An Adaptive Chinese Word Segmentation Method

Zhi Yuan

Institute of Electronic Science and Technology,

University of Electronic Science and Technology of China, Chengdu 611730, China

Abstract: Due to the limitations of the field of training corpus, the Chinese word segmentation based on statistic results in poor self-adaptability in the field. In view of the difficulty of obtaining large-scale annotation corpus in the target area, this paper proposes an area adaptation method that combines domain dictionaries with active learning algorithms. Select a small-scale corpus containing the largest number of unmarked discrepant sentences to prioritize manual annotation, by the statistical analyzing of the difference between the target area text and the existing annotation corpus. Then combine the n-gram statistics in large-scale texts to train the segmentation model in the target area. Finally, the domain adaptiveness is achieved by integrating lexical information into the CRF statistical word segmentation model. Experiments show that this method significantly improves the domain adaptive ability of statistical Chinese word segmentation.

Keywords: Chinese word segmentation; Active learning; CRF; domain adaption

1. Introduction

Chinese word segmentation refers to the process of separating the sequence of Chinese characters that make up a sentence into separate word sequences using delimiters. Chinese word segmentation methods generally can be divided into three categories: based on the dictionary matching method, based on statistical methods and based on neural network model segmentation method [1].

As the field of Chinese word segmentation changes, the proportion of unregistered words will increase, resulting in a significant drop in the accuracy of the Chinese word segmentation system.

Statistics-based methods have high accuracy in word segmentation, but there is a great shortage of cross-cutting areas. The dictionary-based approach uses dictionaries as the primary resource. Such methods do not need to consider the problem of domain adaptability.

In this paper, based on the combination of dictionary and statistical methods, we can make use of dictionaries to make up for the lack of cross-regional statistical methods. At the same time, in order to further improve the accuracy and performance of word segmentation, this paper uses the CRF model to train the common corpus. After obtaining the word segmentation model, we continue to adjust the model by using Active Learning in specific fields so that the language model can obtain cross-domain segmentation ability.

2. Active learning

Active learning algorithm proposed by Professor Angluin [3] Yale University. It selects some of the unlabeled samples to mark, then puts them into a previously existing set of labeled samples, retrains the classifier, and selects another untagged sample using the classifier. By selectively expanding the set of marked examples and cyclic training, the classifier is gradually gaining more generalization ability. Compared with the previous algorithm, it has the characteristics of simulating

the human learning process, so it has been widely concerned. In recent years, it has been widely used in the field of natural language processing such as information retrieval and text classification, and has become one of the most important directions in the field of machine learning.

Active learning algorithm makes the choice of samples in learning process more reasonable by a prior selection algorithm, so that the precision of trained classifier is higher. In the work of sequence annotation applied to natural language processing, more distinguishing information can be obtained under the same annotation cost, which helps to improve the accuracy of the classification model.

3. CRF Chinese word segmentation

Xue Nianwen [4] et al. Proposed in 2003 that Chinese word segmentation should be regarded as sequence annotation. Each word in the sentence is classified according to its position in the word and is divided into four categories: B, M, E, S. Where B is the beginning of each word, M is the middle position of a word, E is the end of a word, and S is the character that can independently form a word. CRF is currently the mainstream sequence labeling algorithm, which has achieved great success in the issue of sequence labeling. For a given sentence $\mathrm{x} = c_1 \dots c_n$ and one of its participle annotation results $\mathrm{y} = y_1 \dots y_n$,where c_i is the input character,$y_i \in \{B, M, E, S\} (1 \le i \le n)$, we can express the probability of y as follows:

$$P_\lambda(y|x) = \frac{1}{z(x)} \exp(\lambda \cdot \sum_{i=1}^{n} \Phi(y_{i-1}, y_i, \mathrm{x})) \quad (1)$$

$z(x)$ is a normalization factor, $\Phi(y_{i-1}, y_i, \mathrm{x})$ is the eigenvector function, and λ is the feature weight vector.

4. An Adaptive Chinese Word Segmentation Method Based on Active Learning and Dictionaries

Suppose there is already a Chinese word segmentation model trained on the corpus of certain field and a high-quality dictionary of the target area. To segment the text in the target area, we need to adjust the Chinese word segmentation model from the original area to the target area.

In order to deal with the segmentation of proper nouns and special sentences in the field, this paper proposes an adaptive Chinese word segmentation based on active learning and dictionaries. With the aid of Active Learning algorithm, small-scale sentences with the most domain characteristics in the target area are selected for manual annotation in the training process of the model, and then the cross-domain nature of the model is improved by merging with the statistical characteristics of the large-scale corpus n-gram in the field. By using the auxiliary information provided by word-specific dictionaries in the word segmentation task, the influence of the domain on the word segmentation can be greatly reduced without changing the original word segmentation model.

4.1 Model retraining

Compared with the distribution of words in the original domain, there are great differences between Chinese characters and word formation patterns in the distribution of words in the target domain. Due to the large number of statements in the target area, artificial standards are more difficult. Therefore, how to filter out such a sentence with more differences becomes the key.

In this paper, the method based on n-gram weighted statistics is used to calculate the n-gram distribution difference of each sentence relative to the original domain. Specific calculation as shown in the formula：

$$\Psi^N(t) = \sum_{k=1}^{N} \frac{\omega_k}{|X_t^k|} \sum_{x \in X_t^k} log \frac{P(x|U,k)}{P(x|L,k)} \quad (2)$$

L represents the original domain statement set, U represents the target domain statement set. X_t^k represents the n-gram set in sentence S. $P(x|D,k)$ represents the probability of x in the n-gram set of statement set D. ω_k represents the weight of adjusting the importance of n-grams of different lengths. This article uses the N = 4 language model. $\Psi^N(t)$ is the score of the sentence S. The higher the score, the statement contains more strings that are not covered by the original domain. The lower the score, the lexical distribution of the sentence is close to that of the original field.

After obtaining the artificial annotation corpus, then according to Figure 1 steps to complete the model retraining.

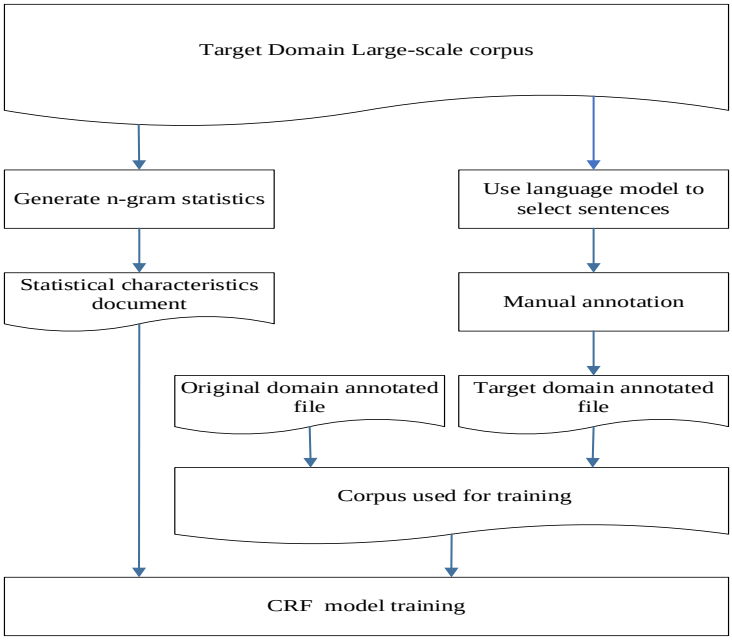

Fig. 1 Model Retraining Based on Active Learning

4.2 Application of domain dictionary

In this paper, statistical Chinese word segmentation model incorporates the relevant features of the dictionary, making statistical Chinese word segmentation model and the dictionary organically.

Since the features used in this paper do not rely on specific words, they use the idea of a maximal match to provide the length information of the longest word containing the current character to the statistical model. Therefore, when a Chinese word segmentation is performed for a specific field, the original word segmentation model does not need to be changed, and only the corresponding dictionary of the field needs to be loaded, thereby greatly reducing the influence of different fields on word segmentation. The specific application of the dictionary is shown in Figure 2.

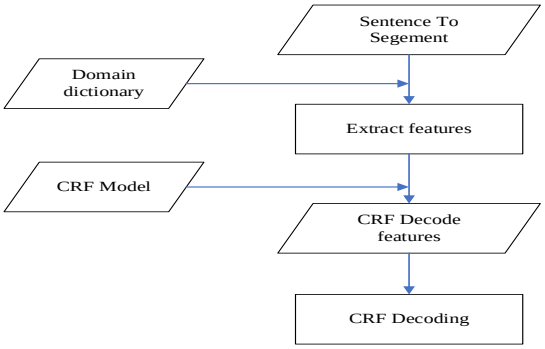

Fig. 2 Application of domain dictionary

5. Experiment

In order to evaluate the performance of the Chinese word segmentation model adapted in the specific field presented in the previous section and verify the effectiveness of the proposed method, a set of experiments was designed on news and financial corpus data.

The experiment is divided into two steps. The first step is to complete the retraining of the participle model. The second step is to test whether the participle model that incorporates a domain dictionary as a CRF feature is valid.

First of all, as the annotation corpus of the original field, the first to 270, 400 to 931 and 1001 to 1151 annotation data in CTB5.0 [5]in the field of information are used.

Then from the financial corpus data randomly selected 300 sentences for manual annotation. Using news corpus and financial corpus for model retraining.

Finally, select the news dictionary in the universal dictionary, and the specific domain dictionary obtains 1000 special terms in the financial field manually. The validity of this method is tested by adding the features of the dictionary of financial field during the segmentation task.

5.1 Experimental results and analysis

First, after the model is retrained, the model proposed in this paper is compared with the model that does not have the field adaptability and the model that has the field adaptability but does not adopt the active learning method. The results are shown in Table 1:

Table 1 Comparison of three systems

Chinese word segmentation system	P/%	R/%	F1/%
Our （Active Learning +n-gram+ Original Domain Corpus+ Target Domain Corpus）	88.3	90.4	89.4
Baseline (Random +n-gram+ Original Domain Corpus + Target Domain Corpus)	88.0	89.6	88.8
No domain adaptive word segmentation system（News Domain）	69.7	78.9	74.1

By comparing the evaluation results of domain-adaptive system and domain-free adaptive word system in the comparison table, we can see that all the evaluation results of domain-adaptive word segmentation system are higher than those of domain-free adaptive word segmentation system. It shows the importance of domain adaptation in improving the performance of Chinese word segmentation system.

By comparing the word segmentation system of the corpus with the Active learning method and

the participle system of the random selection of the corpus in the two participle systems adaptively added in the field of artificial annotation, we can see that the former results are higher than those of the latter. It can be seen that Active Learning has obvious effect in the field of adaptive model training.

After obtaining the preliminary adaptive word segmentation model in the field, continue to join the domain dictionary information without changing the model.

In the financial field test corpus, keep training the CRF participle model unchanged, the use of the dictionary is based on the training of the corpus dictionary added about 1000 financial domain dedicated dictionaries. Table 2 shows the test results in the financial field. As can be seen from the table, using the dictionary as auxiliary information in the execution of the word segmentation task can further improve the accuracy of the word segmentation system.

Table 2 With the Domain Dictionary

Chinese word segmentation system	P/%	R/%	F1/%
OUR （Active Learning +n-gram+ Original Domain Corpus+ Target Domain Corpus+ **Target Domain Dictionary**）	88.7	90.4	89.4
Baseline (Random +n-gram+ Original Domain Corpus + Target Domain Corpus+ **Target Domain Dictionary**)	88.0	89.6	88.8

As can be seen from Table 2, after the information characteristics of the dictionary are incorporated into the statistical model, the performance of the participle has been improved to some extent. On the other hand, the performance of the participle can still be maintained at a certain level after the relocation of the field.

6. Conclusions

In this paper, Active Learning algorithm is introduced in the model training period to reduce the amount of manual annotation tasks while ensuring that the model has some cross-domain capabilities. At the same time, it refers to specific domain dictionaries in word segmentation tasks and improves the precision of word segmentation without changing the model. The results show that the method not only achieved good results in the original field, but also showed good results in the cross-field situation.

To sum up, this article made some explorations on the adaptive task in Chinese word segmentation and achieved preliminary research results. However, there are still many problems with Chinese word segmentation. For example, in the face of different fields, the segmentation of word segmentation is a problem. In the future, we will select other representative areas to conduct a more in-depth exploration of adaptation methods in Chinese word segmentation.

References

[1] Huang Chang-Ning, Zhao Hai. Chinese word segmentation: A decade review. Journal of Chinese Information Processing，2007，21(3)：8—20.)

[2] Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML ’01 Proceedings of the Eighteenth International Conference on Machine Learning, 8(June), 282–289.

[3] Angluin ,D. (1988). Queries and Concept Learning. Machine Learning, 2(4), 319–342. https://doi.org/10.1023/A:1022821128753
[4] Xue, N. (2003). Chinese Word Segmentation as Character Tagging. Computational Linguistics and Chinese Language Processing, 8(1), 29–48.
[5] Li Hang. Statistical learning methods[M]. Beijing: Tsinghua University Press,2012:192-209
[6] Zong Qing. Statistical natural language processing[M]. Beijing: Tsinghua University Press,2008
[7] GB/T13715-1992. Modern Chinese word segmentation specification for information processing[S]. Beijing: China Standard Press,1992：
[8] Yue Zhang, Stephen Clark. Chinese segmentation with a word-based perceptron algorithm[C]//Proceedings of the 45th ACL. 2007:840-847