HMM and CRF Based Hybrid Model for Chinese Lexical Analysis

Degen Huang, Xiao Sun, Shidou Jiao, Lishuang Li, Zhuoye Ding, Ru Wan
Dept. of Computer Science and Engineering, Dalian University of Technology.
Dalian, Liaoning, 116023
huangdg@dlut.edu.cn,suntian@gmail.com,jiaoshidou@gmail.com,
computer@dlut.edu.cn,dingzhuoye@sina.com,wanrulove@sina.com

Abstract

This paper presents the Chinese lexical analysis systems developed by Natural Language Processing Laboratory at Dalian University of Technology, which were evaluated in the 4th International Chinese Language Processing Bakeoff. The HMM and CRF hybrid model, which combines character-based model with word-based model in a directed graph, is adopted in system developing. Both the closed and open tracks regarding to Chinese word segmentation, POS tagging and Chinese Named Entity Recognition are involved in our systems' evaluation, and good performance are achieved. Especially, in the open track of Chinese word segmentation on SXU, our system ranks 1st.

1 Introduction

Chinese presents a significant challenge since it is typically written without separations between words. Word segmentation has thus long been the focus of significant research because of its role as a necessary pre-processing phase for the tasks above. Meanwhile, the POS tagging and Chinese Named Entity Recognition are also the basic steps in Chinese lexical analysis. Several promising methods are proposed by previous researchers. In tradition, the Chinese word segmentation technologies can be categorized into three types, rule-based, machine learning, and hybrid. Among them, the machine learning-based techniques showed excellent performance in many research studies (Peng et al., 2004; Zhou et al., 2005; Gao et al., 2004). This method treats the word segmentation problem as a sequence of word classification. The classifier online assigns either "boundary" or "non-boundary" label to each word by learning from the large annotated corpora. Machine learning-based word segmentation method is adopted in the word sequence inference techniques, such as part-of-speech (POS) tagging, phrases chunking (Wu et al., 2006a) and named entity recognition (Wu et al., 2006b). But there are some cost problems in such machine learning problems, and sometimes choose between word-based and character based is also a dilemma.

In our system, we present a hybrid model for Chinese word segmentation, POS tagging and named entity recognition based on HMM and CRF model. The core of the model is a directed segmentation graph based on the maximum matching and second-maximum matching model. In the directed graph, the HMM model and CRF model are combined, the HMM model is used to process the known words (words in system dictionary); CRF model is adopted to process the unknown word, the cost problem can be solved. Meanwhile, for the CRF model, the character-based CRF model and word-based model are integrated under the framework of the directed segmentation graph, so the integrative CRF model can be more flexible to recognize both the simple and complex Chinese Named Entity with high precision. With the directed segmentation graph, Chinese word segmentation, POS tagging and Chinese Named Entity recognition can be accomplished simultaneously.

2 System Description

With the maximum matching and second-maximum matching (MMSM) model, CRF model, and several post processing strategies, our systems are established. First the MMSM model is applied, based on the system dictionary the original directed segmentation graph is set up. The directed graph is composed by the known words from the system dictionary, which are regarded as the candidate word of the segmentation result. Then some candidate Chinese Named Entity Recognition automata search the directed graph, and find out the candidate Chinese Named Entities into the directed graph based on some generation rules. Then the CRF is applied to the candidate Chinese Named Entities to determine if they are real Chinese Named Entities that should be added into the directed graph. During this procedure, the character-based CRF and word-based CRF are respectively applied to the simple and complex Chinese Named Entities recognition.

In the following section, the Chinese word segmentation, POS tagging and Chinese named entity recognition in open track will be mainly discussed.

2.1 The maximum matching and second-maximum matching model

The maximum matching and second-maximum matching(MMSM) model, which is a segmentation method that keeps the maximum and second-maximum segmentation result from a certain position in a sentence, and store the candidate segmentation results in a directed graph, then some decoding algorithm is adopted to find the best path in the directed graph. With the MMSM model, almost all the possible segmentation paths and most lexical information can be reserved for further use; little space cost is guaranteed by using the directed graph to store the segmentation paths; the context spaces are extended from single-dimension to multi-dimension.

2.2 Conditional Random Fields

Conditional random field (CRF) was an extension of both Maximum Entropy Model (MEMs) and Hidden Markov Models (HMMs) that was firstly introduced by (Lafferty *et al*., 2001). CRF defined conditional probability distribution *P(Y|X)* of given sequence given input sentence where *Y* is the "class label" sequence and *X* denotes as the observation word sequence.

A CRF on *(X,Y)* is specified by a feature vector *F* of local context and the corresponding feature weight λ. The *F* can be treated as the combination of state transition and observation value in conventional HMM. To determine the optimal label sequence, the CRF uses the following equation to estimate the most probability.

Conditional random fields (CRFs) are undirected graphical models trained to maximize a conditional probability (Lafferty et al., 2001). A linear-chain CRF with parameters $\Lambda = \{\lambda_1, \lambda_2, \cdots\}$ defines a conditional probability for a state sequence $y = y_1 \ldots y_T$, given that and input sequence $x = x_1 \ldots x_T$ is

$$P_\Lambda(y \mid x) = \frac{1}{Z_x}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, x, t)\right)$$

Where Z_x is the normalization factor that makes the probability of all state sequences sum to one; $f_k(y_{t-1}, y_t, x, t)$ is ofen a binary-valued feature function and λ_k is its weight. The feature functions can measure any aspect of a state transition, $y_{t-1} \to y_t$, and the entire observation sequence, x, centered at the current time step, *t*. For example, one feature function might have the value 1 when y_{t-1} is the state *B*, y_t is the state *I*, and x_t is some Chinese character.

2.3 Chinese Named Entity Recognition

First, we will introduce our Chinese Named Entity Recognition part for the Open track. Several NER automata are adopted to find out all the candidate NEs in the directed graph, then the CRF model is applied to filter the candidate NEs to check if the specified NE should be added into the graph. To use the CRF, first, we generate some lists from the training corpus.

PSur: the surname of Person Name.

PC: the frequency information of a character in Person Name

PPre: the prefix of Person Name

PSuf: the suffix of Person Name

LF: the frequency information of a character in Local Name

LC: the centre character of Local Name

LPre: the prefix of Local Name

LSuf: the suffix of Local Name

OF: the frequency information of a character in ORG Name

OC: the centre character of ORG Name

OPre: the prefix of ORG Name

OSuf: the suffix of ORG Name

We define the template as follows:

PER: PSur(n)PC(n) PPre(n)PSuf(n), (n = -2, -1, 0, +1, +2)

LOC: LF(n)LC(n)LPre(n)LSuf(n), (n = -2, -1, 0, +1, +2)

ORG: OF(n)OC(n)OPre(n)OSuf(n), (n = -2, -1, 0, +1, +2)

With the CRF we filter the candidate NEs. The candidate NEs are filtered and added into the directed segmentation graph as new nodes with new edges. The NEs includes personal name(PRE), location name(LOC) and organization name(ORG).

The “PER”,”LOC” in open track is the same as in the close track except some external resources. The external resources include external lexicon, name list for word segmentation, and generating the features.

In the “ORG” part, a different method is proposed. We adopt an automatic recognition method of Chinese organization name with the combination of SVM and Maximum Entropy. SVM model is used to decide the latter boundary of a organization name, and then Maximum Entropy is used to confirm the former boundary.

First, a characteristic dictionary is collected from the training corpus. As for the words appeared in the characteristic dictionary, whether it is the characteristic word of an organization name should be decided. As a problem of two value categorization, SVM is applied to complete this task. If it is considered to be a characteristic word, then the former boundary of an organization name is detected. Maximum Entropy can combine different kinds of text information, and solve the problem of the recognition of the more complex former words of the Chinese organization name, so the Maximum Entropy is adopted to confirm the former boundary of ORG. During the NEs recognition and filtering the word and POS tag as main features and adopt a context window of five words.

Because of the complex construction of the Chinese Named Entity, one single statistical model can not solve simple and complex NER simultaneously, such as the character-based CRF model makes lower recognition accuracy for complex NERs, meanwhile, the word-based CRF model will lose many useful features in processing simple NERs. Integrating the character-based and word-based CRF model into one framework is the key to solve all the NERs simultaneously.

In this paper, an integrative model based on CRF is proposed. With the preliminary results of the segmentation and POS tagging, at the bottom of the system, character-based CRF is applied to recognized simple PERs, LOCs, and ORGs; The recognition result will be transformed to the top of the system together with the segmentation and POS tagging result. At the top of system, word-based CRF is used to recognize the nested LOCs and ORGs. The character-based model and word based model are integrated into one framework to recognition the NEs with different complexions simultaneously. The identification results of the bottom-level provide decision support for the high-level, the limitations of the separated character-based model and word-based model are avoided, and improves recognition accuracy of the system.

2.4 Result from the directed graph

After the recognition and filtering of the Chinese Named Entity, the original segmentation directed graph is now with the candidate Chinese Named Entity nodes. Some decoding algorithm is needed to find final path from the directed graph. Here, we revised the Dijkstra minimum cost path algorithm to find out the minimum cost path from the directed graph. The calculation of the cost of the nodes and edges in the directed graph can be found in our related work(Degen Huang and Xiao Sun, 2007). The final path from the directed graph is the result for the Chinese word segmentation, POS tagging and Chinese Named Entity recognition.

3 Evaluations and Experimental Results

3.1 Result of Chinese word segmentation

We evaluated our Chinese word segmentation model in the open track on all the simple Chinese corpus, such as University of Colorado, United States (CTB, 642246 tokens), State Language Commission of P.R.C.,Beijing(NCC, 917255 tokens) and Shanxi University, Taiyuan (SXU 528238 tokens). The OOV-rate is 0.0555, 0.0474 and 0.0512.

The CTB open track is shown in the following table 1. We get the third position in the CTB track by the F result.

Table 1. CTB open track result

CTB	R	P	F
Base	0.8864	0.8427	0.8640
Top	0.9710	0.9825	0.9767
Our	0.9766	0.9721	0.9743
	IV-R	IV-P	IV-F
Base	0.9369	0.8579	0.8956
Top	0.9698	0.9832	0.9764
Our	0.9805	0.9794	0.9800
	OOV-R	OOV-P	OOV-F
Base	0.9920	0.9707	0.9812
Top	0.0273	0.1858	0.0476
Our	0.9089	0.8553	0.8813

The NCC open track is shown in the following table 2. In the NCC open track, we get the third position track by the F result.

Table 2. NCC open track result

NCC	R	P	F
Base	0.9200	0.8716	0.8951
Top	0.9735	0.9817	0.9776
Our	0.9620	0.9496	0.9557
	IV-R	IV-P	IV-F
Base	0.9644	0.8761	0.9181
Top	0.9725	0.9850	0.9787
Our	0.9783	0.9569	0.9675
	OOV-R	OOV-P	OOV-F
Base	0.0273	0.1858	0.0476
Top	0.9933	0.9203	0.9554
Our	0.7109	0.7619	0.7355

The SXU open track is shown in the following table 3. In the SXU open track, we get the first two positions by the F result.

Table 3. NCC open track result

NCC	R	P	F
Base	0.9238	0.8679	0.8949
Top	0.9820	0.9867	0.9844
Our	0.9768	0.9703	0.9735
	IV-R	IV-P	IV-F
Base	0.9723	0.8789	0.9232
Top	0.9813	0.9890	0.9851
Our	0.9872	0.9767	0.9820
	OOV-R	OOV-P	OOV-F
Base	0.0251	0.0867	0.0389
Top	0.9942	0.9480	0.9705
Our	0.7825	0.8415	0.8109

We also participate in the close track in CTB, NCC and SXU corpus. The result is shown in the following table 4.

Table 4. Segmentation Result in close track

	R	P	F	Foov	Fiv
CTB	0.9505	0.9528	0.9517	0.7216	0.9659
NCC	0.9387	0.9301	0.9344	0.5643	0.9524
SXU	0.9594	0.9493	0.9543	0.6676	0.9697

3.2 Result of Chinese NER

We evaluated our named entity recognizer on the SIGHAN Microsoft Research Asia(MSRA) corpus in both closed and open track.

Table 5. NER in MSRA closed track:

Close	R	P	F
PER	90.29%	95.19%	92.68%
LOC	81.85%	92.78%	86.97%
ORG	70.16%	84.05%	76.48%
Overall	80.58%	91.07%	85.5%

Table 6. NER in MSRA open track:

Open	R	P	F
PER	92.06%	95.17%	93.59%
LOC	83.62%	94.24%	88.62%
ORG	74.04%	79.66%	75.65%
Overall	82.38%	90.38%	86.19%

3.3 Result of POS tagging

The POS tagging result of our system is shown in the following table 7.

Table 7. POS tagging in close track

Close	Total-A	IV-R	OOV-R	MT-R
CTB	0.9088	0.9374	0.4866	0.8805
NCC	0.9313	0.9604	0.4080	0.8809
PKU	0.9053	0.9451	0.2751	0.8758

Table 8. POS tagging in open track

Open	Total-A	IV-R	OOV-R	MT-R
CTB	91.2%	93.74%	53.61%	88.05%
NCC	93.26%	96.04%	43.36%	88.09%
PKU	93.29%	95.18%	63.32%	89.72%

4 Conclusions and Future Work

In this paper, the hybrid model in our system is described, An integrative lexical analysis system is implemented, which completes all the steps of the lexical analysis synchronously, by integrating the segmentation, ambiguous resolution, POS tagging, unknown words recognition into one theory framework. The integrative mechanism reduces the conflicts between the steps of the lexical analysis. The experimental results demonstrate that, the integrative model and its algorithm is effective. The system used the automata recognition and CRF-based hybrid model to process the Chinese Named Entity. The Chinese word segmentation, POS tagging and Chinese Named Entity recognition are integrated; the character-based CRF and word-based CRF are integrated, the HMM, CRF and other statistic model are integrated under the same segmentation framework. With this model we participated in the "The Fourth SIGHAN Bakeoff" and got good performance.

References

Degen, Huang and Xiao *An Integrative Approach to Chinese NamedEntity Recognition*, In Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology.

Gao, J., Wu, A., Li, M., Huang, C. N., Li, H., Xia, X., and Qin, H. 2004. *Adaptive Chinese word segmentation*. In Proceedings the 41st Annual Meeting of the Association for Computational Linguistics, pp. 21-26.

Lafferty, J., McCallum, A., and Pereira, F. 2001. *Conditional Random Field: Probabilistic models for segmenting and labeling sequence data*. In Proceedings of the International Conference on Machine Learning.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. *Text chunking using transformation-based learning.* In Proceedings of the 3rd Workshop on Very Large Corpora, pages 82-94. Nocedal, J., and Wright, S. 1999. Numerical optimization. Springer.

Peng, F., Feng, F., and McCallum, A. 2004. *Chinese segmentation and new word detection using conditional random fields*. In Porceedings of the Computational Linguistics, pp. 562-568.

Shi, W. 2005. *Chinese Word Segmentation Based On Direct Maximum Entropy Model*. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

Wu, Y. C., Chang, C. H. and Lee, Y. S. 2006a. *A general and multi-lingual phrase chunking model based on masking method*. Lecture Notes in Computer Science (LNCS): Computational Linguistics and Intelligent Text Processing, 3878: 144-155.

Wu, Y. C., Fan, T. K., Lee Y. S. and Yen, S. J. 2006b. *Extracting named entities using support vector machines*," Lecture Notes in Bioinformatics (LNBI): Knowledge Discovery in Life Science Literature, (3886): 91-103.

Wu, Y. C., Lee, Y. S., and Yang, J. C. 2006c. *The Exploration of Deterministic and Efficient Dependency Parsing*. In Proceedings of the 10th Conference on Natural Language Learning (CoNLL).