Research on domain terminology recognition based on dependency tree-conditional random field

Yanyan Lin, Jieping Lu

School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212000, China
945143660@qq.com

Abstract. In view of the inconsistency of Chinese patent information in manual marking and classification, which leads to problems such as missed detection, partial detection and noise of patent search, this paper proposes a method based on the dependency tree-conditional random field (CRF) identification field terminology. The method is based on the modern grammar theory of dependency, using the existing technology to mark the dependency relationship. Finally, the corresponding technical feature words are identified in the results of the dependency labelling, and the training data is used as the training data to train the conditional random field model to identify the domain terminology. The experimental results show that the acquisition of training data through the dependency tree can improve the accuracy, recall and F value of the recognition results.

1. Introduction
The technical feature words of a patent refer to technical terms with a specific domain meaning, which are used to describe the technical features and technical solutions of the patent. The traditional text-based keyword mining algorithm mainly uses the feature vector model, which first calculates the frequency of occurrence of words in the document set, and then selects words with frequencies greater than a certain threshold as keywords. The main drawback faced by the method of using keywords is that many high-frequency words are not actually technical characteristic words, thus ignoring a part of the low-frequency words that are actually technical characteristic words. The Conditional Random Fields (CRF) model based on the field of natural language processing [1] is a non-directional statistical graph model applied to pattern recognition and machine learning, used as a structured prediction. CRF has been widely used in the areas of part-of-speech tagging, shallow syntactic analysis, and named entity recognition for news/biological corpus. For example, FuYao et al. used the conditional random field algorithm to automatically identify existing concepts and discover new concepts in a specific field [2]; LiNa based on conditional random field algorithm to construct an automatic extraction model of local history aliases [3]. CRF combines the advantages of Hidden Markov Model (HMM) and Maximum Entropy Markov Model (MEMM) to normalize the global features, which can solve the labeling bias problem well and obtain the best discriminant value. However, the disadvantage is that a certain number of manually labeled technical feature words are required as training data for training the machine learning model. Therefore, this paper proposes the technical feature words needed to automatically acquire the training model based on the grammatical relationship for the Chinese patent text, improving the work efficiency and the accuracy of manual labeling.
2. Dependency tree

In computational linguistics, a dependency tree is a grammar tree generated by a dependency grammar. The law of dependency is a modern syntactic theory based on dependency. Dependency is based on semantic analysis, with more consideration of words and grammar. Language units are connected by directed edges, verbs are considered to be the center of the structure in the clause, and all other words are directly or indirectly connected to the verb through the directed edge. The structure of the dependency is determined by the relationship between a prefix and its dependants.

This paper uses the open source syntax dependency tree tool Stanford Parser [4] to extract the grammatical relations in Chinese sentences and generate a Chinese syntactic dependency tree. In the expression of the dependency tree, the grammatical relations in all sentences are represented by semantic triples, that is, two words and the relationship between them. For example: "The power supply for the drive motor can be a wired vehicle power supply, or it can be a self-contained high-energy battery (patent CN200310106427.1)". After Stanford Parser processing, the semantic triplet can be shown in Table 1:

| Table 1 defines the triples in the dependency tree |
|-----------------------------------------------|
| assmod(power-3, drive motor-1)               |
| attr(is-5, power-9)                          |
| assm(drive motor-1, of-2)                   |
| nsubj(is-5, can be -4)                       |
| mmod(is-5, power -4)                        |
| root(ROOT-0, is-5)                          |
| assamod(power-9, wired-6)                   |
| nn(power-9, car-8)                          |

Where each row represents a triple, such as the triplet "assmod (power-3, drive motor-1)" in the first row, assmod is the semantic relationship name, and the power supply and the drive motor are respectively for this semantic relationship. Head" and "dependent". The semantic triplet described above can be directly mapped to a dependency tree represented by Fig.1, where the words in the sentence act as nodes of the tree and the semantic relationships act as edges of the tree.

![Fig. 1 Dependency Tree (Patent CN200310106427.1)](image)

According to the semantic relationship in the 46 Chinese dependency trees defined by Chang et al. [5], combined with the ratio of various semantic relations in the Chinese corpus, this study selects six
semantic relations that are useful for extracting technical feature words, as shown in Table 2. Show.
The nominal technical feature words can be composed of various combinations of nn, dobj, and nsubj. After analyzing a large number of patent texts, it is concluded that the nominal technical feature words are usually composed of three modes: “nn+nn+...”, “nsubj+dobj”, “nn+nsubj”, and the attribute feature words are composed of “amod” and "top+attr" is composed of two modes.

| Semantic relations | Chinese description | Semantic relations | Chinese description |
|--------------------|---------------------|--------------------|---------------------|
| nn                 | compound noun modifier | nsubj             | noun subject        |
| dobj               | direct object        | amod               | adjective modifier  |
| top                | Subject              | attr               | attribute modifier  |

Table 2 Six semantic relationships defined in the dependency tree

The proposed semantic triple-based feature word mining method is applied to the abstract text of the patent specification (short length, language concise, technical feature words are obvious), and the implementation process mainly includes noise reduction on the imported patent text (such as extra blank) Characters, text entry errors, various types of numbers before the characters, etc.), word segmentation, and then use Stanford Parser for dependency labeling, and finally extract the corresponding technical feature words in the results of the dependency labeling.

3. Conditional random field

The conditional random field was first proposed by J. Lafferty et al. In recent years, it has been widely used in the fields of natural language processing and image processing. The definition of the conditional random field is described as follows: Let $G = \langle V, E \rangle$ denote an undirected graph in which the node corresponds one-to-one with the element in the marker variable $y$, $y_v$ denotes a marker variable corresponding to the node $v$, and $n(v)$ denotes a knot The neighboring node of point $v$, if each variable $y_v$ of graph $G$ satisfies Markov property,

$$
P(y_v|x, y_{\ell \setminus \{v\}}) = P(y_v|x, y_{n(v)}) \quad (1)$$

Then $(y, x)$ constitutes a conditional random field.

For Chinese patents, the idea of domain term recognition based on CRF algorithm is: If each sentence is treated as a sequence of characters, the process of automatically labeling the corresponding feature words in the clause is considered to automatically use the sequence labeling algorithm to automatically sequence the characters. The process of labeling, and the content marked is the technical feature word to be identified. Define the following set of tags to mark the feature word [6] to be extracted: {B, I, O}. Where B is the beginning of the feature word, I is the feature word, and O is not the feature word. These three tags are used to manually mark certain training data, and then train the sequence labeling algorithm model.

In addition, after selecting the feature set, the quality of the feature template will greatly affect the results of the experiment. Table 3 lists the feature template used to train the sequence labeling model. The template contains 20 features, F1-F20, among which Two or more feature combinations are connected by "/".

| Feature | Feature content | Feature | Feature content |
|---------|----------------|---------|----------------|
| F1      | target word    | F11     | F3/F1          |
| F2      | target word meaning | F12 | F1/F7 |
4. Conditional random field

4.1. Experimental data
This paper selects 600 invention patents (in the field of illumination) with IPC classification number F21 on the Chinese patent database service platform of China Intellectual Property Network. Among them, 480 patent information is used as training corpus, and the rest is used as test corpus. The specific corpus information is shown in Table 4.

| Corpus             | Training set | Test set       |
|--------------------|--------------|----------------|
| lighting           | 480 patent information | 120 patent information |
| patent information | about 82,000 words | about 20,000 words |

4.2. Evaluation method
This paper uses the three general evaluation indicators in the field of natural language processing, namely accuracy (P), recall rate (R) and F value, to evaluate the results of term recognition. The specific definition is as follows:

\[
\text{Accuracy}(P) = \frac{\text{number of domain terms correctly identified}}{\text{total number of domain terms identified}} \times 100\% \quad (2)
\]

\[
\text{Recall rate}(R) = \frac{\text{number of correctly identified domain terms}}{\text{total number of domain terms in corpus}} \times 100\% \quad (3)
\]

\[
F \text{ value} = \frac{(\beta^2+1) \times P \times R}{\beta^2 \times P + R} \quad (4)
\]

When comprehensively evaluating the recognition results, P and R should be considered at the same time, but it is difficult to see the two values at the same time. Therefore, it is usually evaluated by combining two values (F value), and \( \beta \) is the relative weight of recall rate and accuracy, generally take 1.

4.3. Experiment procedure
The experimental process is shown in Fig.2.
The domain terminology identification method based on dependency tree-CRF proposed in this paper: firstly use the open source tools such as Apache POI and Java API to read the patent data stored in excel format or XML format, and also extract the patent data of specific parts, such as the specification part. Then use the open source syntax dependency tree tool Stanford Parser to extract the grammatical relations in Chinese sentences, according to the specified semantic triples, the grammatical relations in all sentences and generate the Chinese syntactic dependency tree, as shown in Figure 1. Based on the artificial expert evaluation, the labeled technical feature word set is used as the training data of the CRF. The machine learning CRF model is trained by referring to the method of the literature [7] to mark the patent text of the test corpus, and the term candidate feature set of the test corpus is generated.

4.4. Experimental results and analysis

Two experiments were performed here, which are CRF model domain terminology recognition based on manually labeled training data and domain terminology recognition based on dependency tree-CRF. Each experiment was done 10 times, and the accuracy, recall and F value of the results were compared, as shown in Table 5.

Table 5 Data results of two experiments

| Number of experiments | Accuracy (%) | Recall rate (%) | F value (%) |
|-----------------------|--------------|-----------------|-------------|
|                       | Experiment 1 | Experiment 2    | Experiment 1 | Experiment 2 | Experiment 1 | Experiment 2 |
| 1                     | 91.45        | 94.11           | 85.65        | 89.94        | 84.84        | 85.6         |
| 2                     | 91.87        | 96.84           | 86.51        | 92.73        | 88.22        | 88.99        |
| 3                     | 92.01        | 95.04           | 86.34        | 92.84        | 84.63        | 84.96        |
| 4                     | 92.15        | 94.48           | 85.99        | 92.04        | 87.58        | 87.99        |
5  92.11  95.8  85.78  93.93  84.2  84.86
6  91.25  96.92  86.45  94.61  85.84  86.02
7  91.45  95.43  86.77  90.94  87.96  88.48
8  88.29  94.16  84.56  92.31  88.29  88.66
9  88.14  96.18  84.33  91.81  87.83  88.14
10 88.66  96.15  84.52  93.08  87.83  99.88

Average  90.738  95.56  85.69  92.42  86.753  88.76

The accuracy comparison, the recall rate and the F value of the two experiments are shown in Fig.3. By comparison, the evaluation value of the second semantic-based term extraction method is higher than that of the experimental one. The terminology extraction method based on the dependency tree-CRF based on semantics is effective in the accuracy, the recall rate and the F value of the term extraction.

![Fig.3 Comparison of mean values of two experiments under different evaluation indicators](image)

### 5. Conclusion

The above research on dependency tree is based on the part of speech and grammar to extract technical feature words, more consideration of word meaning, part of speech, and statement. In the face of massive patent text data, CRF requires a large amount of human resources to support the manual labeling of a certain number of technical feature words as training corpus to train the CRF model, and the use of the dependency tree method can easily extract technical feature words. And the dependency tree algorithm based on part of speech and grammar does not ignore low-frequency feature words. When there is a lack of highly reliable parsing tools, using a dependency-tree-CRF learning model is a good alternative. In addition, this method is not only applicable to processing instructions, but also can be extended to process patented texts in other parts without the need for manual marking.

### Acknowledgments

This work was financially supported by the Jiangsu Provincial Graduate Innovation Research Project (KYCX18_2334) and the Ministry of Industry and Information Technology Intellectual Property Promotion Program (GZN09-2017).

### References

[1] Lafferty J, McCallum A. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning, 2001.
[2] Fu Yao, Wan Jing, Xing Lidong. Discovery of specific domain concepts based on conditional random field and information entropy [J/OL]. Computer Application Research: 1-5[2019-01-18].http://kns.cnki.net/kcms/detail/51.1196.TP.20181224.1632.012.html.

[3] Li Na. The Construction of Automatic Extraction Model of Ancient Records of Local Records Based on Conditional Random Fields[J]. Journal of Chinese Information Processing, 2018, 32(11): 41-48+61.

[4] http://nlp.stanford.edu/software/lex-parser.shtml

[5] Chang P, Tseng H, Jurafsky D. Discriminative reordering with Chinese grammatical relations features. In: Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, 2009, pp. 51-59.

[6] Zhou Junsheng, Dai Xinyu, Yin Cunyan et al. Automatic Identification of Chinese Institution Names Based on Cascading Condition Random Field Model[J]. Chinese Journal of Electronics, 2006, 34(5): 804-809.

[7] Shi Shuicai, Wang Wei, Han Yanwei, Lu Xueqiang. Research on domain terminology recognition based on conditional random field [J]. Computer Engineering and Applications, 2013, 49(10): 147-149+155.