algorithm and seed word extraction method to construct an original domain dictionary. Moreover, neural network method is applied to extend domain dictionary. In experiments, different methods were employed to extract the domain-specific terms and their performances were compared in automobile field, the results reveal the effectiveness of our method for construction of domain-specific dictionary.

Introduction

Domain dictionary is the set of terms or means of expressions, each element of a domain dictionary is called domain term. Domain terms have specific semantics and belong to a specific domain. The study of domain terms acquisition has important theoretical and practical significance, and its research results can be directly applied to many aspects of Natural Language Processing.

At present, the lack of semantic knowledge has become a significant obstacle to natural language application, especially in special field. A large number of studies have been made on automated dictionary construction based on NLP technology. However, issues due to complexity of natural language, for instance the ambiguous/synonym term problems, still remain on NLP [1], therefore, It is necessary for us to improve the efficiency of building domain dictionaries by using new technologies. In our research, we first extract the automobile domain corpus of the Chinese Wikipedia, train the word2vec model word vector with the web link structure, and extract the terms in the domain by word clustering. These terms are made up of the primary domain dictionary. Finally, we use neural network to identify domain terms in different types of domain corpus to expand domain dictionary.

The rest of this paper is organized as follows. In section 2, we introduce some researches on automated dictionary construction. In section 3, we describe the proposed method and show the result of our experiments in section 4. We draw a conclusion in section 5.

Related Work

The research of domain dictionary construction focus on the extraction of domain words, extracting the domain knowledge by different extraction algorithms and methods.
Some researchers construct a domain term dictionary with the ontology construction technology. Zhang Chunxia [2] proposed a hybrid extraction of domain concepts, learning extraction terms through upper and lower relationships, and an ontology driven description flow extraction method. Chen Yu [3] propose a method for new term expansion based on domain ontology. This method combines linguistic rules and statistical methods to get the infection degree from a word to a document. It optimizes the term candidate result, uses domain ontology to recognize concept in document and to calculate correlation degree between word, document and specific domain. In recent years, the method of link structure based on Web has become the focus of research. The main purpose of Web structure mining is to get useful information by analyzing the hyperlink structure of Web. Chen [4] proposed a way to automatically build domain dictionaries by analyzing hyperlink structures. Wiki mining has recently become a new research field. MasahiroIto [5] Proposed a dictionary construction method based on link co-occurrence analysis. Experimental results show that combining TFIDF with link co-occurrence can achieve the best performance. Kotaro Nakayama [1] proposed lfibf algorithm for computing the correlation between two nodes in Wikipedia. Yin Wenke [6] proposed a method of automatic construction of domain lexicon based on weighted undirected link structure graph clustering.

Methodology

We proposed an efficient domain dictionary construction methods based on word2vec and link structure in Chinese Wikipedia. In this section, we will take the automobile field as an example to describe methods and related technologies in detail.

Dictionary Building Model

**Wikipedia.** Wikipedia is one of the largest online encyclopedias, which takes the mechanism of online cooperating editing. It is high-quantity, wide-coverage, evolving and semi-structural, and has become a well corpus of semantic knowledge bases.

In Chinese Wikipedia, all kinds of links, subtitles and classifications included in the text and entries of the word can be regarded as a semi structured organization of semantic information. On the basis of automatically constructing the source selection of domain dictionaries, compared to the weakly structured web pages, the Wikipedia words are easier to handle, and the difficulty of mining is reduced, the higher quality of the words can guarantee the accuracy of the semantic. In addition, the evolution of Wikipedia over time can largely avoid lagging behind in domain dictionary updating and reduce maintenance cost.

**Word2vec.** Word2vec is an efficient tool for characterizing words as real numeric vectors, its open source released by Google in 2013. The model can translate the words into vectors by using a deep learning algorithm. It was widely used in the academic world because of its efficiency, convenience and support for word semantic similarity computing, and this article will use word2vec to compute word similarity and combine link structure to generate the word vector model in the field of automobile.

**Our model.** In our research, Wikipedia is used in two aspects. On the one hand, all the articles of Chinese Wikipedia are preprocessed, participle and simplified to form a large Chinese corpus as training data. On the other hand, the relevant web pages in the automotive field are extracted from Wikipedia to build an automotive field knowledge base.

Wikipedia has a very dense link structure. In order to make full use of Wikipedia's link structure, we constructed a domain link structure diagram based on the automotive Chinese Wikipedia.

The corpus is a set of concepts and the hyperlinks among them, it can be expressed by a directed graph \( G = (V,E) \), we use a lfibf algorithm proposed by Kotaro Nakayama [1] to calculate the two node correlation in our automotive corpus. In the process of training word2vec words vector, we try to train a word vector representation model that is suitable for the current task and make full use of Wikipedia link structure. For an anchor text \( w_t \) in Wikipedia, we extract the words corresponding to the most relevant words calculated by lfibf algorithm in the link structure as its relevant words. In our model, the aim of learning is to maximize the logarithmic likelihood function is shown in Eq.1.
\[ L = \sum_{w \in C} \log p(w_t | \text{Context}(w_t) + \text{Relevant}(w_t)) \]  

In this formula, \( w_t \) represents a hyperlink text in Wikipedia. \( \text{Context}(w_t) \) represents the context of \( W \) in the Wikipedia text, and \( \text{Relevant}(w_t) \) represents the node words that are most related to the current node in the Wikipedia link structure by lfi-df algorithm.

In the above work, we trained the word vector model incorporating Wikipedia link structure, and extracted the link structure of Chinese Wikipedia. As we all know, domain words in Wikipedia corpus mainly consist of two categories, one is hyperlink terminology, and each term corresponds to an article. The other is the common words in the article. Some entries have not been created yet. We use two methods to extraction domain terminology in the corpus, for the anchor text in the automotive domain corpus, we use pre-training vector to express the anchor text in Wikipedia, then use the CPMw (Clique Percolation Method with weights) algorithm to do word clustering. Finally, we use the clustering results and domain top words to construct domain dictionaries. For the domain words that are not created in the Wikipedia text, we use seed words extraction to extract domain vocabularies. Firstly, we select \( n \) seed words in the dictionary constructed by CPMw algorithm, and calculate the direct similarity between words and words based on trained word vectors. For every seed word \( s_i \), calculate the most similar word in the corpus to form a set \( C_i \). Then compare the terms of different sets of \( C_i \), and merge them into domain dictionary.

**The Expansion of the Domain Dictionary**

In previous processing, we chose a general word segmentation tool in the training process of word vector. Some incorrect segmentation results may lead to the fact that domain terms can't be extracted. There are still some omissions in the domain dictionaries built by Wikipedia. In the research, we regard the expansion of domain lexicon as a sequence labeling problem. We identify the domain terms in the new corpus through deep neural network. If domain terms are not included in domain dictionary, the terms are added to the domain dictionaries.

The Bi-directional LSTM effectively uses the context information of the domain text, excavates more hidden features, not only can extract terms, but also has a certain ability to discover new words.

In our research, we cut the corpus into word, then define a tagging system suitable for this task, and then use the basic domain dictionary to tag the domain terms in the corpus. In our tagging system, \('s'\) represents the head of the term, \('e'\) represents the end of the term, \('m'\) represents the middle of the term, and \('o'\) is not part of the term.. For a sequence \( X \) with a length of \( m \), suppose the result of annotation is \( [y_1, y_2, ..., y_m] \), \( y_i = s/e/m/o \). We use the trained Bi-directional LSTM model to identify the domain terms in the new corpus. If the term is not in the domain dictionary, it will be added to the domain dictionary.

**Experiment**

In this section, we describe the dataset and evaluation metrics in our method. To evaluate the effectiveness of our approach, we conducted several experiment in automobile field on the proposed algorithms and other algorithms for comparison and discuss the result.

**Corpora and Evaluation Metrics**

In our research, we use the web resources including automobile Chinese Wikipedia, x-car automobile knowledge base and automobile brand commentary corpus.

We employ the widely-used CP (Concept precision) \[5\] to measure the performance of all the methods users evaluated whether the associated terms presented by the system are relevant or not by ranking them into 3 levels (1: irrelevant, 2: Moderate, 3: Relevant). CP is defined by the Eq.2.

\[
CP = \frac{\text{Number of retrieved relevant concepts}}{\text{Number of total retrieved concepts}}
\]  

(2)
Construction of Domain dictionary

The first groups of experiments was conducted to construct a primary domain dictionary in the field of automobile. We presented six sets of associated terms extracted by different algorithms.

In the corpus of Chinese Wikipedia in the automotive field, we implemented the Yin’s paper [6] as the first sets of experiments. We implemented the tfibf algorithms from Kotaro’s paper [5] as the second sets of experiments. The algorithms takes full advantage of Wikipedia's link structure based on the idea of TFIDF.

The last four experiments are the methods proposed in this article. First, we use Wikipedia data to train word2vec model, and use this data to train Wikipedia data with link structure as contrast experiment. In the construction of domain dictionaries, seed words similarity extraction algorithm and CPMw word clustering algorithm are used respectively.

Since the first two algorithms are based on the link structure of Wikipedia, all the words in the domain dictionary come from Wikipedia's entries, and the two methods get 427 and 536 domain words respectively. The last four algorithms are based on word2vec model, we choose 1000 words with the highest score to form a domain dictionary.

We let the user evaluate the relevancy of the terms by 3 levels. Totally 10 people participated in the experiment and terms in domain dictionary are evaluated. Top-n means the concept precision on average for the top n terms in the ranking (sorted by the scores of different algorithms). Table 1 shows the results.

Table 1. CP Comparison for Different Algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Top 25</th>
<th>Top 50</th>
<th>Top 100</th>
<th>Top 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSI+CPMw</td>
<td>95.0%</td>
<td>88.0%</td>
<td>76.0%</td>
<td></td>
</tr>
<tr>
<td>Lfibt</td>
<td>80.0%</td>
<td>74.0%</td>
<td>62.0%</td>
<td></td>
</tr>
<tr>
<td>W2V+CPMw</td>
<td>92.0%</td>
<td>87.0%</td>
<td>79.0%</td>
<td>71.3%</td>
</tr>
<tr>
<td>W2V+Link+seed</td>
<td>96.0%</td>
<td>86.0%</td>
<td>85.0%</td>
<td>77.6%</td>
</tr>
<tr>
<td>W2V+Link+CPMw</td>
<td>96.0%</td>
<td>88.0%</td>
<td>82.0%</td>
<td>72.2%</td>
</tr>
<tr>
<td>W2V++link+CPMw+seed</td>
<td>96.0%</td>
<td>90.0%</td>
<td>82.0%</td>
<td>76.9%</td>
</tr>
</tbody>
</table>

In the Table 1, W2V represents the methods of Word2vec, Link refer to integration of connection structures in the re building dictionary process.

As can be seen from the table, the algorithms used in this paper are superior to other algorithms in TOP25 and Top50. The first two methods mainly use the link structure of Wikipedia. Because the method proposed in this paper uses the word2vec word vector of large-scale field corpus training, it can build a larger dictionary. Compared with the four experiments, we can see that the word vector trained by Wikipedia’s link structure is better. Because CPMw algorithm allows a node to belong to multiple clusters, the experimental results show that CPMw word clustering is better than seed word extraction algorithm. In the last experiment, we regard the CPMw algorithm result as seed words, the results shows this methods achieve best performance.

Expansion of Domain dictionary

Second groups of experiments was conducted to expanding the domain dictionary in the new corpus. First, we use the primary domain dictionary built in the first set of experiments to annotate the automobile related corpus. We use the above-mentioned corpus randomly to select 70% as training set and 30% as test set. In order to avoid the effect of inaccuracy of word segmentation on experimental results, this group of experiments was tagged with the smallest unit of word. And train the model in a large scale corpus in the automobile field. We use the model to identify the domain term in article of test corpus.

In the experiment, we randomly select 10 articles in different automotive domain corpus, then use the trained model to identify domain terms. Most of the terms identified are words in the Automotive Dictionary, and a small part of the term is new. We repeat the above process 10 times, and then take the average of statistical results. The following Table 2 is the averaged CP of the new words.
Table 2. Different algorithms averaged CP in different corpus.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Articles</th>
<th>New Words</th>
<th>Averaged CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikiperia-auto</td>
<td>10</td>
<td>6.7</td>
<td>88.1%</td>
</tr>
<tr>
<td>Xcar-auto-base</td>
<td>10</td>
<td>12.2</td>
<td>65.6%</td>
</tr>
<tr>
<td>Auto-comment</td>
<td>10</td>
<td>23.5</td>
<td>76.5%</td>
</tr>
</tbody>
</table>

We found that the first experiment identify less new words, because the domain dictionary built in our work is based on Chinese Wikipedia. The appearance of these words are mainly due to the Chinese word segmentation errors of word segment tools. The accuracy of the new words identified by Bi-directional LSTM is higher than others. The corpus of Xcar automobile base mainly contains some knowledge of automobile, and the number of new concept is more than that of Wikipedia. However, the accuracy of this word is relatively low. The third experiment is the comment data on the automobile news, which is quite different from the source of the domain dictionaries. So the result of this experiment is similar to that of the second experiment. In brief, our method to expand domain dictionary has certain effect.

Conclusion and Future Works
This paper first constructs a corpus of automobile domain based on Chinese Wikipedia, then proposes a word2vec model training method which combines the Wikipedia link structure. Then uses the CPMw to cluster the words based word2vec model, after that, the basic dictionary in the field of automobile is obtained. In this paper, we takes automobile field as an example to conduct different contrast experiments. Experiment results shows that the quality of domain dictionary constructed by this method is better than other algorithms.

In the extension of the domain dictionary, the basic word list is first tagged based on the word, and then the Bi-directional LSTM model is used to identify the domain terms. The experimental results show that this method can identify some terminology that is not in the domain dictionary.

Wikipedia is a very large scale Web-based encyclopedia. And it is a great value in knowledge extraction. In future work, we will exploit the deep learning technology and make full use of the characteristics of Wikipedia to solve more practical problems.

Acknowledgement
This project was supported by the Funding Project for Natural Science Foundation of China (Grant No. 61671070) and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (Grant No.ICDD201708).

References