Machine Learning Research
2017; 2(3): 105-109
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20170203.14

Chinese Word Segmentation Based on Conditional Random Field

Junxia Deng[1], Hong Zhang[2, 3, 4, 5, 6, 7, 8, 9, 10], Shanzai Li[2, 3, 4, 5, 6, 7, 8, 9, 10]

[1]International Economics and Trade, Gengdan Institute of Beijing University of Technology, Beijing, China
[2]School of Information, Beijing Wuzi University, Beijing, China
[3]Chinese Academy of Sciences, Bioinformatics Research Center, Beijing, China
[4]Chinese Academy of Sciences, Power System Research Center, Beijing, China
[5]Chinese Academy of Sciences, Partial Differential Equation and Its Application Center, Beijing, China
[6]Chinese Academy of Sciences, Statistical Science Research Center, Beijing, China
[7]Chinese Academy of Sciences, Center for Optimization and Applied Research, Beijing, China
[8]Chinese Academy of Sciences, Stochastic Analysis and Research Center, Beijing, China
[9]Chinese Academy of Sciences, Academy of Mathematics and Systems Science, Beijing, China
[10]School of Mathematical Sciences, Peking University, Beijing, China

Email address:
dr.yuwenjunxian@gmail.com (Junxia Deng)

To cite this article:
Junxia Deng, Hong Zhang, Shanzai Li. Chinese Word Segmentation Based on Conditional Random Field. *Machine Learning Research.* Vol. 2, No. 3, 2017, pp. 105-109. doi: 10.11648/j.mlr.20170203.14

Received: February 6, 2017; **Accepted**: February 27, 2017; **Published**: April 17, 2017

Abstract: This paper systematically describes the definition, model structure, parameter estimation and corpus selection of the conditional random field model, and applies the conditional random field to the Chinese word segmentation and the Chinese word segmentation method. In this paper, a large number of experiments have been carried out using conditional random fields. The experimental corpus has been tested by Changjiang Daily for many years. Experiments are carried out to analyze the influence of the choice of conditional random field model parameters and the selection of Chinese character annotation sets on the experimental results. Furthermore, the condition of random field model can be used to add the advantages of arbitrary features, and some new features are added to the model. Word probability, the paper explores the probability characteristic of word location. Experiments on the corpus show that the introduction of the word position probability feature has improved the accuracy, recall and the value of Fl.

Keywords: Natural Language Processing, Chinese Word Segmentation, Hidden Markov Model, Maximum Entropy Model, Conditional Random Field, Automatic Proofreading

1. Introduction

1.1. Annotation Method

The so-called Chinese word segmentation, is the process of word segmentation as each Chinese character classification process, by marking each character in the sentence to segmentation. For example, Xue N divides Chinese characters into four categories according to their different positions in Chinese words, and then divides them into Chinese characters by using the maximum entropy model. Peng F establishes a Chinese character segmentation model based on CRF. In addition to using some common features, But also used a lot of domain knowledge. Zhou J built a hybrid method of Chinese word segmentation around CRF model.

Common Chinese character tagging method is based on the Chinese characters appear in the words of different locations marked different labels. For example, "O" can be used to represent individual Chinese characters, "B" means Chinese characters appear in the head, "I" means Chinese characters appear in the middle or the end of the word. So the word segmentation problem is transformed into a pure sequence data labeling problem, you can use a lot of sequence tag algorithm for word segmentation. Therefore, Chinese word

segmentation method has become a frequently used method to study word segmentation.

Figure 1 is an example of the use of Chinese characters marked word segmentation. Enter the sentence as "This is Wuhan." We first set three candidate marks "O", "B" and "I" for each Chinese character and add a start node "BOS" to the head, An end node "EOS". The feature that appears on each node is then calculated, using the feature weights to compute the most probable of all paths from "BOS" to "EOS". In order to reduce the computation cost, we adopt some rules to eliminate some unnecessary paths. According to the meaning of "O", "B" and "I", the rules are summarized as follows:

1, the sentence of the first Chinese character tag can not be I, the last character of the mark can not be B. Because the mark B must be followed by a mark and only the mark I, and must be marked in front of the mark I, and marked as B or I.

2, marking O cannot appear behind the mark I, can only be O or B.

3, Mark B can only be followed by the mark I.

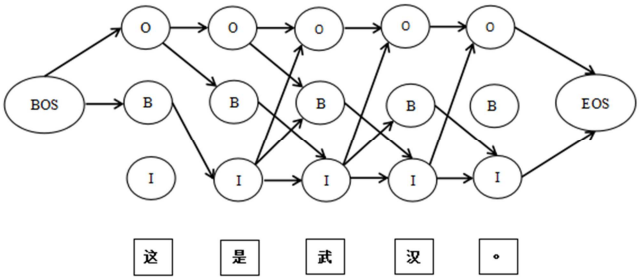

Figure 1. *Chinese characters marked word segmentation.*

Finally, according to the mark of each Chinese character, the mark result is "OOBIO", so the result is "这 / 是 / 武汉 /."

1.2. Feature Selection

One of the most important problems in using Chinese character annotation is feature selection. Although the CRF model can theoretically accommodate arbitrary features, even long-distance features, too many feature selections result in degraded system performance, so features are often extracted only in limited contexts. For example, you can take the context of the context of the two characters as a feature.

2. CRF Word Segmentation System

CRF word segmentation system is a word segmentation system based on conditional random field, and the Chinese word segmentation method is adopted.

2.1. Word System Process

CRF word segmentation process of Chinese word segmentation shown in Figure 2:

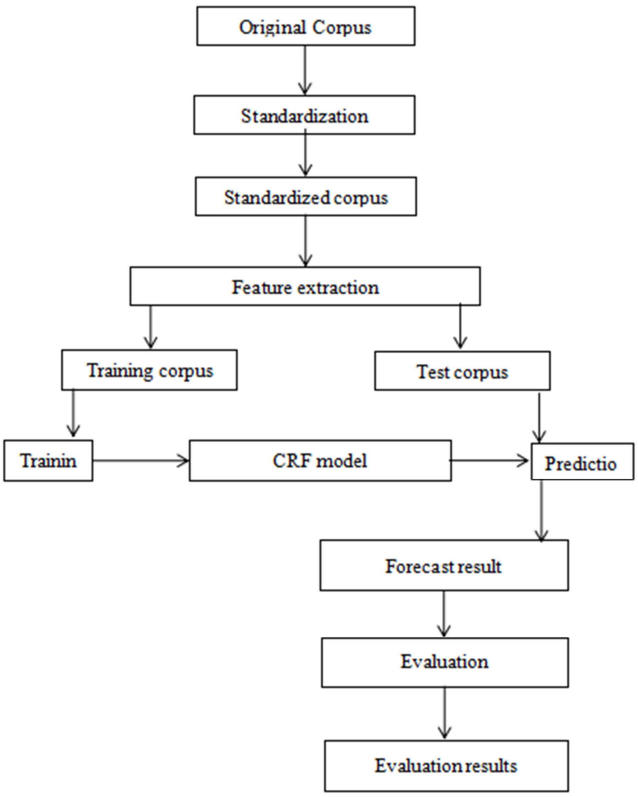

Figure 2. *CRF-based Chinese word segmentation process.*

In order to construct the CRF model, we must first use a standardized process to convert the original corpus into a standard form. The standard corpus form used here dictates that each line in the corpus contains only one word, and the information associated with the word is followed by a tab stop followed by the word. Secondly, feature extraction is carried out to generate training corpus and test corpus which can be recognized by CRF model tool. The format is as follows: each line includes a word and some characters and markings related to the word, characters and characteristics, And between the feature and the tag are separated by tabs. Then the training corpus is trained to generate a CRF model, and some training parameters such as iteration number are added in the training process. Using the CRF model generated by the training, the test corpus is tested and a prediction result is obtained. Finally, the evaluation program is used to evaluate the prediction results and get the evaluation results.

2.2. Feature Selection

Since the selection of features has a great impact on the results of CRF segmentation, it is a key issue to consider which features to choose. As mentioned earlier, the CRF makes it easy to add any feature in the observed sequence to the model, so that not only the transfer and emission characteristics of the traditional HMM sequence model can be incorporated into the model, but also some other The feature information associated with the observation sequence or with the language itself is added to the model.

The following Table 1, Table 2, Table 3 is used in the experiment some of the characteristics of the template.

Table 1. *Feature instances (basic feature templates).*

Numbering	Feature	Characteristic significance
1	W-2	The forward second word of the current word
2	W-1	The forward first word of the current word
3	W0	The current word
4	W1	Returns the first word of the current word
5	W2	Returns the second word of the current word
6	W-1W0	The first word of the current word, the current word
7	W0W1	Current word, currently the first word backwards
8	W-2W-1W0	The current forward two words, the current word
9	W-1W0W1	Forward one word, current word, backward one word
10	W0W1W2	Current word, current backward two words

Table 2. *Feature Examples.*

Numbering	Feature	Characteristic significance	For example
1	Ni	The Chinese character at position i is a number	0 1 2 3
2	Ci	Position i of the Chinese characters are Chinese numbers	一 二 三 四
3	Li	Position i of the Chinese characters are letters (including capitalization)	a b c A B C
4	Pi	The Chinese character at position i is a separate punctuation mark	、，。！《》？
5	Si	The Chinese character at position i is a non-individual word punctuation mark	@ %...
6	Ti	The character i in position i is the time word	year, month, day, hour, minute

Table 3. *Feature templates III (Word position probability feature template).*

Numbering	Feature	Characteristic significance	For example
1	X	The probability of the i-character position alone word: pX> 95%	又吗哦啊啥也
2	Y	The probability of the i-character of the position as the prefix: pY> 95%	昨狭抨第竞耽
3	Z	The probability of the i-character as the suffix: pZ> 95%	丸役袄侣丸瞰
4	R	The probability of the i-character individual word of the position: 85% <= pR <= 95%	吧枚刘磊躺却
5	U	Position of the i-character as a prefix of the probability: 85% <= pU<= 95%	参巩适农遗遵
6	V	The probability of the i-character as the suffix: 85% <= pV<= 95%	貌型胁帘午岸
7	D	The probability of the i-character position of the individual word: pD <= 5%	言辽改信申仪
8	E	The probability of the i-character of the position as the prefix: pE <= 5%	业络王场姆杨
9	F	The probability of the i-character as the suffix: pF<= 5%	增隐晋香浓伊

In order to deal with the long-distance information, this paper takes the context distance as 2. Table 2 and Table 3 are two new feature templates. The feature template in Table 2 is a classification of Chinese characters, the template well dealing with numbers, letters and punctuation marks such as error-prone not logged in the above table, the Table 1 feature template is a basic feature template, Use word-based.

The probabilities of the position probabilities in Table 3 are extracted from the training corpus and the probability of each position is calculated according to the following probability formula:

P (individual word) = number of occurrences of the individual idiom of the word / total number of occurrences of the word * 100% (1)

P (prefix) = number of occurrences of the word as a prefix / total number of occurrences of the word * 100% (2)

P (suffix) = number of occurrences of the word as a suffix / total number of occurrences of the word * 100% (3)

After many experiments and comparisons, the probability of more than 85% or probability of less than 5% of the word as a location probability feature, but also on the selected word for some filtering, the elimination of some unnecessary words, such as At the same time, select the probability of greater than or equal to 85% of the word is subdivided into the probability of 85% <p <95% and p> 95% of the two sets.

Appropriate increase of p> 95% of the number of occurrences is in order to improve the characteristics of the sample expectations, and achieved good word effect.

In this paper, CRF automatic word segmentation experiments, the use of features include the following two: a single word characteristic: a position on the word characteristics. For example, "the previous word is a number, the current word is a quantifier", "the second word of the current word is the number, the first word is the number, the current word is the number, the latter number is the number, Two words are time words "and other characteristics.

Combinations: A combination of 2 or 5 different word or string features at different positions. For example, "the previous word is a number, the current word is a quantifier", "the second word of the current word is the number, the first word is the number, the current word is the number, the latter number is the number, Two words are time words "and other characteristics.

3. Word Segmentation Experiment

3.1. The Choice of Experimental Corpus

The main corpus used in this paper is the training corpus and test corpus of Changjiang Daily. The corpus is from 1950 to 2005, and the scale is 2564168000 sentences. The coding method is GB code. The corpus content mainly comes from newspaper news. The format consists of a sentence segment consisting of words marked with spaces.

3.2. Experimental Evaluation Standard

The overall goal of the automatic Chinese word segmentation system is to establish an open, modern Chinese word segmentation system with high versatility and practicability. The performance of a word segmentation system mainly depends on two aspects: segmentation precision and word segmentation speed. Word segmentation accuracy, also known as word segmentation accuracy, is the core performance index system. The performance of Chinese automatic word segmentation is evaluated by the following three indexes: correct rate (P), recall rate (R) and F value. Where, P refers to the accuracy of word segmentation; R refers to the word recall rate; F value refers to the P and R integrated value. The formula is as follows:

Correct rate P = number of words correctly recognized / total number of system output words * 100% (4)

Recall rate IP correctly identify the number of words / test words in the total number of words * 100% (5)

F value F=*P*R/(P+R)*100％ (6)

The speed of word segmentation is another important index of word segmentation performance. The main factors that affect the speed of word segmentation are the structure of word segmentation dictionary and word segmentation algorithm. The query speed of the word dictionary depends on the organization structure of the dictionary. As the automatic word segmentation of the various types of knowledge to be obtained from the word dictionary, the system in the word processing needs frequent query word dictionary, word dictionary query speed will directly affect the speed of the word segmentation system. In different applications, the performance requirements of the word segmentation system have different emphases. In addition, automatic word segmentation system should also be easy to expand, maintainability and portability; to support different regions, different application areas of different application goals; vocabulary and processing functions, processing methods can be flexible combination of loading and unloading, thereby enhancing the system Processing precision and processing speed; also, to build a "information processing with the modern Chinese word segmentation standard" to match the common or common modern Chinese word segmentation.

3.3. CFR Performance Test of Word Segmentation System

CRF The word segmentation system is a word segmentation system based on the conditional random field, and uses the feature template one, the feature template two and the feature template three. Table 4 below shows the results of the CRF word segmentation system on the Yangtze River Daily Test Set and the contribution of each feature template to the results.

Table 4. *CRF system test results.*

Model	Correct rate	Recall rate	F-value
Feature Template 1	0.873	0.882	0.877
+Feature Template 2	0.902	0.899	0.900
+Feature Template 3	0.931	0.912	0.921

As can be seen from Table 4, with the characteristics of each template to join, the model test results gradually increased. The results of "+ feature template 3" model are obviously better than that of "+ feature template 2" model, that is, under the condition of adding feature template 3, F-score is 4.4% higher than that of feature template 3, Played a better effect. Moreover, since the word position probability feature is extracted completely from the training corpus, some of the participle criterion information of the corpus is extracted to a certain extent, so that when the test set is tested, Corpus.

3.4. CFR Comparison of Word Segmentation System with Other Models

In order to compare the availability of the conditional random field model, we also established two word segmentation models: Hidden Markov Model (HMM) segmentation model and Maximum Entropy (MEM) segmentation model. (CRF) word segmentation model, the experiment uses the combination of "feature template one", "feature template two" and "feature template three" in the common daily closed test set Test, the performance comparison of the results shown in Table 5.

Table 5. *The result of each word system.*

Model	Correct rate	Recall rate	F-value
Hidden Markov Model	0.915	0.896	0.905
Maximum entropy model	0.901	0.874	0.897
Conditional random field model	0.931	0.912	0.921

It can be seen from Table 5 that the results of the CRF system are better than those of other models under the same conditions as the training corpus and the test corpus. The experimental results show that conditional random field is an efficient segmentation method.

This chapter first briefly introduces the CRF tools, experiment corpus and standard of experimental evaluation in Chinese word segmentation experiments. And then use these tools and the corpus carried out a number of experiments. The experiment not only demonstrates the influence of the choice of conditional random field model parameters and Chinese character annotation set on the experimental results, but also

verifies the validity of the new features and the feasibility of the new method. In the experiment, only some feature information is used, and most of the features are extracted from the training corpus, we have achieved good results. As the Yangtze River Daily corpus is from the newspaper news, for the news corpus in the special format, such as title, poetry, weather forecasting, etc., to our model training has a certain impact, so if the corpus in the handling of these disturbances, Our model should have better performance.

Acknowledgments

I would like to thank Hong Zhang (ShanZai Li) for guiding me and in the process of writing this article.

References

[1] John Lafferty, Andrew McCallum, F Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th international Conference on Machine Leaning. San Francisco, USA. 2001: 282-289.

[2] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [A]. Processing of the International Conference on Machine Learning (ICML-2001) [C]. Williams college, MA, 2001: 282-289.

[3] Pinto D, McCallum A, Wei X et al. Table extraction using conditional random fields [A]. Proceedings of the 26th ACM SIGm [C], Toronto, Canada, 2003: 235-242.

[4] David Palmer A Trainable Rule-based Algorithm for Word Segmentation 1997.

[5] Berkeley, California, A new statistical formula for Chinese text segmentation incorporating contextual information. United States Pages: 82-89 Year of Publication: 1999.

[6] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition proceedings of The IEEE 77 (2): pp. ZS7-286, 1989.

[7] Zhou, GD., Su J. Named entity Recognition using all HMM-based chunk tagger. 2002.

[8] E. T. Jaynes. information Theory and Statistica Imeehanics. 1957.

[9] J. R. Crran and S. Clark Investigatigating GIS and Smoothing for Maximum Entropy Tggers. Proceedings of the llh Conference of the Europe Chapter of the Association of Computation Linguistics (EACL), Pages 91—98, Budapest, Hungary. 2003.

[10] Tan Y'Yao T, Chea Q ET al. Applying conditional random fields to Chinese shallow parsing

[11] Proceedings of CICLing-2005 [c], Mexico City, Mexico, 2005: 167-176.

[12] Kudo T, Yamamoto K, Matsumoto Y. Applying Conditional Random Fields to Japanese Morphological Analysis [A]. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004) [C], Barcelona, 2004: 230-237.

[13] Zhou J, Dai X, Ni R et al. A hybrid approach to Chinese word segmentation around CRFs [A]. Proceedings of the Fourth SIGHAN Workshop on Chinese language Processing [C], Jeju Island, Korea, 2005: 196-199.