Statistical Chinese Word Segmentation using Domain Dictionaries

Hengjun Wang[1, a *], Nianwen Si[1,b] and Xiaopeng Li[1,c]

[1] Zhengzhou Institute of Information Science and Technology, Henan, China

[a] wanghengjun@163.com, [b]snw1608@163.com, [c]peng001123@sina.com

***Abstract* — Chinese word segmentation is the basic task for natural language processing. Also, many related research studies on Chinese word segmentation have gained considerable accuracies, but they are usually limited to specific fields. To deal with the domain adaptability problem, in this paper we propose an effective statistical model which combines the basic Conditional Random Field method with C-value based domain dictionary to make word segmentation. The Conditional Random Field makes rough segmentation to obtain primitive results, then the model uses the constructed domain dictionary to make refined segmentation based on previous results. Experimental results show that the proposed model achieves competitive accuracy on news and blog corpus.**

Keywords - Chinese word segmentation, Natural language processing, Conditional Random Field, C-value, Domain adaptability.

I. INTRODUCTION

With the rapid development of artificial intelligence and machine learning technology, researches in natural language processing has gained many processes, and been applied into lots of intelligent field[1],[2]. In natural language processing community, Chinese word segmentation is the basic process for higher order tasks, only correct word segmentation could help to achieve correct machine understanding. Sequence labeling methods for example Hidden Markov Model[3] and Conditional Random Field[4], have been widely used in many natural language processing tasks such as word segmentation[5][6], part-of-speech tagging[7],[8][9], named entity recognition[10][11] and semantic role labeling[12][13].

In all kinds of sequence labeling models, Conditional Random Field(CRF) model is widely acknowledged and researched[14] because it could make use of more contexture information and obtains higher accuracy[15][16]. For Chinese word segmentation, CRF model has been studied and applied in many researches for general fields segmentation. However, in some special domains, CRF model could not gain high accuracy as they do in general fields due to the existing of special domain terms. To make the word segmentation model better adapted for the special domain, this paper proposes an effective model which combines CRF model with domain dictionary. Based on domain dictionary, the basic CRF model will further improve the accuracy for some special domains.

II. RELATED WORK

Chinese word segmentation is the basic task of natural language processing, which plays an important role in Chinese information processing technology. After about twenty's development, Chinese word segmentation technology has made many progresses, lots of models and algorithms related were proposed[17][18], as well as some useful word segmentation applications. In recent years, with the development of machine learning and statistical theory, and their application in word segmentation, the accuracy thus being improved significantly. Currently, the algorithm of word segmentation can be divided into two categories: rule-based word segmentation and statistic-based word segmentation.

Rule-based word segmentation algorithm is the traditional method, the main idea of which is to split the sentence according to the rich and integrated word dictionary, the most representative algorithms of this kind are Forward Maximum Matching method(FMM) and Backward Maximum Matching method(BMM). Intuitively, the core elements of these word segmentation algorithms are the split rules and word dictionary, and the algorithm process is relatively simple and have low time complexity. However, simply utilizing word matching method will cause low accuracy, could not handle the complexity and diversity of language, especially facing with some Out Of Vocabulary words(OOV).

With the rapid development of statistical learning methods, more and more researchers attempt to introduce these machine learning algorithms into word segmentation field[19]. These algorithms usually develop a segmentation model, then train the model with manually annotated training corpus, after iterating several training process the algorithm will chose the model which obtains the best performance on development dataset. Finally, the trained model could make word segmentations over the raw sentence. Due to the advantage of high accuracy and speed, statistic-based word segmentation algorithms are widely

acknowledged and adopted now in many natural language processing systems.

However, despite that these models has achieved considerable accuracies, most of them just focus on the word segmentation in general field such as economical corpus and news corpus. While transformed into other domains especially for some specified areas, they don't work well compared with the general field, because the model training is usually constrained by the scale of the training corpus[20][21]. Therefore, some researchers start to study word segmentations over specific field[22][23][24].

III. WORD SEGMENTATION MODEL

In this paper, we propose a Chinese word segmentation model which combines Conditional Random Field(CRF) model and word dictionary to make word segmentation on raw sentence. The CRF model could make rough segmentation according to sequence labeling method, then we use C-value statistic to make further elaborate modification, in this way, the word segmentation model could obtain high accuracy rate. Figure 1 illustrates the word segmentation model architecture.

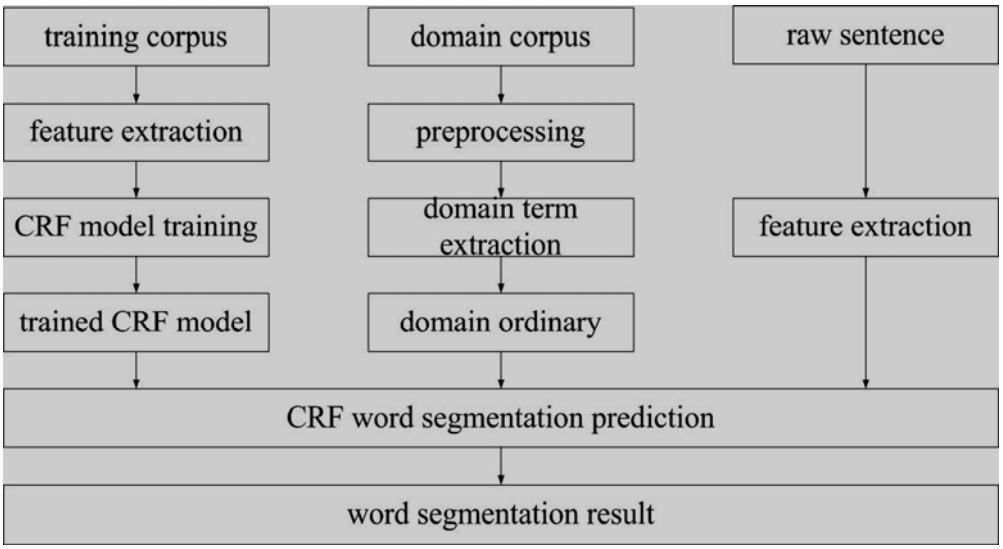

Fig. 1 CRF-based word segmentation architecture.

The above figure shows the main process of the word segmentation based on CRF and dictionary. In the process of establishing the CRF model, the model will first extracts features according to designed feature template, then the model will be trained on standard training corpus. In the process of constructing domain dictionary, we use some statistics such as C-value to extract domain term for form the specific dictionary. As for testing, the trained model will also extract features and then make sequence predictions for word segmentation.

A. CRF-based word segmentation

1) Conditional Random Field

Xue et al.[25] proposed to treat word segmentation process as sequence labeling, their main idea is to classify the word in the sentence into several categories, and make predictions for every character of the word with four labels B, M, E, S, where B denotes the beginning of the word, M denotes the middle and E denotes the end of the word. For the special word which consists of one word such, they use S to label this single word.

CRF model is firstly proposed by John Lafferty et al.[4] in 2001, it is the main sequence labeling model now based on statistic theory, CRF has been applied into many field such as Chinese word segmentation, POS tagging and Named Entity Recognition. Being the main approach for word segmentation, CRF allows users to utilize more features such as word uniform, pos tags and their combination with adjacent words, users could design the features by themselves to better improve the model performance.

CRF is an undirected graph discriminative model in essential, for the observation sequence $x = x_1 x_2 ... x_n$, where $x_i (i = 1, 2, ..., n)$ denotes the ith word of sentence x and n is sentence length, CRF will make predictions for state sequence $y = y_1 y_2 ... y_n$ under the conditional probability.

$$p(y \mid x, \lambda) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \sum_j \lambda_i f_j(y_{i-1}, y_i, x, i)\right). \quad (1)$$

Where $f_j(y_{i-1}, y_i, x, i)$ is the feature function with positive real number. The state sequence y will be obtained for the given observation sequence x, and λ_i is the weight vector corresponding to jth feature function. $Z_\lambda(x)$ is the normalization factor to make the sum of all the predicted probability to be one.

2) Feature Extraction

The choice of feature is pretty important for CRF model. Different features will generate different influence for the word segmentation model. Based on previous works which adopt CRF model to make word segmentation, in this paper, we choose similar model features. They are word n-gram feature, character category feature and the position feature of each character in one word. The specific description of each feature we use is in the following tables.

In all kinds of feature for CRF model, the word uniform itself is a kind of good feature, this has been proved in many previous work. In this word we design word unigram, bigram, and trigram features as the word uniform feature.

TABLE 1. WORD UNIFORM FEATURE

feature type	feature description
word n-gram feature	C_{-2} , C_{-1} , C_0 , C_1 , C_2
	C_{-2} C_{-1}, C_{-1} C_0, C_0C_1, C_1C_2
	C_{-2} C_{-1} C_0, C_{-1} C_0 C_1, $C_0C_1C_2$

Where C represents the word in the sentence, the subscript of C denotes the relative location of C. For example, C_0 denotes the current word to be tagged, and C_{-1} denotes the former word of C_0. Word n-gram features usually give the CRF model more contexture information which could be much helpful for word segmentation tagging.

Commonly, we often treat Chinese word as several categories in sequence labeling model, such as word, punctuation, English and number. In this paper, we treat part of speech tags as several categories, the following table describes a part of them.

TABLE 2. WORD CATEGORY FEATURE

word category	feature description
noun	N
verb	V
adjective	A
preposition	P
conjunction	C
punctuation	W
alphabet	NX

Compared with other CRF models which make word segmentation, this paper choose two word position schemes to make prediction for tags of each word, and compare their performance which could better improve the accuracy of prediction. They are four-word position tagging (4-tags, B, M, E, S) and six-word position tagging(6-tags, B, Mi, M, E, S)(i=1, 2, 3…). The following table describes the tagging scheme in detail.

TABLE 3. WORD POSITION FEATURE

tagging scheme	tagging set	four-words tagging	six-word tagging
four tags	B, M, E, S	B, M, M, E	B, M, M, M, M, E
six tags	B, Mi, M, E, S (i=1,2,3…)	B, M1, M, E	B, M1, M2, M3, M, E

Where B denotes the beginning of the word and E denotes the end of the word, M denotes the middle of the word. Especially, when some single words exist in the sentence, for example the prepositions "和", "且" and "或", they will be tagged with S. Particularly, when the long words appear in the sentence, then 4-tags scheme and 6-tags scheme will have different tags for them, 6-tags scheme will give relative order for each word which contains more feature information compared with 4-tags scheme. We will test these two tagging schemes in our experiment separately to see their performance.

B. Statistic-Based Domain Dictionary

1) C-value Variable

Frantizi et al.[26] propose to use C-value/NC-value variable to make term extraction and achieve good performance. Liang et al.[27] proposed to combine C-value variable and mutual information to make term extraction, their method shows pretty good accuracy for geological information field. In this work, we use only C-value variable to make domain term, since the statistical CRF model has gained pretty good performance, the C-value variable will be used to construct domain dictionary under the help of artificial work.

The computation of C-value variable can be described as follows

$$C-value = \begin{cases} \log_2 |s| f(s) , & s \text{ is not contained} \\ \log_2 (f(s) - \frac{1}{P(T_s)} \sum_{w \in T} f(w)) , & \text{otherwise} \end{cases} \quad (2)$$

In the above formula equation, s denotes the possible term string and the corresponding length is |s|. f(s) is the

frequency of string s in corpus. Ts is the term set which contains string set. P(Ts) is the number of term which contains the term s, and w denotes the term which contains term s in the term set Ts. Intuitively, this formula shows the C-value will grows in the direct proportion with the length |s| and frequency f(s). if string s is a complex term, then the frequency of s should be computed with the subtraction between f(s) and complex term frequency.

2) Domain Dictionary

Developing a domain dictionary is an effective way which can further improve the word segmentation accuracy[28]. In this work, we propose a method which use C-value variable to develop the domain dictionary under the help of artificial rule, the process is following.

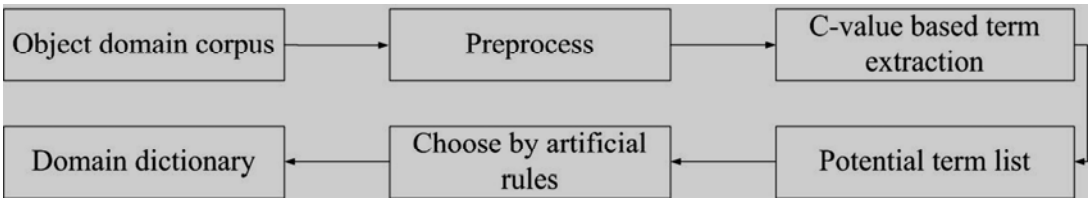

Fig. 2 the construction of domain dictionary.

The above process firstly make preprocess for object domain corpus, then based on C-value variable, the model compute the C-value to make term extraction in order to shape the list which contains all potential domain terms. After this, the list may contains many useless term which is not helpful for the word segmentation, so we should filter them with artificial rule, finally the domain dictionary is constructed.

IV. EXPERIMENT

In the experiment, we use news corpus which was annotated by Peking University from People's Daily in January of 1998 as training corpus. The scale of the training corpus is nearly 200 million which can train the CRF model well. For the detail process, we firstly use CRF++ toolkit to train the model with training corpus, then the domain dictionary will be constructed based on C-value variable. At last, the model will combine the trained CRF and domain dictionary to make word segmentation.

A. Evaluation Criterion

In this paper, we use the common evaluation criterion which contains three aspects: P denotes accuracy, R denotes recall and F measure, the formally described as follows.

Accuracy rate:

$$P = \frac{\text{number of correct segmentations}}{\text{number of all segmentations}} \times 100\%. \quad (3)$$

Recall rate:

$$R = \frac{\text{number of correct segmentations}}{\text{number of gold segmentations}} \times 100\%. \quad (4)$$

F-measure:

$$F = \frac{2PR}{P+R} \times 100\%. \quad (5)$$

Where the accuracy P denotes the proportion of correct of segmentations in all segmentations. R denotes the proportion of correct of segmentations in the gold segmentations. F-measure is the synthesize of accuracy rate P and recall rate R.

B. Results and Comparisons

To compare the model performance with different tagging scheme, we use two tagging schemes in the experiment separately. The first one is four position tagging scheme, the second one is six position scheme. We test our model in small scale of news corpus and blog corpus which are annotated manually. The news corpus is used in the test to evaluate the validity of the proposed method , the blog corpus is used as the domain corpus which contains many new words which appears in recent years while these new words don't exist in training corpus, thus it could be used to evaluate our model's adaptability.

The news corpus contains 2063 sentences and 7053 words, and the blog corpus contains 2537 sentences and 6571 words. The basic CRF model is just utilizing the standard Conditional Random Field to make word segmentation. The model with 4-tags represents that the proposed model with domain dictionary using 4-tags scheme. The model with 6-tags represents that the proposed model with domain dictionary using 6-tags scheme. The experiment result is in table 4 and table 5.

TABLE 4. RESULTS ON SOGOU NEWS CORPUS

word segmentation model	accuracy rate(P)	recall rate(R)	F-measure
basic CRF model	91.32%	84.70%	87.89%
this work with 4-tags	92.76%	86.37%	89.45%
this work with 6-tags	93.51%	87.96%	90.65%

TABLE 5. RESULTS ON BLOG CORPUS

word segmentation model	accuracy rate(P)	recall rate(R)	F-measure
basic CRF model	90.37%	83.76%	86.94%
this work with 4-tags	91.45%	85.89%	88.58%
this work with 6-tags	91.96%	86.63%	89.22%

Clearly, the proposed model in this paper perform better both in 4-tags and 6-tags scheme, the accuracy rate, recall rate and F-measure is higher than basic CRF model. Intuitively, we think the domain dictionary helps a lot. Compared with 4-tags scheme, 6-tags scheme improves 0.75% in Sogou news corpus and 0.51% in blog corpus, because the longer tagging scheme show better prediction accuracy for the long domain term, so they obtain higher accuracy in long term segmentation. The experimental result show that the proposed gained competitive performance for domain word segmentation.

V. CONCLUSION

In this paper, we propose a statistical model for Chinese word segmentation. In our model, we firstly utilize Conditional Random Field to make rough segmentations for the sentence, to further improve the accuracy of result, we make use of two tag schemes: 4-tags scheme and 6-tags scheme. The basic CRF model is trained on common corpus to obtain the rough result, then we use C-value variable to develop domain dictionary to further improve the segmentation accuracy. The C-value firstly use statistical variable to extracts domain terms list, then the list will be filtered by manually designed rules to shape the final domain term dictionary. Experiments on small scale annotated new corpus and blog corpus show that the proposed model achieves competitive result on accuracy and recall rate.

REFERENCES

[1] Huang C, Zhao H. Chinese word segmentation: A Decade Review[J]. Journal of Chinese information processing. 2007, 21(3):8-19.

[2] Sun, Xu, et al. "Probabilistic Chinese word segmentation with non-local information and stochastic training." Information Processing & Management 49.3(2013):626-636.

[3] Chellappa, Rama, and A. Jain. "Markov random fields. Theory and application." -1(1993):242-261.

[4] Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models For Segmenting And Labeling Sequence Data[C]// 2001:282--289.

[5] Chen, Lei, et al. "A Double-layer Word Segmentation Combined with Local Ambiguity Word Grid and CRF." Transactions on Computer Science & Technology 2.1(2013):1-8.

[6] Zhao, Hai, and C. Kit. "Scaling Conditional Random Field with Application to Chinese Word Segmentation." International Conference on Natural Computation IEEE Computer Society, 2007:95-99.

[7] TONG Xiao Jun SONG Guo Long LIU Qiang ZHANG Li JIANG Wei School of Computer Science amp Technology, N. University, and Shenyang. "Research on the Model of Integrating Chinese Word Segmentation with Part-of-speech Tagging." Computer Science34.9(2007):174-177.

[8] Hong, Ming Cai. "A Chinese Part-of-speech Tagging Approach Using Conditional Random Fields." Computer Science 33.10(2006):148-151.

[9] Wei, Jiang. "Conditional Random Fields Based POS Tagging." Computer Engineering & Applications (2006).

[10] Zhang Z, Ren F, Zhu J. A Comparative Study of Features on CRF-based Chinese Named Entity Redognition[C], National Conference on information retrieval and Content Security, 2008

[11] Guo J. Research of Named Entity Recognition Based on Conditional Random Fields[D], Shen Yang Institute of Aeronautical Engineering. 2007.01

[12] Ji-Hong, L. I., et al. "Automatic Labeling of Semantic Roles on Chinese FrameNet." Journal of Software 28.21(2010):597-611.

[13] Song, Yijun, et al. "Semantic Role Labeling of Chinese FrameNet Based on Conditional Random Fields." Journal of Chinese Information Processing (2014).

[14] Fosler-Lussier, Eric, et al. "Conditional Random Fields in Speech, Audio, and Language Processing." Proceedings of the IEEE 101.5(2013):1054-1075.

[15] Chi, Chengying. "A Chinese Word Segmentation Approach Using Conditional Random Fields." Journal of Information (2008).

[16] "Review of Chinese Automatic Word Segmentation." Library & Information Service (2011).

[17] Li Q, Chen Y, Sun J. A new dictionary mechanism for Chinese word segmentation[J]. Journal of Chinese Information Processing, 2003, 17(4):13-18.

[18] Cao Y, Cao Y, Jin M, Liu C. Information retrieval oriented adaptive Chinese word segmentation system[J]. Journal of Software, 2006, 17(3):356-363.

[19] Chu Y, Liao M, Song J. Integrated Chinese words segmentation and labeling based on statistic method[J]. Computer Systems and Applications, 2009, 18(12):55-58.

[20] Zhang M, Deng Z, Che W, Liu T. Combining Statistical Model and Dictionary for Domain Adaption of Chinese Word Segmentation[J]. Journal of Chinese Information Processing, 2011:8-12.

[21] Liu Z, Ding D, Li C. Chinese word segmentation method for short Chinese text based on conditional random fields[J]. Journal Tsinghua University(Science & Technology), 2015(8):906-910.

[22] Xu H, Zhang Y, Yang X. Active Learning Based Domain Adaptation for Chinese Word Segmentation[J]. Journal of Chinese Information Processing, 2015, 29(5):55-62.

[23] Han D, Chang B. Approaches to domain adaptive Chinese segmentation model[J]. Chinese Journal of Computers, 2015, 38(2):272-281.

[24] Xiu C. The Research and Implementation of Method for Domain Chinese Word Segmentation[D], Beijing University of Technology. 2013.06

[25] Xue N. Chinese word segmentation as character tagging[J]. Computational Linguistics and Chinese Language Processing, 2003, 8(8:1):29-48.

[26] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International Journal on Digital Libraries, 2000, 3(2):115-130.

[27] Liang Y, Zhang W, Zhang Y. Term Recognition Based on Integration of C-value and Mutual Information[J]. Computer Applications and Software. 2010, 27(4):108-110.

[28] Li Chao, Wang H, Zhu M, Zhang L, Zhu J. Exploiting Domain Interdependence for Multi-Word Terms Extraction[J]. Journal of Chinese Information Processing, 2009:94-98.