A Method of Constructing Feature Lexicon Based on Word Level

You Yu, Yu Fu

Department of Information Security, Naval University of Engineering, Wuhan 430033, China

ABSTRACT. *In view of the complexity of text categorization and search in the era of big data, Based on the diversity of Chinese words, and the task of constructing feature lexicon in text classification and searching, this paper designs a feature lexicon method based on word level. By learning the existing samples and identifying new words using CRF model, discriminating the importance of the words, reasonably dividing the word level and assigning weights, constructing an efficient and accurate feature lexicon, this method could obtain stable word segmentation effects, and effectively improve the accuracy of subsequent classification.*

KEYWORDS: *feature lexicon, word level, word segmentation, CRF*

1. Introduction

Compared with English, there is no strict delimiter between Chinese words, increasing the difficulty of Chinese word segmentation. The main reason is that there is no unified and universal participle specification, namely feature lexicon. With the continuous development of computer networks and languages, the data in the open field is increasing, and some new hot words are constantly emerging, such as names, technical names, and even online terms. However, due to the limited of lexicon, it is impossible to add new words to the original lexicon in time, which increases the difficulty of word segmentation. For different Chinese text segmentation, the focus of the word segmentation is also different. For example, the focus of word segmentation should be between different subjects when classifying medical texts, such as internal medicine, surgery, ENT. Therefore, in order to improve the efficiency and accuracy of the word segmentation algorithm, a reasonable feature lexicon must be constructed.

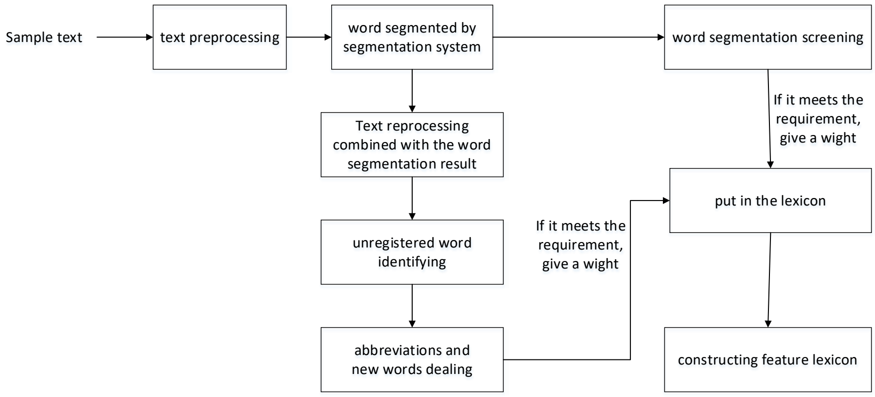
To construct the feature lexicon is to continuously filter, add, and delete the word in the feature lexicon. At present, many countries have developed semantic dictionaries for computers. The method of English texts include: WordNet, MindNet, and FrameNet [1]. The methods for Chinese text mainly include: ICTCLAS word segmentation system, IKAnalyzer Chinese word segmentation device, Paoding

Analysis, and there are some dictionary-based word segmentation systems, including Synonym Lin (Extended) [2], Modern Chinese Semantic Dictionary, Modern Chinese Semantic Dictionary, and HowNet [3]. The above methods all implement word segmentation through the method of embedding the lexicon in the system, but the ability to analyze and identify ambiguous words and unregistered words is insufficient. There are also many scholars who propose some methods for word segmentation directly. Xue proposes a word segmentation method based on HMM model [4]. Reference [5] proposed a method for automatic Chinese word segmentation using coverage ambiguity detection and statistical language models, and used multiple iterations to train Chinese word level statistical models. Liu Qun proposed a Chinese word segmentation method based on cascading hidden Markov model [6]. Liu Wenpeng proposed a Chinese word segmentation method based on the thesaurus and Bayes' theorem. And calculated the word segmentation scheme according to probability [7]. Reference [8] proposed a segmentation method based on BiLSTM by learning Chinese semantic vectors from a large number of corpora, and applying word vectors to word segmentation. The literature [9] proposed a segmentation method based on CRF and optimized the tag selection and feature template; although the above method improves the accuracy of word segmentation, it lacks pertinence and the calculation amount is large. Qian Tao et al. proposed a word segmentation model based on migration and text normalization to achieve text normalization by augmenting migration behavior [10]. Reference [11] proposed the domain adaptive method of word segmentation, which effectively improved the recognition of cross-domain words, but nit work on word segmentation in the same field.

Based on the diversity of Chinese text classification words segmentation and the construct tasks for feature lexicon in text classification and search operations, this paper proposes a word-based feature lexicon construction method by learning the text, identifying the words and giving them reasonable feature weights according to their characteristics.

2. A method of feature lexicon construction based on word level

As a basis for word segmentation, the feature lexicon must be comprehensive and streamlined to effectively improve the accuracy and efficiency of subsequent operations. For different requirements of word operations, such as part-of-speech classification, subject classification, grade identification, digital recognition, etc., the requirements of the feature lexicon are also different, and the importance of different features in the lexicon is also different. This paper proposes a method of constructing feature lexicon based on word level, which is divided into three levels according to the importance of different words in subsequent operations: level 1 - the most important level, words that can directly reflect the characteristics of the text Level 2 - general level, words related to the text topic; level 3 - general words, words that have little effect on the text and cannot be removed by screening. The operation is subsequent performed at the word level, which reduces confusion errors caused by a large number of word redundancy, thereby improving the accuracy and efficiency in subsequent operations. The specific operation process is shown in the figure below.


Sample text
text preprocessing
word segmented by segmentation system
word segmentation screening
If it meets the requirement, give a wight
Text reprocessing combined with the word segmentation result
put in the lexicon
If it meets the requirement, give a wight
unregistered word identifying
abbreviations and new words dealing
constructing feature lexicon

Finger. 1 The flow chart of constructing the feature lexicon

Step 1: Text preprocessing. Filtering useless words before text segmentation can reduce the dimensionality of text features, reduce unnecessary operations, and improve efficiency.

Step 2: Use the word segmentation system for word segmentation. Constructing the feature lexicon is the process of adding feature words to the lexicon. Obtain the word segmentation results by using the existing word segmentation system to perform preliminary word segmentation on the preprocessed text.

Step 3: Statistics and screening of word segmentation results. First, recall the words that are misclassified in the word segmentation, for example, “information security” will be mistakenly divided into two words: “information” and “security”. The recall can effectively improve the accuracy of the thesaurus.

Step 4: Identify the unregistered words. After reprocessing the processed text, identify unregistered words in the remaining text, mainly including the abbreviation recognition and the new word recognition.

Step 5: Replace the abbreviations identified in Step4 with their native words, and count the frequency of all the words and the remaining strings.

Step 6: Classify the word level in the feature lexicon. According to the different lexicon construction requirements, the word-level division requirements are also different. The words are classified according to their importance.

The above is the specific steps of constructing the feature lexicon based on the word level. Among them, the unregistered word recognition is still a difficult problem in Chinese text processing and how to achieve word-level division accurately is also a key point which is directly related to the final operation effect.

3. Unregistered word identifying

Unregistered word recognition is an essential part of Chinese text processing, and usually includes named entity recognition and other new word recognition. The unregistered words that need to be identified in this paper mainly include two parts: abbreviated word and new word.

3.1 Abbreviation recognition

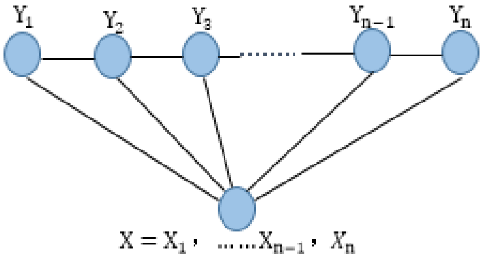
For abbreviations in the normative text, there are some rules, which can be obtained by training the related texts, and then get the correspondence between the standard words and the abbreviations. Some of its rules are as follows: Conditional random field (CRF) method. We can know that “CRF” is the same word as “Conditional random field”.

In the process of recognizing abbreviations through automatic text learning, we should add manual intervention to assist the supervision, and count replacement probability for each original standard words and abbreviations, and then determine whether the abbreviations can be replaced by the original words, which can make the replacement probability is more accurately.

3.2 New word recognition

At present, there are two main methods for identifying new words: statistical-based methods and rule-based methods [12]. The statistical-based method is mainly judged by the probability of the occurrence of statistical words and other relevant features. The rule-based method is mainly based on the rules which set by relevant knowledge, mainly used to obtain technical terms. In this paper, a discriminant probability model, the conditional random field (CRF) model [8], is used to identify new words, which combines the characteristics of the maximum entropy model and the hidden Markov model. It is an undirected graph model and can get the global optimal solution.

In the definition of CRF, there is no identical requirement for the structure of variables X and Y. In reality, X and Y usually have the same structure. Therefore, a special form, the linear chain CRF model the CRF model is used in this paper. Its structure is as followed:



Finger. 2 The linear chain CRF model

In linear CRF, the feature functions are divided into two categories, one is defined on the node Y that characteristic functions are only related to the current node, and the other is the local characteristic function defined in the Y context, which is not only related to the current point but also related to the previous node. In general, the processing of text needs to consider the connection between contexts, so the conditional probability of a state sequence y can be expressed as:

$$\mathrm{P}\left(y \middle| x, \omega\right) = \frac{1}{Z(x)} \cdot \exp(\sum_{i=1}^{N}\sum_{j}\omega_j \cdot f_j(y_{i-1}, y_i, x, i)) \tag{1}$$

Where i is position information; y_j (y_i-1), x_i, x, i) is a characteristic function, indicating state transfer information of y on i, and j is a label of the characteristic function; Z(x) is a normalization factor, for:

$$\mathrm{Z}\left(x\right) = \sum_{y}\exp(\sum_{i=1}^{N}\sum_{j}\omega_j \cdot f_j(y_{i-1}, y_i, x, i)) \tag{2}$$

Finally, the optimal sequence of y under given x conditions can be calculated, which can be determined by the Viterbi algorithm:

$$y = \arg\max P(y \mid x, \omega) \tag{3}$$

In addition to the basic characteristic such as length, the features used in CRF learning are as follows:

(1) TF-IDF of words

$$\mathrm{TF - IDF = TF{*}IDF} \tag{4}$$

(2) Left information entropy of words

$$\mathrm{H\_L(i)} = -\frac{1}{n}\sum_{a\epsilon A}C(a, i)log\frac{C(a, i)}{\mathrm{n}} \tag{5}$$

Where A represents the set of words located to the left of the word i, and C(a, i) is the number of occurrence of a and i.

(3) Right information entropy of words

$$\mathrm{H\ \_\ R(i)} = -\frac{1}{n}\sum_{b\epsilon B} C(b,\, i) log\, \frac{C(b,\, i)}{n} \tag{6}$$

Where B represents the set of words located to the right of word i, and $\mathrm{C(b,i)}$ is the number of occurrence b and i.

4. Word level division method

According to different lexicon construction requirements, the different features in the thesaurus are different, and the requirements for word-level division are also different. In order to reduce the confusion caused by a large number of word redundancy and improve the accuracy and efficiency in subsequent operations, this paper proposes a word-based feature lexicon construction method, which divide the words into three levels according to the importance of words in the subsequent: the words that best reflect the characteristics of the text are divided into level 1 words, that is, the important level; the words that can reflect the text subject are divided into level 2 words, that is, the general level; the remaining words are divided into level 3 words, that is, ordinary level.

After the word segmentation of the training text which category is known, count the frequency of occurrence of each word, and analyze its position in the text, and then give a weight, as follows:

```
$\boldsymbol{while\ count(i) \neq 0 \quad do}$:
$\boldsymbol{P_{i\_high} = \max\{p_{ij}\}}$;
$\boldsymbol{P_{i\_low} = \min\{p_{ij}\}}$;
$\boldsymbol{if\ P_{i\_low} < TL\ or\ \ P_{i\_high} > TH\ , delete\ i}$;
$\boldsymbol{for\ i\ , if\ \frac{n_{ij}}{\sum_j nij} = 1\ , i \to vital\ , \pi\_z_i = 1\ , remove\ i}$;
$\boldsymbol{for\ i\ , if\ \frac{N_{ij}}{num_j} \gg \sum_{k\neq j} \frac{N_{ik}}{num_k}\ ,\ \ i \to minor\ , \pi\_z_i = \frac{N_{ij}}{num_j} \cdot \left(1 - \sum_{k\neq j} \frac{N_{ik}}{num_k}\right) , remove\ i}$;
$\boldsymbol{for\ i\ , i \to general\ ,\ \ \pi\_z_i = \min\{P_{i\_low}\ , 1 - P_{i\_high}\}\ , remove\ i}$;
$\boldsymbol{end}$;
```

Where $\mathrm{P_{i\_low}}$ is the minimum frequency at which the word i appears in the class j text, TL and TH are threshold values, and $\mathrm{n_{ij}}$ is the frequency at which the word i appears in the class j text, π_z_i is the absolute weight of word i, $\mathrm{N_{ij}}$ is the number of texts which words i appears in the class j text, $\mathrm{num_j}$ is the number of class j text, and count (i) is the number of segmented words. Repeat the above steps until all the words have been processed, and finally normalize the weights of the words:

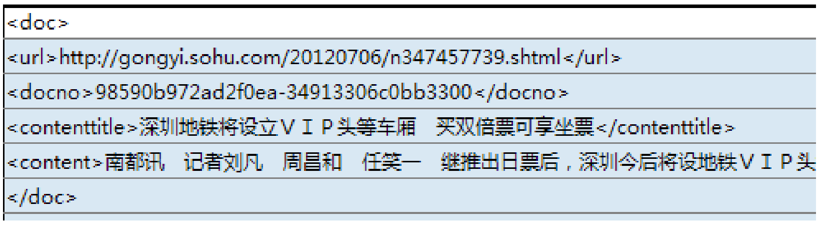
$$\pi_i = \pi\_z_i \cdot \lambda \tag{7}$$

Where λ is the normalization factor.

5. Experiment

The corpus used in this paper is from Sogou CS (SogouCS) [13], which is from the news data of Sohu News from June to July 2012 including domestic, international, sports, social, entertainment and other 18 channels. The data contained are: URL, title, body content. The news was selected as experimental data from the five categories of finance, military, sports, entertainment. They were randomly selected according to the ratio of 8:2, and the former is used for learning, the latter is used for testing.

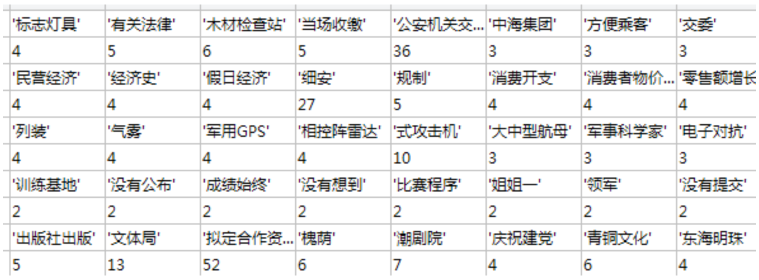
The downloaded Sohu news data is a .dat package which size is 1.43GB. Firstly, the data is divided and classified, and it is divided into single news according to the .dat data format, and it reads the type of the news from its url and classifies it. In the operation, we can find that the provided data contains a small amount of noise data (obsolete news) and we should filter it. After completing the above operations, selects150 news items from financial, military, sports, entertainment and society for experiment.



```
<doc>
<url>http://gongyi.sohu.com/20120706/n347457739.shtml</url>
<docno>98590b972ad2f0ea-34913306c0bb3300</docno>
<contenttitle>深圳地铁将设立ＶＩＰ头等车厢　买双倍票可享坐票</contenttitle>
<content>南都讯　记者刘凡　周昌和　任笑一　继推出日票后，深圳今后将设地铁ＶＩＰ头
</doc>
```

Finger. 3 News data format

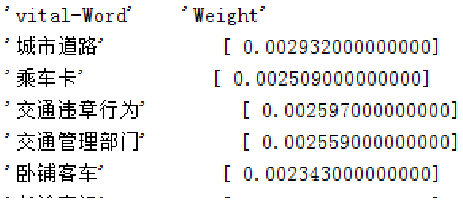
Select the text content in the news and preprocess the text. The usual method is to remove all punctuation and stop words which usually do not carry any useful information, such as auxiliary words and "because", "so" and other words that reflect the structure of the sentence. And then remove some of the mark information, mainly for web page text and other markup language text. Use the ICTCLAS word segmentation system to do preliminary word segmentation processing, and use the method that is proposed in this paper to identify the new words in the remaining texts. The partial new word recognition results are shown in the following figure:



'标志灯具'	'有关法律'	'木材检查站'	'当场收缴'	'公安机关交...	'中海集团'	'方便乘客'	'交委'
4	5	6	5	36	3	3	3
'民营经济'	'经济史'	'假日经济'	'细安'	'规制'	'消费开支'	'消费者物价...	'零售额增长
4	4	4	27	5	4	4	4
'列装'	'气雾'	'军用GPS'	'相控阵雷达'	'式攻击机'	'大中型航母'	'军事科学家'	'电子对抗'
4	4	4	4	10	3	3	3
'训练基地'	'没有公布'	'成绩始终'	'没有想到'	'比赛程序'	'姐姐一'	'领军'	'没有提交'
2	2	2	2	2	2	2	2
'出版社出版'	'文体局'	'拟定合作资...	'槐荫'	'潮剧院'	'庆祝建党'	'青铜文化'	'东海明珠'
5	13	52	6	7	4	6	4

Finger. 4 The statistical results of new words

And then divide the words into there levels, and a total of 408 level 1 words, 514 level 2 words, and 791 level 3 words are obtained. The hierarchical lexicon is as shown in the following figure:



```
'vital-Word'    'Weight'
'城市道路'            [ 0.002932000000000]
'乘车卡'             [ 0.002509000000000]
'交通违章行为'           [ 0.002597000000000]
'交通管理部门'           [ 0.002559000000000]
'卧铺客车'            [ 0.002343000000000]
```

Finger. 5 Words and its weight in the lexicon

Select the text content in the test news data, and use the above word segmentation method to segment the test data text. At the same time, directly use the ICTCLAS word segmentation system and CRF method to process the word segmentation. The word weight is given according to the frequency of occurrence of the word, and do category classification experiments for this news. Comparing several methods, the specific results are shown in the following table:

Table 1 Comparison of experimental results

Method	ICTALAS	CRF	Feature lexicon based on word level
Precision/%	93.3	95.5	96.5
Recall/%	92.2	95.2	96.7
F-Measure	0.927	0.953	0.966

According to the experimental results, it can be found that the feature lexicon based on word level has the best classification effect. In terms of the number of word segmentation, the word lexicon based on word level construction has the least words, because a large number of new words are recognized and the abbreviations are replaced by their original word. The experiment proves that the method of constructing feature lexicon based on word level can improve the accuracy of subsequent text operations.

6. Conclusion

Based on the diversity of Chinese words and the task of constructing the feature lexicon in text classification and search operations, this paper proposes a feature lexicon construction method based on word level. This method reduces the complexity compared with other methods, and it is good for the classification, searching and comparison of subsequent texts, and determines the different contributions of different words by dividing the word level, avoiding the existence of a large number of redundant words in the text which affect the subsequent processing results. This method can improve the accuracy and efficiency of subsequent operations.

References

[1] Li Hui (2015). A Review on the Research of Word Similarity Algorithm. Modern Information, vol.35, no.4, p.172-177.

[2] Xiong Huixiang, Ye Jiaxin (2018). Research on Hierarchical Structure Construction of Socialized Tags Based on TongYiCiCiLin. Journal of Information, vol.37, no.1, p.126-131.

[3] Dong Zhendong, Dong Qiang. HowNet [DB/OL].[2012-9-11]. http://www.keenage.com.

[4] XUE N (2003). Chinese Word Segmentation as Character Tagging. Computational Linguistics Chinese Language Pro-cessing, vol.8, no.1, p. 29-48.

[5] Wang Xianfang, Du Limin (2003). Using Chinese Coverage Ambiguity Detection and Statistical Language Model for Automatic Chinese Word Segmentation.Journal of Electronics & Information Technology, no.9, p.1168-1173.

[6] Liu Qun, Zhang Huaping, Yu Hongkui, et al (2004). Chinese Lexical Analysis Using Hhidden Markov Model. Computer Research and Development, vol.41, no.8, p.1421-1429.

[7] Liu Wenpeng (2012). Research on Chinese Word Segmentation Based on the Thesaurus and Bayes' theorem. Huazhong University of Science and Technology.

[8] Zhang Honggang, Li Huan (2017). Chinese Word Segmentation Method Based on Two-way Llong-term Memory model.Journal of South China University of Technology(Natural Science Edition),vol.45, no.3, p.61-67.

[9] Liu Zewen, Ding Dong, Li Chunwen (2015). Chinese Word Segmentation Method for Short Chinese Text Based on conditional random field. Journal of Tsinghua University(Science and Technology), vol. 55, no.8, p. 906-910+915.

[10] Qian Tao, Ji Donghong, Dai Wenhua (2015). A Model of Weibo Participle and Text Normalization Based on Migration. Journal of South China University of Technology (Natural Science Edition), vol.43, no.11, p.47-53.

[11] Han Dongxi, Chang Bing (2015). Approach to Domain Adaptive Chinese Segmentation Model. Chinese Journal of Computers, vol.38, no.7, p.272-281.

[12] Chen Shouqin (2017). Research on Chinese Short Text Unregistered Word Discovery and Sentiment Analysis Method. Beijing University of Technology.

[13] Sohu News Data [DB/OL]. [2012]. https://www.sogou.com/labs/resource/cs.php.