A Study of Chinese Lexical Analysis Based on Discriminative Models

Guang-Lu Sun Cheng-Jie Sun Ke Sun and Xiao-Long Wang
Intelligent Technology & Natural Language Processing Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, 150001, Harbin, China
{glsun, cjsun, ksun, wangxl}@insun.hit.edu.cn

Abstract

This paper briefly describes our system in The Fourth SIGHAN Bakeoff. Discriminative models including maximum entropy model and conditional random fields are utilized in Chinese word segmentation and named entity recognition with different tag sets and features. Transformation-based learning model is used in part-of-speech tagging. Evaluation shows that our system achieves the F-scores: 92.64% and 92.73% in NCC Word Segmentation close and open tests, 89.11% in MSRA name entity recognition open test, 91.13% and 91.97% in PKU part-of-speech tagging close and open tests. All the results get medium performances on the bakeoff tracks.

1 Introduction

Lexical analysis is the basic step in natural language processing. It is prerequisite to many further applications, such as question answer system, information retrieval and machine translation. Chinese lexical analysis chiefly consists of word segmentation (WS), name entity recognition (NER) and part-of-speech (POS) tagging. Because Chinese does not have explicit word delimiters to mark word boundaries like English, WS is essential process for Chinese. POS tagging and NER are just like those of English.

Our system participated in The Fourth SIGHAN Bakeoff which held in 2007. Different approaches are applied to solve all the three tasks which are integrated into a unified system (ITNLP-IsLex). For WS task, conditional random fields (CRF) are used. For NER, maximum entropy model (MEM) is applied. And transformation-based learning (TBL) algorithm is utilized to solve POS tagging problem. The reasons using different models are listed in the rest sections of this paper. We give a brief introduction to our system sequentially. Section 2 describes WS. Section 3 and section 4 introduce NER and POS tagging respectively. We give some experimental results in section 5. Finally we draw some conclusions.

2 Chinese word segmentation

For WS task, NCC corpus is chosen both in close test and open test.

2.1 Conditional random fields

Conditional random fields are undirected graphical models defined by Lafferty (2001). There are two advantages of CRF. One is their great flexibility to incorporate various types of arbitrary, non-independent features of the input, the other is their ability to overcome the label bias problem.

Given the observation sequence *X*, on the basis of CRF, the conditional probability of the state sequence Y is:

$$p(Y/X) = \frac{1}{Z(X)} exp\left\{\sum_k l_k f_k(y_{i-1}, y_i, X, i)\right\} \quad (1)$$

$$Z(X) = \sum_{y \in Y} exp\left\{\sum_k l_k f_k(y_{i-1}, y_i, X, i)\right\} \quad (2)$$

Z(x) is the normalization factor. $f_k(y_{i-1}, y_i, X, i)$ is the universal definition of features in CRF.

2.2 Word segmentation based on CRF

Inspired by Zhao (2006), the Chinese WS task is considered as a sequential labeling problem, i.e., assigning a label to each character in a sentence given its contexts. CRF model is adopted to do labeling.

6 tags are utilized in this work: B, B1, B2, I, E, S. The meaning of each tag is listed in Table 1. The

raw training file format from NCC can be easily to convert to this 6 tags format.

An example: 向/S 广/B 东/B1 省/B2 高/I 级/I 人/I 民/I 法/I 院/E 提/B 出/E 上/B 诉/E 。/S.

Table 1 Tags of character-based labeling

Tag	Meaning
B	The 1st character of a multi-character word
B1	The 2nd character of a multi-character word
B2	The 3rd character of a multi-character word
I	Other than B, B1, B2 and last character in a multi-character word
E	The last character of a multi-character word
S	Single character word

The contexts window size for each character is 5: C_{-2}, C_{-1}, C_0, C_1, and C_2. There are 10 feature templates used to generate features for CRF model including uni-gram, bi-gram and tri-gram: C_{-2}, C_{-1}, C_0, C_1, C_2, $C_{-1}C_0$, C_0C_1, $C_{-2}C_{-1}C_0$, $C_{-1}C_0C_1$, and $C_0C_1C_2$.

For the parameters in CRF model, we only do work to choose cut-off value for features. Our experiments show that the best performance can be achieved when cut-off value is set to 2.

Maximum likelihood estimation and L-BFGS algorithm is used to estimate the weight of parameters in the training module. Baum-Welch algorithm is used to search the best sequence of test data.

For close test, we only used CRF to do segmentation, no more post-processing, such as time and date finding, was done. So the performance could be further improved.

For open test, we just use our NER system to tag the output of our close segmentation result, no more other resources were involved.

3 Chinese name entity recognition

For NER task, MSRA is chosen in open test. Chinese name dictionary, foreign name dictionary, Chinese place dictionary and organization dictionary are used in the model.

3.1 Maximum entropy model

Maximum entropy model is an exponential model that offers the flexibility of integrating multiple sources of knowledge into a model (Berger, 1996). It focuses on the modeling of tagging sequence, replacing the modeling of observation sequence.

Given the observations sequence *X*, on the basis of MEM, the conditional probability of the state sequence Y is:

$$p(Y \mid X) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j f_j(Y, X) \right) \quad (3)$$

$$Z(X) = \sum_Y \exp\left(\sum_j \lambda_j f_j(Y, X) \right) \quad (4)$$

Table 2 Feature templates of NER

Feature template		Description
C_i		The word tokens in the window i =-2, -1, 0, 1, 2
T_i		The NE tags i = -1
C_iC_{i-1}		The bigram of C_i i = -1, 1
P_i		The POS tags of word tokens i = -1, 0, 1
$P_{-1}P_1$		The combination of POS tags
$T_{-1}C_0$		The previous tag and the current word token
$W(C_i)$	B	C_i is Chinese family name
	C	C_i is part of Chinese first name
	W	C_i is Chinese whole name
	F	C_i is foreign name
	S	C_i is Chinese first name
	O	other
$W(C_{i-1})W(C_i)$		The bigram of W(Ci) i = -1, 1
IsInOrgDict(C_0)		The current word token is in organization dictionary
IsInPlaceDict(C_0)		The current word token is in place dictionary

Being Similar to the definition of CRF, Z(x) is the normalization factor. $f_j(Y, X)$ is the universal definition of features.

3.2 Name entity recognition based on MEM

Firstly, we use a segmentation tool to split both training and test corpus into word-token-based texts. Characters that are not in the dictionary are scattered in the texts. NE tags using in the model follow the tags in training corpus. Other word tokens that do not belong to NE are tagged as *O*. Based on the segmented text, the context window is also set as 5. Inspired by Zhang's (2006) work, there are 10 types of feature templates for generating features for NER model in Table 2.

When training our ME Model, the best performance can be achieved when cut-off value is set to 1.

Maximum likelihood estimation and GIS algorithm is used to estimate the weight of parameters in the model. The iteration time is 500.

4 Chinese part-of-speech tagging

For POS tagging task, NCC corpus and PKU corpus are chosen both in the close test and open test.

4.1 Transformation-based learning

The formalism of Transformation-based learning is first introduced in 1992. It starts with the correctly tagged training corpus. A baseline heuristic for initial tag and a set of rule templates that specify the transformation rules match the context of a word. By transformating the error initial tags to the correct ones, a set of candidate rules are built to be the conditional pattern based on which the transformation is applied. Then, the candidate rule which has the best transformation effect is selected and stored as the first transformation rules in the TBL model. The training process is repeated until no more candidate rule has the positive effect. The selected rules are stored in the learned rule sequence in turn for the purpose of template correction learning.

4.2 Part-of-speech tagging based on TBL

POS tagging is a standard sequential labeling problem. CRF has some advantages to solve it. Because both corpora have relative many POS tags, our computational ability can not afford the CRF model in condition of these tags. TBL model is utilized to replace with CRF.

We compute the max probability of current word's POS tag in training corpus. The POS tag which has max occurrence probability for each word is used to tag its word token. By this method, we got the initial POS tag for each word.

The rule templates which are formed from conjunctions of words match to particular combinations in the histories of the current position. 40 types of rule templates are built using the patterns. The cut-off value of the transformation rules is set to 3 (Sun, 2007).

For open test, our NER system is used to tag the output of our POS tagging result. Parts of NE tags are corrected.

5 Evaluation

Following the measurement approach adopted in SIGHAN, we measure the performance of the three tasks in terms of the precision (P), recall (R), and F-score (F).

5.1 Word segmentation results

Table 3 Word segmentation results on NCC corpus

NCC	close test	open test
R	.9268	.9268
C_r	.00133447	.00133458
P	.926	.928
C_p	.00134119	.00132534
F	.9264	.9273
R_{oov}	.6094	.6265
P_{oov}	.4948	.5032
F_{oov}	.5462	.5581
R_{iv}	.9426	.9417
P_{iv}	.9527	.9546
F_{iv}	.9476	9481

The WS results are listed on the Table 3. Some errors could be caused by the annotation differences between the training data and test data. For example, "阿珍" (A Zhen) was considered as a whole word in training data, while "阿兰" (A Lan) was annotated as two separate word "阿" (A) and "兰" (Lan) in the test data. Some post-processing rules for English words, money unit and morphology can improve the performance further, Following are such errors in our results: "vid eo",

"日 元" (Japan yen), "不 三 不 四" (not three not four).

For open test, we hoped to use NER module to increase the OOV recall. But the NER module didn't prompt the performance very much because it was trained by the MSRA NER data in Bakeoff3. The difference between two corpora may depress the NER modules effect. Also, the open test was done on the output of close test and all the errors were passed.

5.2 Name entity recognition results

The official results of our NER system on MSRA corpus for open track are showed in Table 4. As it shows, our system achieves a relatively high score on both PER and LOC task, but the performance of ORG is not so good, and the Avg1 performance is decreased by it. The reasons are: (1) The ORG sequences are often very long and our system is unable to deal with the long term, a MEMM or CRF model may perform better. (2) The resource for LOC and ORG are much smaller than that of PER. More sophisticated features such like "$W(C_i)$" may provide more useful information for the system.

Table 4 NER results on MSRA corpus

MSRA	P	R	F
PER	.9498	.9549	.9524
LOC	.9129	.9194	.9161
ORG	.8408	.7469	.7911
Avg1	.9035	.8791	**.8911**

5.3 Part-of-speech tagging results

We evaluate our POS tagging model on the PKU corpus for close and open track and NCC corpus for close track based on TBL. Table 5 is the official result of our system. In PKU open test, NER is used to recognize name entity of text, so its result is better than that of close test. The IV-R result is relative good, but the OOV-R is not so good, which drops the total performance. The reasons lie in: (1) TBL model is not good at tagging out of vocabulary words. CRF model may be a better selection if our computer can meet its huge memory requirements. (2) Our NER system is trained by MSRA corpus. It does not fit the PKU and NCC corpus.

Table 5 POS results on PKU and NCC corpus

Corpus	Total-A	IV-R	OOV-R	MT-R
PKU close test	.9113	.9518	.2708	.8958
PKU open test	.9197	.9512	.4222	.899
NCC close test	.9277	.9664	.2329	.9

6 Conclusions

Chinese lexical analysis system is built for the SIGHAN tracks which consists of Chinese word segmentation, name entity recognition and part-of-speech tagging. Conditional random fields, maximum entropy model and transformation-based learning model are utilized respectively. Our system achieves the medium results in all the three tasks.

References

A. Berger, S. A. Della Pietra and V. J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 1996. 22(1), pages 39-71.

G. Sun, Y. Guan and X. Wang. A Maximum Entropy Chunking Model With N-fold Template Correction. *Journal of Electronics*, 2007. 24(5), pages 690-695.

J. Lafferty, A. McCallum and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of ICML-2001*, Williams College, Massachusetts, USA. 2001. pages 282-289.

S. Zhang, Y. Qin, J. Wen, X. Wang. Word Segmentation and Named Entity Recognition for SIGHAN Bakeoff3. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Sydney, Australia. 2006. pages 158–161.

H. Zhao, C. Huang, and M. Li. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia. 2006. pages 162–165.