Domain Adaptive Chinese Word Segmentation Method for power data security classification

Yang Dong\textsuperscript{1,}\*, Yang Zheng\textsuperscript{1}, Xu Zhang\textsuperscript{1}, Jing Gao\textsuperscript{1} and Zheng Fan\textsuperscript{2}

\textsuperscript{1}State Grid Tianjin Information & Telecommunication Company, Tianjin, 300000, China
\textsuperscript{2}State Grid Tianjin Power Company, Tianjin, 300090, China
\textsuperscript{*}Corresponding author’s e-mail: dysay@126.com

Abstract. Chinese word segmentation (CWS) is an important task for Chinese NLP, and also an essential pre-processing step to establish a word-root database for security classification of power data, covering different domains such as laws & regulations, power. It is impracticable to label a large number of training corpus for each domain, which brings great challenge to the supervised statistical learning method to carry out effective CWS. Therefore, a Chinese word segmentation approach based on dictionary and semi-supervised conditional random field (SS-CRF) is presented. At first, a CRF model for CWS is trained with self-training and active learning algorithms and used to conduct CWS task. Then the dictionary features are introduced to correct the result of CRF based segmentation by adopting RMM algorithm. Experiments on a cross domain segmentation task show that the proposed method can effectively improve the domain-adaptive performance of CWS.

1. Introduction
With increasing development of the construction of the power Internet of things, how to ensure data security is particularly important. At present, research on encryption, desensitization and leakage prevention of power data security has been carried out at home and abroad, but the data security classification technology has not been developed in depth, and the research on the term root of power data security by using NLP technology is also relatively insufficient. However, CWS is the basis for the establishment of term-root database, involving several different domains, such as laws and regulations, power business.

Word segmentation is an essential fundamental task in Natural Language Processing (NLP) for those languages without word delimiters, e.g., Chinese [1]. In recent years, there has been a lot of work in the research field of CWS. Existing approaches for CWS roughly fall into three categories: dictionary-based methods, sequence modeling methods with statistical techniques and neural-network-based methods. It is difficult to deal with the changes in language usages for the dictionary-based method due to the bottleneck brought by the expansion of dictionaries. Statistical methods such as Hidden Markov Model (HMM) [2]and conditional random field (CRF) [3], are reported to achieve high accuracy with large-scale annotated dataset for they can easily learn new words from corpora. In these CWS methods, a key problem is to build the feature vector for each character in sentences, which is constructed via a manual engineering and so is very time-consuming and hard-work. Besides, human annotated corpora of special domains are of resource-intensive. Therefore, the performance of this kind of method is practically not satisfying in domain-specific system. In order to improve the performance of domain adaptation for...
CWS, many approaches are proposed, such as feature expanding [4], active learning [5], and using different tag sets [6]. In [7], a method is presented, where the external dictionaries are able to incorporate into the statistical model to realize domain adaptation for CWS. However, these methods have some limitations because domain-specific dictionaries and the annotated corpora are not available readily.

More recently, deep neural networks have been employed for CWS, for example, convolution neural network [8], gated recursive neural network [9], long short-term memory (LSTM) [10], BLSTM-CRF [11]. Some research reports that these neural networks show better performance than popular statistical algorithms. However, these methods usually depend on numerous labeled sentences. It is very difficult to correctly segment the sentences containing these words which are scarce or absent in training data by using these methods.

To deal with the practical case where labeled data is limited but considerable unlabeled data is readily available, this paper combines the semi-supervised statistical method with the dictionary method to improve the adaptation performance with the help of a small amount of labelled data. Firstly, a semi-supervised CRF word segmentation model incorporated with active learning algorithm is trained with the labeled and unlabeled data, and then this CWS model is applied to test the target text to yield a segmentation result. At last, the dictionary-based maximum inverse matching method is applied to perform a second word segmentation with the help of domain dictionaries to raise the recognition accuracy of professional terms. Empirical studies show that the method proposed in this paper is able to effectively improve the adaptation performance of CWS, especially when the labeled training data is insufficient. Our contributions can be summarized as follows:

(1) We combine the semi-supervised CRF and the dictionary-based method for improving the domain-adapation performance of CWS in case of insufficient annotated corpora in specific domain.

(2) We integrate active learning algorithm into self-training algorithm of the semi-supervised CRF for better performance.

2. Semi-supervised conditional random fields

2.1. Conditional random fields (CRF)

CRF is a probabilistic framework and also a type of discriminative model, where the parameters are estimated by maximizing the joint probability over observation and label sequences. CRF has been studied in depth and performed well for sequence labeling task since it was first proposed by (John Lafferty,2001). Linear-chain CRF is the typical type of the CRF models and most widely used in machine learning tasks, in which a conditional probability for a labeling sequence \( Y=(y_1, y_2, \ldots, y_n) \) given in an input sequence \( X=(x_1, x_2, \ldots, x_n) \) is defined as :

\[
P_\theta(Y / X) = \frac{1}{Z_X} \exp \left\{ \sum_i \sum_k \theta_k f_k(y_i, x, i) \right\}
\]

Where \( f_k(y_{i-1}, y_i, x, i) \) is a feature function, \( \theta_k \) is a learned weight associated with the feature function and \( Z_X \) is the normalization factor over all state sequence, which is calculated according to the following formula:

\[
Z_X = \sum_y \exp \left( \sum_i \sum_k \theta_k f_k(y_{i-1}, y_i, x, i) \right)
\]

For an input \( X \), the most probable label sequence can be efficiently determined by adopting the Viterbi algorithm.

\[
\theta' = \arg \max_{\theta} [ P_\theta(Y / X) + \alpha U(\theta) ]
\]

Where \( U(\theta) \) is the regularization term, being used to prevent over-fitting; \( \alpha \) is a hyper-parameter which is used to balance the proportion between the logarithmic likelihood term and the regularization term.

The parameter optimization in (3) is usually conducted by using L-BFGS algorithm, which simplifies the computing complexity and alleviates memory burden through the approximation of BFGS algorithm.
2.2. Self-training algorithm and active learning algorithm
Semi-supervised learning is a statistical machine learning method which combines supervised and unsupervised learning. It can train and classify a large number of unlabeled corpus and a small number of labeled corpus without manual intervention. Therefore, in case of small amount of labeled corpus available, semi-supervised learning is a good approach for obtaining a model with strong generalization ability. As the simplest but effective semi-supervised learning algorithm, self-training algorithm is adopted in this paper. However, these pseudo labels may contain quite some wrong labels, which can impair the performance of subsequent training procedures due to accumulative errors.

On the other hand, active learning algorithm is to actively select unlabeled sample labels which are beneficial to model training so as to reduce the labeling cost and the computational scale as much as possible. It has been successfully applied for NLP task. Therefore, the active learning algorithm is incorporated into semi-supervised learning with self-training technique to select high-quality pseudo labels in this paper.

2.3. Formatting author affiliations Semi-supervised CRF learning
In this paper, the samples that cannot be labeled correctly by the pre-training model are called positive data and stored in the unlabeled data set. Here, the positive data will be selected according to confidence value. In CRF model, the conditional probability \( P(Y|X) \) is smaller, the probability of achieving the optimal solution \( Y \) given observation sequence \( X \) is smaller. Therefore, \( P(Y|X) \) can represent the confidence score of labeling results. Here multiple groups of data are selected as positive data, whose \( P(Y|X) \) is smaller than the specified threshold.

The flowchart is shown in Figure 1 for the semi-supervised CRF learning algorithm proposed in this paper. The procedure for learning the CRF model is as follows:
- Input: labeled training sample set \( L \) and unlabeled training sample set \( U \).
- Step1: Obtain a small number of labeled corpus as \( L \);
- Step2: Train CRF by using \( L \) to achieve the model \( C_k \);
- Step3: Tag \( U \) with Model \( C_k \) and estimate the confidence score \( P(Y|U|X) \);
- Step4: Select multiple groups of data with confidence score lower than the set threshold in \( U \) as positive data and mark as \( U_P \);
- Step5: Label \( U_P \) manually and then mark them as \( u \);
- Step6: Add \( u \) to \( L \) and delete \( u \) from \( U \);
- Step7: Iterate the above process for \( K \) times until the convergence condition is satisfied. Here the convergence condition is set as:

\[
\left| P^{(k)}(Y|U|X) - P^{(k-1)}(Y|U|X) \right| \leq \varepsilon
\]  

Output: the CRF model \( C_{k+K} \).

3. Domain adaption CWS method
The matching methods based on dictionary have achieved great success in CWS. Compared with the word segmentation corpus used in the statistical model, dictionaries are easier to obtain. But the previous methods based on dictionary perform worse than statistical model in solving the ambiguity problem of
word segmentation. Therefore, the statistical CWS model and the dictionary are combined to improve the accuracy and domain adaptability of CWS in this paper.

3.1. Tag Set and Feature Template

CRF is applied for CWS by conducting a sequence labeling task, that is, labeling the position of each Chinese character in word which the character belongs to. This paper used the 4-tag set [3]. A context of three characters for feature generation and feature templates represented in Table 1 are adopted in this paper, which are called as basic features.

Table 1. Feature templates

| Type           | Feature                        | Description                                      |
|----------------|--------------------------------|--------------------------------------------------|
| Unigram        | C_{i-1}, C_0, C_i              | C_0 denotes the current character; C_i (C_{i-2})  |
| Bigram         | C_{i-1} C_0, C_0 C_i, C_{i-2} C_i | denotes the character i positions to the right (left) of the current character. |
| Punctuation    | IsP(C_0)                       | Current character is punctuation                 |
| Character type | T(C_{-2}), T(C_{-1}), T(C_1), T(C_2) | Types of character: date, numeral, alphabet, others |

3.2. Feature representation of dictionary information

In order to represent features of dictionary information in CRF model, for a given sentence X= (x_1, x_2, ..., x_n) and the dictionary D, two functions about the jth character c_j (1 ≤ j ≤ n) are defined as follows:

\[ f_\ell(X, j, D) = \max l, \ s.t. \ \left\{ \begin{array}{l} W = C_{j-l} \cdots C_j \in D \\ 1 ≤ j - l - 1 \end{array} \right. \]  \tag{5}  

\[ f_m(X, j, D) = \max l, \ s.t. \ \left\{ \begin{array}{l} W = C_{j-l} \cdots C_j \in D \\ 1 ≤ s - l - 1 < j \\ j < s ≤ n \end{array} \right. \]  \tag{6} 

Where W denotes a word. f_\ell(X, j, D) represents the maximum length of a word which is obtained at the position j in a sentence X by using reverse maximum matching (RMM) according to dictionary D. f_m(X, j, D) represents the maximum length of the word including the character C_j but not ending with C_j and ending with C_s (s > j), which is obtained in a sentence X by using RMM according to D.

The extended features related to dictionary D introduced in this paper are shown in Table 2. Assuming the current position is j, [f_\ell]_i = f_\ell(X, j+i, D), [f_m]_i = f_m(X, j+i, D). For example, a sentence is "供电服务质量好", if the words related to this sentence in the dictionary include "供电", "服务", "供电服务", "质量", "服务质量", "好", and the current character is "务", then [f_\ell]_i = 0, [f_\ell]_i = 4, [f_\ell]_i = 0, [f_m]_i = [f_m]_i = 4 .

Table 2. Dictionary feature used in CRF model

| Dictionary feature | Description |
|--------------------|-------------|
| Unigram [f_\ell], [f_m] (i = -1, 0, 1) |
| trigram [f_\ell], [f_m], [f_m], [f_m], [f_m], [f_m], [f_m], [f_m], [f_m], [f_m] |

3.3. Domain adaption CWS

The following Figure 2 demonstrates the framework of domain-adaption-CWS method which introduces the domain dictionary into the proposed semi-supervised CRF model. Firstly, the training and test corpus are preprocessed and then feature selection and extraction are carried out. Secondly, the semi-supervised learning method described above is employed to learn CRF model with training corpus so as to generate an optimal CRF model, in which the L-BFGS algorithm is employed for parameter estimation of the CRF. Thirdly, the CRF model is used to segment words of the test corpus with the Viterbi algorithm.
Finally, the domain dictionaries are introduced to correct the first segmentation result with the tri-tree RMM algorithm, and then output the final segmentation results.

![Figure 2. Framework of domain adaption CWS.](image)

4. Experiments

4.1. Experimental Settings

The labeled training corpus used in this paper is PKU corpus provided by SIGHAN CWS bake-off 2005, which is the news corpus of People's Daily. The size of the corpus is 7548kb. Since there is no authoritative and unified corpus of laws and regulations on network security and electric power business, 200 relevant documents are collected manually, such as “Network security law of the people’s Republic of China”, “Provisions on the confidential scope of State Grid Corporation of China”, “Provisions on classified scope of State Grid Corporation of China”, with a total of 63058 words. Then, Antconc 3.5.7 is used to make the unlabeled corpus based on the collected texts. The domain dictionary includes Chinese law dictionary and power term one. By implementing the new-word-recognition algorithm proposed in reference [12], the new words are mined in the test corpus, and then the top 300 words are picked up manually to update the domain dictionary.

CRF++, a famous CRF framework written in C++, is implemented to carry out training and inference work in this paper. Precision (P), recall (R) and F1-score (F1) are used as the evaluation metric of our model. Three groups of experiments were conducted to verify the effectiveness of the proposed CWS method.

Test 1: The CRF model for CWS was trained by using this method presented in this paper, where the threshold value of confidence is 25%, the data with confidence lower than 25% are manually annotated and then added to the labeled training set for the next iteration. Then the CRF model is used to segment the test text.

Test 2: The domain dictionary was introduced to correct the results achieved by Test1.

Test 3: The semi-supervised CRF with self-training method was used to learn the training corpus iteratively. The samples with confidence higher than 85% were directly added to the training sample set for the next iteration until convergence. After the first CWS is completed by the CRF model, the domain dictionary is introduced for correction.

4.2. Experimental Results and Analyses

When the hyper-parameter $\alpha$ varies between 0.001 and 0.1, $F_1$ in the network security laws and regulations corpus increases first and then decreases, and is maximal at $\alpha = 0.01$; when $\alpha$ changes from 0.0001 to 0.01, $F_1$ in the power business corpus shows the same changing trend, and reaches a maximum at $\alpha = 0.001$. When $\alpha$ is the optimal value, the inference results in three experiments are shown in table 3. The following conclusions can be drawn from table 3:

(1) Adding dictionary and integrating active learning algorithm into self-training semi-supervised learning are both able to improve the performance of CWS.

(2) In test 3, the CWS result achieved by using the semi-supervised method with self-learning algorithm is not good. Because network security and power engineering both have their unique text characteristics while the labeled corpus used by pre-training CRF model belong to the common domain, the accuracy of word segmentation in specific domain will be reduced. Besides, errors will be accumulated in the process of iterative learning, which will lead to poor performance of the classifier.

(3) In test 2, the method proposed in this paper is used to segment text, in which the data with low confidence are manually labeled to reduce the accumulation of errors in the process of iterative learning.
The test result shows that the semi-supervised learning method proposed in this paper can work better on CWS in specific domain.

Table 3. CWS results of different methods

| Testing corpus                      | Test | $\alpha$ | $R$ (%) | $P$ (%) | $F_{1}$ (%) |
|-------------------------------------|------|----------|---------|---------|-------------|
| Laws and regulations of network safety | 1    | 0.01     | 86.8    | 88.6    | 87.7        |
|                                     | 2    | 0.01     | 89.9    | 91.3    | 90.6        |
|                                     | 3    | 0.01     | 64.4    | 65.4    | 64.9        |
| Power business                      | 2    | 0.001    | 88.2    | 89.1    | 88.6        |
|                                     | 3    | 0.001    | 61.5    | 62.8    | 62.1        |

5. Conclusions

This paper proposes a semi-supervised CRF model combining self-training learning with active learning, which can improve the performance of cross-domain CWS. In process of training CRF model with the self-training algorithm, the data with low confidence are manually annotated and added to the labeled samples to reduce the accumulation of errors of iterative learning. Besides, in order to further improve the domain adaptability of CWS, the dictionary features are integrated into CRF model to correct the CRF segmentation results by adopting RMM algorithm. Empirical studies show that the proposed method can effectively improve the domain-adaptive performance of CWS, especially when the labeled training data is insufficient.

In our future work, we would like to automatically mine various domain related words to expand domain dictionary so that our word segmentation approach can adapt to various domains.

References

[1] Xue, N.W. (2003) Chinese word segmentation as character tagging. Computer. Linguist. Chin. Lang. Process, 8(1): 29–48.
[2] Eddy, S.R. (2001) Hidden markov models, Curr. Opin. Struct. Biol. 6(3): 361–365.
[3] Chi C.Y., Yu C.Y., Zhan X.G. (2008) A Chinese Word Segmentation Approach Using Conditional Random Fields, Journal of Information, 5: 79–87.
[4] Shi Y.X., Wang M.Q. (2007) A dual-layer CRFs based joint decoding method for cascaded segmentation and labeling tasks, In: Proceedings of IJCAI, India, pp.1707–1712.
[5] Li S.S., Zhou G.G, Huang C.R. (2012) Active learning for Chinese word segmentation, In: Proceedings of COLING, Japan, pp. 683–692,
[6] Zhao H., Li C.N., Lu M. (2006) Effective tag set selection in Chinese word segmentation via conditional random field modeling, In: Proceedings of PACLIC, Wuhan, pp. 87–94
[7] CHENG Y.S., SHI Y.T. (2018) Domain specific Chinese word segmentation, Computer Engineering and Applications, 54 (17): 30-34.
[8] Zheng X.Q., Chen H.Y., Xu T.Y. (2013) Deep learning for Chinese word segmentation and POS tagging, In: Proceedings of EMNLP, pp. 647–657.
[9] Xu. J., Sun X. (2016) Dependency-based gated recursive neural network for Chinese word segmentation, In: Proceedings of ACL, pp. 567–572
[10] Chen X.C., Qiu X.P., Zhu C.X., Liu P.F. (2015) Long short-term memory neural networks for Chinese word segmentation, In: Proceedings of EMNLP, pp. 1197–1206.
[11] YAO M.J., LI H.J., Lu H.H. (2018) Chinese word segmentation method based on BI-LSTM-CRF neural network for sequence labeling, Modern Electronics Technique, 42(1): 95-99.
[12] Luo Z.Y., Song R., (2007) Adaptive Method for Chinese New Word Identification Based on Multi-features, Journal of Beijing University of Technology, 33(7): 718-724.