An Improved CRF based Chinese Language Processing System for SIGHAN Bakeoff 2007

Xihong Wu, Xiaojun Lin, Xinhao Wang, Chunyao Wu, Yaozhong Zhang and **Dianhai Yu**
Speech and Hearing Research Center
State Key Laboratory of Machine Perception,
Peking University, China, 100871
{wxh,linxj,wangxh,wucy,zhangyaoz,yudh}@cis.pku.edu.cn

Abstract

This paper describes three systems: the Chinese word segmentation (WS) system, the named entity recognition (NER) system and the Part-of-Speech tagging (POS) system, which are submitted to the Fourth International Chinese Language Processing Bakeoff. Here, Conditional Random Fields (CRFs) are employed as the primary models. For the WS and NER tracks, the n-gram language model is incorporated in our CRFs based systems in order to take into account the higher level language information. Furthermore, to improve the performances of our submitted systems, a transformation-based learning (TBL) technique is adopted for post-processing.

1 Introduction

Among 24 closed and open tracks in this bakeoff, we participated in 23 tracks, except the open NER track of MSRA. Our systems are ranked 1st in 6 tracks, and get close to the top level in several other tracks.

Recently, Maximum Entropy model(ME) and CRFs (Low et al., 2005)(Tseng et al., 2005) (Hai Zhao et al., 2006) turned out to be promising in natural language processing tracks, and obtain excellent performances on most of the test corpora of Bakeoff 2005 and Bakeoff 2006. Compared to the generative models, like HMM, the primary advantage of CRFs is that it relaxes the independence assumptions, which makes it able to handle multiple interacting features between observation elements (Wallach et al., 2004).

However, the ME and CRFs emphasize the relation of the basic units of sequence, like the Chinese characters in these tracks. While, the higher level information, like the relationship of the words is ignored. From this point of view, the n-gram language model is incorporated in our CRFs based systems in order to cover the word level language information.

Based on several pilot-experimental results, we found that the tagging errors always follow some patterns. In order to find those error patterns and correct the similar errors, we integrated the TBL post-processor in our systems. In addition, extra training data, which is transformed from People Daily Corpus (Shiwen Yu et al., 2000) with some auto-extracted transition rules, is used in each corpus for the open tracks of WS.

The remainder of this paper is organized as follows. The scheme of our three developed systems are described in section 2, 3 and 4, respectively. In section 5, evaluation results based on these systems are enumerated and discussed. Finally some conclusions are drawn in section 6.

2 Word Segmentation

The WS system mainly consists of three components, CRFs, n-gram language model and post-processing strategies.

2.1 Conditional Random Fields

Conditional Random Fields, as the statistical sequence labeling models, achieve great success in natural language processing, such as chunking (Fei Sha et al., 2003) and word segmentation (Hai Zhao et al., 2006). Different from traditional generative

model, CRFs relax the constraint of the independence assumptions, and therefore turn out to be more suitable for natural language tasks.

CRFs model the conditional distribution $p(Y|X)$ of the labels Y given the observations X directly with the formulation:

$$P_\lambda(Y|X) = \frac{1}{Z(X)} exp\{\sum_{c\in C}\sum_{k}\lambda_k f_k(Y_c, X, c)\} \quad (1)$$

Y is the label sequence, X is the observation sequence, $Z(X)$ is a normalization term, f_k is a feature function, and c is the set of cliques in Graphic.

In our tasks, $C = \{(y_{i-1}, y_i)\}$, X is the Chinese character sequence of a sentence.

To label a Chinese character, we need to define the label tags. Here we have six types of tags according to character position in a word (Hai Zhao et al., 2006):

$$tag = \{B_1, B_2, B_3, I, E, S\}$$

"B_1, B_2, B_3, I, E" represent the first, second, third, continue, and end character positions in a multi-character word, and "S" is the single-character word tag.

The unigram feature templates used here are:

$$C_n\ (n = -2, -1, 0, 1, 2)$$
$$C_nC_{n+1}\ (n = -2, -1, 0)$$
$$C_nC_{n+1}C_{n+2}\ (n = -1)$$

Where C_0 refers to the current character and $C_{-n}(C_n)$ is the *n*th character to the left(right) of the current character. We also use the basic bigram feature template which denotes the dependency on the previous tag and current tag.

2.2 Multi-Model Integration

In order to integrate multi-model information, we use a log-linear model(Och et al., 2002) to compute the posterior probability:

$$\begin{aligned} Pr\quad (W|C) &= p_{\alpha_1^M}(W|C) \\ &= \frac{exp[\sum_{m=1}^{M}\alpha_m h_m(W,C)]}{\sum_{W'} exp[\sum_{m=1}^{M}\alpha_m h_m(W',C)]} \end{aligned} \quad (2)$$

Where W is the word sequence, and C is the character sequence. The decision rule here is:

$$\begin{aligned} W_0 &= argmax_W\{Pr(W|C)\} \\ &= argmax_W\{\sum_{m=1}^{M}\alpha_m h_m(W,C)\} \end{aligned} \quad (3)$$

The parameters α_1^M of this model can be optimized by standard approaches, such as the Minimum Error Rate Training used in machine translation (Och, 2003). In fact, the CRFs approach is a special case of this framework when we define $M = 1$ and use the following feature function:

$$h_1(W, C) = logP_\lambda(Y|X) \quad (4)$$

In our approach, the logarithms of the scores generated by the two kinds of models are used as feature functions:

$$\begin{aligned} h_1(W,C) &= logP_{crf}(W,C) \\ &= log\prod_{w_i} P_\lambda(w_i|C) \end{aligned} \quad (5)$$

$$h_2(W,C) = logP_{lm}(W) \quad (6)$$

The first feature function(Eq.5) comes from CRFs. Instead of computing the score of the whole label sequence Y with character sequence X through $P_\lambda(Y|X)$ directly, we try to get the posterior probability of a sub-sequence to be tagged as one whole word $P_\lambda(w_i|C)$. Then we combine all the score of words together. The second feature function(Eq.6) comes from n-gram language model, which aims to catch the words information.

The log-linear model with the feature functions described above allows the dynamic programming search algorithm for efficient decoding. The system generates the word lattice with posterior probability $P_\lambda(w_i|C)$. Then the best word sequence is searched on the word lattice with the decision rule(Eq.3).

Since arbitrary sub-sequence can be viewed as a candidate word in word lattice, we need to deal with the problem of OOV words. The unigram of an OOV word is estimated as:

$$Unigram(OOV\ Word) = p^l \quad (7)$$

where p is the minimal value of unigram scores in the language model; l is the length of the OOV word, which is used as a punishment factor to avoid overemphasizing the long OOV words (Xinhao Wang et al., 2006).

2.3 Post-Processing Strategies

The division and combination rule, which has been proved to be useful in our system of Bakeoff 2006 (Xinhao Wang et al., 2006), is adopted for the post-processing in the system.

2.4 Training Data Transition

For the WS open tracks, the unique difference from closed tracks is that the additional training data is supplemented for model refinement.

For the Simplified Chinese tracks, the additional training data are collected from People Daily Corpus with a set of auto-extracted transition rules. This process is performed in a heuristic strategy and contains five steps as follows:
(1) Segment the raw People Daily texts with the corresponding system for the closed track of each corpus.
(2) Compare the result of step 1 with People Daily Corpus to get the conflict pairs. For example,

{*pair1: 江泽民 vs. 江泽 民*}
(Zhemin Jiang)
{*pair2: 两手抓 vs. 两 手 抓*}
(catch with two hands)

In each pair, the left phrase follows the People Daily Corpus segmentation guideline, while the right one is the phrase obtained from step 1.
(3) Divide the pairs into two sets: **the first set** contains the pairs with right phrase appearing in the target training data; the other pairs are in **the second set**.
(4) Select sentences which contain the left phrase of the pairs in **the second set** from People Daily Corpus.
(5) Transform these selected sentences by replacing their phrase in the left side of the pair in **the first set** to the right one. This is used as our transition rules.

3 Named Entity Recognition

The named entity recognition track is viewed as a character sequence tagging problem in our NER system and the log-linear model mentioned above is employed again to integrate multi-model information. To find the error patterns and correct them, a TBL strategy is then used in the post-processing module.

3.1 Model Description

In this NER track, we employe the log-linear model and use the logarithms of the scores generated by the two types of models as feature functions. Besides CRFs, another model is the class-based n-gram language model:

$$h_1(Y,X) = logP_{crf}(Y,X) = logP_\lambda(Y|X) \quad (8)$$

$$h_2(Y,X) = logP_{clm}(Y,X) \quad (9)$$

Y is the label sequence and X is the character sequence.

CRFs are used to generate the N-best tagging results with the scores of whole label sequence Y on character sequence X by $P_\lambda(Y|X)$. And then, the log-linear model is used to reorder the N-best tagging results by integrating the CRFs score and the class-based n-gram language model score together.

CRFs

In this track, one Chinese character is labeled by a tag of ten classes, which denoting the beginning, continue, ending character of a specified named entity or a non-entity character. There are three types of named entities in these tracks, including person name, location name and organization name.

In CRFs, the basic features used here are:

$$C_n\ (n = -2, -1, 0, 1, 2)$$
$$C_nC_{n+1}\ (n = -2, -1, 0, 1)$$
$$C_nC_{n+2}\ (n = -1)$$

Besides basic unigram features, the bigram transition features considering the previous tag is adopted with template $C_n\ (n = -2, -1, 0, 1, 2)$.

Class-Based N-gram Language Model

For the class-based n-gram language model, we define that each character is a single class, while each type of named entity is viewed as a single class. With the character sequence and label sequence, the class sequence can be generated. Take this sentence for instance:

但伊卜拉依莫夫并不满足
(But Ibrahimov is not satisfied)

Table 1 shows its class sequence. Class-based n-gram language model can be trained with class sequence.

3.2 TBL

Since the analysis on our experiments shows that the tagging errors always follow some patterns in NER track, TBL strategy is adopted in our system to find these patterns and correct the similar errors.

character sequence	但	伊	卜	拉	依	莫	夫	并	不	满	足
label sequence	N	Per-B	Per-C	Per-C	Per-C	Per-C	Per-E	N	N	N	N
class sequence	但	*PERSON*						并	不	满	足

Table 1: A class sequence example

Transformation-based learning is a symbolic machine learning method, introduced by (Eric Brill, 1995). The main idea in TBL is to generate a set of transformation rules that can correct tagging errors produced by the initial process.

There are four main procedures in our TBL framework: An initial state assignment which is operated by the system we described above; a set of allowable templates for rules, ranging from words in a 3 positions windows and name entity information in a 3-word window with their combinations considered, and rules which are learned according to the tagging differences between training data and results generated by our system, at last, those rules are introduced to correct similar errors.

4 POS Tagging

The POS tagging track is to assign the part-of-speech sequence for the correctly segmented word sequence. In our system, for the CTB corpus, the CRFs are adopted; however for the other four corpora, considering the limitations of resources and time, the ME model is adopted. To improve the performance of ME model, the POS tag of the previous word is taken as a feature and the dynamic programming strategy is used in decoding.

In the closed track, the features include the basic features and their combined features. Firstly the previous and next words of the current word are taken as the basic features. Secondly, based on the analysis of the OOV words, the first and last characters of the current word, as well as the length of the current word are proven to be effective features for the OOV POS. Furthermore since the long distance constraint word may impact the POS of current word (Yan Zhao et al., 2006), in the open track, a Chinese parser is imported and the word depended on the current word is extracted as feature.

5 Experiments and Results

We have participated in 23 tracks, except the open NER track of MSRA. CRFs, ME model and n-gram language model are adopted in these systems. Our implementation uses the CRF++ package[1] provided by Taku Kudo, the Maximum Entropy Toolkit[2] provided by Zhang Le, and the SRILM Toolkit provided by Andreas Stolcke (Andreas Stolcke et al., 2002).

5.1 Chinese Word Segmentation

In the closed tracks, CRFs and bigram language model are trained on the given training data for each corpus. In order to integrate these two models, it is necessary to train the corresponding parameter α_1^M with Minimum Error Rate Training approache based on a development data. Since the development data is not provided in this bakeoff, a ten-fold cross validation approach is employed to implement the parameter training. A set of parameters can be trained independently, and then the mean value is calculated as the estimation of each parameter.

Table 2 gives the results of our WS system for closed tracks.

	baseline	+LM	+LM+Post
CTB	94.7	94.7	94.8
NCC	92.6	92.4	92.9
SXU	94.7	95.7	95.8
CITYU	92.9	93.7	93.9
CKIP	93.2	93.7	93.7

Table 2: Word segmentation performance on F-value with different approach for the closed tracks

In the open tracks, as we do not have enough time to finish the parameter estimation on the new data, our system adopt the same parameters α_1^M used in closed tracks. The unique difference from closed

[1] http://chasen.org/taku/software/CRF++

[2] http://homepages.inf.ed.ac.uk/s0450736/maxent toolkit.html

tracks is that extra training data is added for each corpus to improve the performance. For the Simplified Chinese tracks, additional data comes from People Daily Corpus which is transformed by our transition strategy. At the same time, for the Traditional Chinese tracks, additional data comes from the training and testing data used in the early Bakeoff. However, we implement two systems for the CTB open track. The system (a) takes the training and testing data used in the early Bakeoff as additional data, and System (b) takes the translated People Daily Corpus as additional data. Table 3 gives the results of our open WS system.

	baseline	+LM	+LM+Post
CTB(a)	99.2	99.2	99.3
CTB(b)	95.6	95.1	97.0
NCC	93.7	93.0	92.9
SXU	96.4	87.0	95.8
CITYU	95.8	90.6	91.0
CKIP	94.5	94.8	95.1

Table 3: Word segmentation performance on F-value with different approach for the open tracks

The result shows that the system performance is sensitive to the parameters α_1^M. Although we train the useful parameter for closed tracks, it plays a bad role in open tracks as we do not adapt it for the additional training data.

5.2 Named Entity Recognition

In the closed NER tracks, CRFs and class-based trigram language model are trained on the given training data for each corpus. The same approach employed in the WS tracks is adopted to train the corresponding parameter α_1^M in our NER systems. Meanwhile, the TBL rules trained via five-fold cross validation approach are also used in post-processing procedure. Table 4 reports the results of our closed NER system.

5.3 POS Tagging

The experiments show that the CRFs/ME method is superior to the TBL method, and the concurrent errors for these two methods are less than 60%. Therefore we adopted TBL to correct the output results of CRFs/ME: If the output tags of CRFs/ME and TBL are not consistent and the output probability of CRFs/ME is below a certain threshold, the TBL results are fixed. Here the 90% of the training set is taken as the training data and remained 10% is separated as the development data to get the threshold, which is 0.60 for the CRFs, and 0.90 for the ME. In addition, the POS tagged corpus of the Chinese Treebank 5.0 from LDC is added to the training data for CTB open track. In our system, the Berkeley Parser (Slav Petrov et al., 2006) is adopted to obtain the long distance constraint words. The performance achieved by the methods described above on each corpus are reported in Table 5.

	baseline	+LM	+LM+Post
MSRA	89.3	89.7	89.9
CITYU	79.3	80.6	80.5

Table 4: Named entity recognition F-value through different approaches for the closed tracks

	CRFs/ME	TBL	CRFs/ME +TBL	CRFs/ME +TBL +Syntax
CTIYU	88.7	87.7	89.1	89.0
CKIP	91.8	91.4	92.2	92.1
CTB	94.0	92.7	94.3	96.5
NCC	94.6	94.3	94.9	95.0
PKU	93.5	93.2	94.0	94.1

Table 5: POS tagging performance on total-accuracy with different approach

6 Conclusion

In this paper, we have briefly described our systems participating in the Bakeoff 2007. In the WS and NER systems, the log-linear model is adopted to integrate CRFs and language model, which improves the system performances effectively. At the same time, system integration approach used in the POS system also proves its validity. In addition, a heuristic strategy is imported to generate additional training data for the open WS tracks. Finally, several post-processing strategies are used to further improve our systems.

References

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.* pp. 161-164. Jeju Island, Korea.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, Christopher Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.* pp. 168-171. Jeju Island, Korea.

Hai Zhao, Chang-Ning Huang and Mu Li. 2006. An Improved Chinese Word Segmentation System with Conditional Random Field. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing.* pp. 162-165. Sydney, Australia.

Hanna M. Wallach. 2004. Conditional Random Fields: An Introduction. *Technical Report, UPenn CIS TR MS-CIS-04-21.*

Shiwen Yu, Xuefeng Zhu and Huiming Duan. 2000. Specification of large-scale modern Chinese corpus. *Proceedings of ICMLP'2001.* pp. 18-24. Urumqi, China.

Fei Sha and Fernando Pereira. 2003. Shallow Parsing with Conditional Random Fields. *Proceedings of Human Language Technology/NAACL.* pp. 213-220. Edmonton, Canada.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* pp. 295-302. Philadelphia, PA.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL).* pp. 160-167. Sapporo, Japan.

Xinhao Wang, Xiaojun Lin, Dianhai Yu, Hao Tian, Xihong Wu. 2006. Chinese Word Segmentation with Maximum Entropy and N-gram Language Model. *the Fifth SIGHAN Workshop on Chinese Language Processing.* pp. 138-141. Sydney, Australia.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: a case study in Part-of-Speech tagging. *Computational Lingusitics.* 21(4).

Yan Zhao, Xiaolong Wang, Bingquan Liu, and Yi Guan. 2006. Fusion of Clustering Trigger-Pair Features for POS Tagging Based on Maximum Entropy Model. *Journal of Computer Research and Development.* 43(2). pp. 268-274.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. *Proceedings of International Conference on Spoken Language Processing.* pp. 901-904. Denver, Colorado.

Slav Petrov, Leon Barrett, Romain Thibaux and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL.* pp. 433-440. Sydney, Australia.