Chinese Word Segmentation and Named Entity Recognition Based on Conditional Random Fields

Xinnian Mao
France Telecom R&D Center (Beijing), Beijing, 100080, P.R.China
xinnian.mao@orange-ftgroup.com

Saike He
University of Posts and Telecommunications, Beijing, 100876, P.R.China

Sencheng Bao
University of Posts and Telecommunications, Beijing, 100876, P.R.China

Yuan Dong[1,2]
[1]France Telecom R&D Center (Beijing), Beijing, 100080, P.R.China
[2]University of Posts and Telecommunications, Beijing, 100876, P.R.China
yuan.dong@orange-ftgroup.com

Haila Wang
France Telecom R&D Center (Beijing), Beijing, 100080, P.R.China
haila.wang@orange-ftgroup.com

Abstract

Chinese word segmentation (CWS), named entity recognition (NER) and part-of-speech tagging is the lexical processing in Chinese language. This paper describes the work on these tasks done by France Telecom Team (Beijing) at the fourth International Chinese Language Processing Bakeoff. In particular, we employ Conditional Random Fields with different features for these tasks. In order to improve NER relatively low recall; we exploit non-local features and alleviate class imbalanced distribution on NER dataset to enhance the recall and keep its relatively high precision. Some other post-processing measures such as consistency checking and transformation-based error-driven learning are used to improve word segmentation performance. Our systems participated in most CWS and POS tagging evaluations and all the NER tracks. As a result, our NER system achieves the first ranks on MSRA open track and MSRA/CityU closed track. Our CWS system achieves the first rank on CityU open track, which means that our systems achieve state-of-the-art performance on Chinese lexical processing.

1 Introduction

Different from most European languages, there is no space to mark word boundary between Chinese characters, so Chinese word segmentation (CWS) is the first step for Chinese language processing. From another point that there is no capitalization information to indicate entity boundary, which makes Chinese named entity recognition (NER) more difficult than European languages. And part-of-speech tagging (POS tagging) provides valuable information for deep language processing such as parsing, semantic role labeling and etc. This paper presents recent research progress on CWS, NER and POS tagging done by France Telecom Team (Beijing). Recently, Conditional Random Fields[1] (CRFs) (Lafferty et al., 2001) have been successfully employed in various natural language processing tasks and achieve the state-of-the-art performance, in our system, we use it as the basic framework and incorporate some other post-processing measures for CWS, NER and POS tagging tasks.

2 Chinese Named Entity Recognition

NER is always limited by its lower recall due to the imbalanced distribution where the NONE class dominates the entity classes. Classifiers built on such dataset typically have a higher precision and a lower recall and tend to overproduce the NONE

[1] We use the CRF++ V4.5 software from http://chasen.org/~taku/software/CRF++/

class (Kambhatla, 2006). Taking SIGHAN Bakeoff 2006 (Levow, 2006) as an example, the recall is lower about 5% than the precision for each submitted system on MSRA and CityU closed track. If we could improve NER recall but keep its relatively high precision, the overall F-measure will be improved as a result. We design two kinds of effective features: 0/1 features and non-local features to achieve this objective. Our final systems utilize these features together with the local features to perform NER task.

2.1 Local Features

The local features are character-based and are instantiated from the following temples:
Unigram: C_n (n=-2,-1, 0, 1, 2).
Bigram: C_nC_{n+1} (n=-2,-1, 0, 1) and $C_{-1}C_1$.

Where C_0 is the current character, C_1 the next character, C_2 the second character after C_0, C_{-1} the character preceding C_0, and C_{-2} the second character before C_0.

2.2 0/1 Features

In order to alleviate the imbalanced class distribution, we assign 1 to all the characters which are labeled as entity and 0 to all the characters which are labeled as *NONE* in training data. In such way, the class distribution can be alleviated greatly, taking Bakeoff 2006 MSRA NER training data for example, if we label the corpus with 10 classes, the class distribution is 0.81(B-PER):1.70(B-LOC):0.95(B-ORG):0.81(I-PER):0.88(I-LOC):2.87(I-ORG):0.76(E-PER):1.42(E-LOC):0.94(E-ORG):88.86(NONE), if we change the label scheme to 2 labels (0/1), the class distribution is 11.14 (entity):88.86(*NONE*). We train the 0/1 CRFs tagger using the local features alone. For the 0/1 features, during the training stage, they are assigned with 2-fold cross validation, and during the testing stage, they are assigned with the 0/1 tagger.

2.3 Non-local Features

Most empirical approaches including CRFs currently employed in NER task make decision only on local context for extract inference, which is based on the data independent assumption. But often this assumption does not hold because non-local dependencies are prevalent in natural language (including the NER task). How to utilize the non-local dependencies is a key issue in NER task. Up to now, few researches have been devoted to this issue; existing works mainly focus on using the non-local information for improving NER label consistency (Krishnan and Manning, 2006). There are two methods to use non-local information. One is to add additional edges to graphical model structure to represent the distant dependencies and the other is to encode the non-locality with non-local features. In the first approach, heuristic rules are used to find the dependencies (Bunescu and Mooney, 2004) or penalties for label inconsistency are required to handset ad-hoc (Finkel et al., 2005). Furthermore, high computational cost is spent for approximate inference. In order to establish the long dependencies easily and overcome the disadvantage of the approximate inference, Krishnan and Manning (2006) propose a two-stage approach using CRFs framework with extract inference. They represent the non-locality with non-local features, and extract them from the output of the first stage CRF with local context alone; then they incorporate the non-local features into the second CRF. But the features in this approach are only used to improve label consistency in European languages. Similar with their work encoding the non-local information with non-local feature, and we also exploit the non-local features under two-stage architecture. Different from their features are activated on the recognized entities coming from the first CRF, the non-local features we design are used to recall more missed entities which are seen in the training data or unseen entities but some of their occurrences being recognized correctly in the first stage, so our non-local features are activated on the raw character sequence.

Different NER in European languages, where entity semantic classification is more difficult compared with boundary detection, in Chinese, the situation is opposite. So we encode different useful information for Chinese NER two subtasks: entity boundary detection and entity semantic classification. Three kinds of non-local features are designed; they are fired on the token sequences if they are matched with certain entity in the entity list in forward maximum matching (FMM) way.

Token-position features (NF1): These refer to the position information (start, middle and last) assigned to the token sequence which is matched with the entity list exactly. These features enable us to capture the dependencies between the identical candidate entities and their boundaries.

Entity-majority features (NF2): These refer to the majority label assigned to the token sequence which is matched with the entity list exactly. These features enable us to capture the dependencies between the identical entities and their classes, so that the same candidate entities of different occurrences can be recalled favorably, and their label consistencies can be considered too.

Token-position & entity-majority features (NF3): These features capture non-local information from NF1 and NF2 simultaneously. They take into account the entity boundary and semantic class information at the same time.

Figure 1 shows the flow of using non-local features under CRFs framework in two-stage architecture. The first CRF is trained with local features alone, and then we test the testing data with the first CRF and get the entities plus their type from the output. The second CRF utilizes the 0/1 features and the non-local features derived from the entity list which is merged by the output of the first CRF from the testing data and the entities extracted directly from the training data. We compare the three kinds of non-local features on MSRA and CityU closed track in SIGHAN 2006 and we find that the NF3 is the best (Mao etc, 2007). So we only incorporate the NF3 into our final NER system.

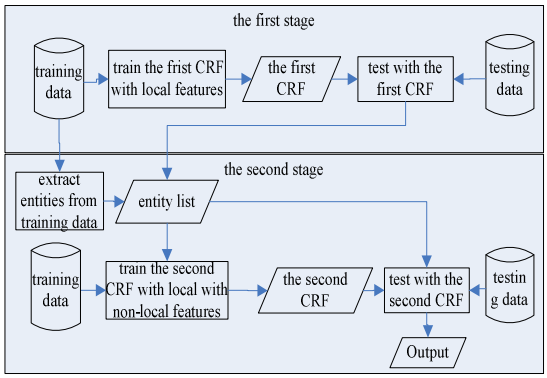

Figure 1. The flow using non-local features in two-stage architecture

2.4 Results

We employ BIOE1 label scheme for the NER task because we found it performs better than IOB2 on Bakeoff 2006 (Levow, 2006) NER MSRA and CityU corpora. Table 1 presents the official results on the MSRA and CityU corpus. The F-measure on MSRA open track is so high just because the testing data in Bakeoff 2007 is part of its Bakeoff 2006 training dataset and we utilize this corpus for training the final CRFs classifier. The F-measure on CityU open track is not much superior to its closed track because we only use its Bakeoff 2006 corpus to train the 0/1 CRFs, but not use the Bakeoff 2006 corpus to train final classifier.

Run ID	F-Score	Run ID	F-Score
cityu_c	84.99	cityu_o	87.92
msra_c	92.81	msra_o	99.88

Table 1: The official results on NER closed(c) tracks and open(o) tracks

3 Chinese Word Segmentation

Type	Feature
Unigram	C_n (n=-2,-1, 0, 1, 2).
Bigram	C_nC_{n+1} (n=-2,-1,0, 1)
Jump	$C_{-1}C_1$
Punc	Pu (C_0)
Date, Digit, Letter	$T_{-1}T_0T_1$

Table 2: The features used in our CWS systems

Table 2 lists the features we used in our CWS systems. After the raw corpus is processed by CRFs, two other post-processing measures are performed. We utilize transformation-based error-driven learning (TBL)[2] to further improve CWS and perform consistency checking among different occurrences of a particular character sequence. For TBL, we use the template defined in (He et al.). Our CWS system participate almost all the tracks and table 3 lists the official results.

Run ID	F-Score	Run ID	F-Score
cityu_c_a	94.43	cityu_o_a	96.97
cityu_c_b	error (94.31)	cityu_o_b	96.86
ckip_c_a	93.17	ckip_o_a	93.25
ckip_c_b	93.06	ckip_o_b	93.64
ctb_c_a	94.86	ctb_o_a	97.93
ctb_c_b	94.74	ctb_o_b	97.28
ncc_c_a	92.99	sxu_c_a	95.46
ncc_c_b	92.89	sxu_c_b	95.17

Table 3: The official results on CWS closed(c) tracks and open(o) tracks

In the table 3, run (a) means that we only perform consistency checking; run (b) means that

[2] We use the TBL software from http://nlp.cs.jhu.edu/~rflorian/fntbl/index.html

TBL is performed after consistency checking is done. We make a mistake on cityu_c_b because we rename cityu_c_a as cityu_c_b, so the two results are the same, after we correct the mistake and score again; we achieve an F-measure of 94.31%.

In the closed tracks, we first train initial CRFs with 3-fold cross-validation; then we test the training data (three parts) with the three trained CRFs, we train the TBL learner on the training data compared it with the testing result from the initial CRFs. The consistency checking is inspired by (Ng and Low, 2004). Table 4 lists the corpus used to train the CRFs and TBL learner in the open tracks.

	CRFs	TBL
CityU	2005,2006,2007	2003
CKIP	2007	2006
CTB	2006,2007	2007

Table 4. Corpora used to train the CRFs classifier and the TBL learner

In the open track, we collect the consistency list from all its correspondent Bakeoff corpora, the gazetteer extract from People Daily 2000 and idioms, slang from GKB. From the table 3 in the closed test, we can confirm that TBL may not improve CWS performance, while in most cases, performance will surely draw back. The reason lies in the fact that the learning capability of CRFs is superior to that of TBL, if they are trained with the same corpus, TBL may modify some correctly tags by CRFs. This can be seen from Table 3 that results without TBL (in run (a)) are almost superior to that with TBL (in run (b)).

4 Part-of-speech Tagging

For POS tagging task, apart from the local features same as used in NER, two other features are designed to improve the performance.

- Ambiguous part-of-speech: this feature is true when the word has more than 2 kinds of part-of-speech.
- Major part-of-speech: The feature is assigned as the major part-of-speech for any word. We do not assign the value to the new words.

Table 5 shows the performance in the closed tracks. Because we only used the simple features and do not process the unknown word specially, our performance is not satisfactory.

Run ID	F-Score	Run ID	F-Score
cityu_c	87.93	ctb_c	92.03
ckip_c	87.93	ncc_c	91.72
ctb_c	92.03		

Table 5: The official results on POS tagging in closed tracks

References

R. Bunescu and R. J. Mooney. 2004. Collective Information Extraction with Relational Markov Networks. In *Proceedings of the 42nd ACL*, 439–446.

J. Finkel, T. Grenager, and C. D. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 42nd ACL*, 363–370.

Nan He, Xinnian Mao, Yuan Dong, Haila Wang, 2007. Transformation-based Error-driven Learning as Post-processing for Chinese Word Segmentation, In *Proceedings of the 7th International Conference on Chinese Computing*, 46-51, Wuhan, China.

N. Kambhatla. 2006. Minority Vote: At-Least-N Voting Improves Recall for Extracting Relations. In *Proceeding of the 44th ACL*, 460–466.

V. Krishnan and C. D Manning. 2006. An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. In *Proceedings of the 44th ACL*, 1121–1128.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th ICML*, 282–289, San Francisco, CA.

G. Levow. 2006. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In *Proceedings of SIGHAN-2006*, 108-117. Sydney, Australia.

Xinnian Mao, Xu Wei, Yuan Dong, Saike He and Haila Wang, 2007. Using Non-local Features to Improve Named Entity Recognition Recall, In *Proceedings of the 21th Pacific Asia Conference on Language, Information and Computation*, 303-310, Seoul, Korea.

Hwee Tou Ng, Jin Kiat Low, 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All at Once? Word-based or Character based? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Spain.