France Telecom R&D Beijing Word Segmenter for Sighan Bakeoff 2006

Wu Liu
France Telecom R&D Beijing
wu.liu@francetelecom.com

Heng Li
France Telecom R&D Beijing
heng.li@francetelecom.com

Yuan Dong
Beijing University of Posts and Telecommunications
yuandong@bupt.edu.cn

Nan He
Beijing University of Posts and Telecommunications
hn.ft.pris@gmail.com

Haitao Luo
Northeastern University of China
luoht@ics.neu.edu.cn

Haila Wang
France Telecom R&D Beijing
haila.wang@francetelecom.com

Abstract

This paper presents two word segmentation (WS) systems and a named entity recognition (NER) system in France Telecom R&D Beijing. The one system of WS is for open tracks based on n-gram language model and another one is for closed tracks based on maximum entropy approach. The NER system uses a hybrid algorithm based on Class-based language model and rule-based knowledge. These systems are all augmented with a set of post-processors.

1 Introduction

The FTRD team participated in MSRA Open, MSRA Closed and CityU Closed tracks of the WS bakeoff and MSRA Open track of the NER bakeoff, and achieved the state-of-the-art performance in these tracks. Analysis of the results shows that each component of these systems contributed to the scores.

2 System Description

2.1 MSRA Open track of WS

The system used in open track of WS is based on the system (Li 2005) participated in the second international WS bakeoff. We mainly modify the factoid detection rules and add the GKB (The Grammatical Knowledge-base of Contemporary Chinese) dictionary. The system also has a few postprocessors. The main postprocessors include named entity recognizers and TBL (Transformation-Based Learning) component.

2.1.1 Basic system

In our basic system, Chinese words can be categorized into one of the following types: lexicon words, morphological words, factoids, name entities. These types of words were processed in different ways in our system, and were incorporated into a unified statistical framework of the trigram language model. The details about the basic system are reported in (Li 2005).

2.1.2 Factoid detection

The factoid rules used in the basic system were summarized according to the MSRA training data. The Tokenization Guidelines of Chinese Text (V5.0) was provided by MSRA in this bakeoff. We used the Guidelines to rewrite the factoid rules, and the performance had the distinct improvement.

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 122–125,
Sydney, July 2006.

2.1.3 Named entity identification

The named entity recognizer is the one participated in the NER bakeoff, as shown in figure 1. In the section **2.3**, we will describe in detail.

2.2 System Used in Close tracks

The system used in closed tracks of WS is based on maximum entropy approach. The system also has a few postprocessors. The main postprocessors include combining the separated words and TBL component.

2.2.1 Basic system

The basic system is similar to (Ng and Low, 2004). We used the Tsujii laboratory maximum entropy package v2.0 (http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/) to train our models. For CityU closed track, the basic features are the same as (Ng and Low, 2004). For MSRA closed track, we used two sets of basic features. The one is similar to (Ng and Low, 2004) and we change the window size of another one from 2 to 3, so we trained two models for MSRA closed track and submitted two results.

2.2.2 Post processing

Firstly, we extracted one lexicon from each training data. For MSRA closed track, the postprocessor only combined the words which appeared in the lexicon but were separated in the test result. For CityU closed track, we firstly used the factoid tool provided by the open system of WS to combine the separated factoid words, and then we used the lexicon to combine the separated words, at last the TBL was applied to the test result.

2.3 MSRA Open track of NER

The system used a hybrid algorithm which can combine a class-based statistical model (Gao 2004) with various types of rule-based knowledge very well. All the words were categorized into three types: Lexicon words (LWs), Factoid words (FTs), Named Entity (NEs). Accordingly, three main components were included to identify each kind of named entities: basic word candidates, NE combination and Viterbi search, as shown in Figure 1.

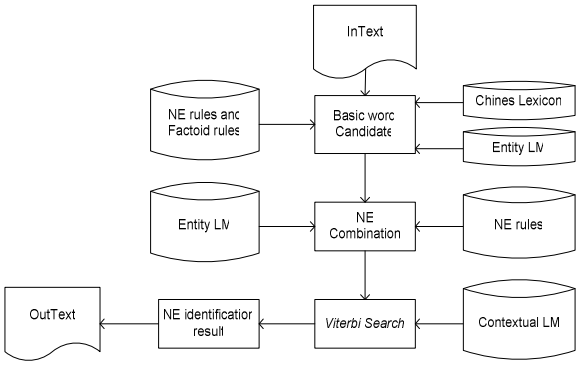

Figure 1 FTRD NE Recognizer

The recognizer was applied to open track of WS and we used it to participate in the MSRA open track of NER. The system also had a TBL post-processor.

2.4 TBL

In our system, the open source toolkit fnTBL (http://nlp.cs.jhu.edu/~rflorian/fntbl/index.html) is chosen. Coping with word segmentation task, we utilized a method called "LMR" tagging which was the same as (Nianwen Xue and Libin Shen 2003). Two rule template sets were used in our system. The complicated one had 40 templates, which covered various kinds of words position and tag position occurrence, i.e., considering contextual information of words and tags. For example, rule "pos_0 word_0 word_1 word_2 => pos" could generate rules containing information about current word, current word's tag, the next word and the word after next. The other rule template neglected tag information, it took only contextual word information into account. For an instance, "word_0 word_1 word_2 => pos". The task of WS applied the two rule template sets, and the task of NER only applied the complicated one. In the **Section 3**, we will compare the two rule template sets.

3 Evaluation

3.1 Open tracks

3.1.1 MSRA Open track of WS

In this open track, we used one lexicon of 294,382 entries, which included the entries of 42,430 MDWs (Morphological Derived Words) generated from the GKB dictionary, 12,487 PNs, 22,907 LNs and 29,032 ONs, 10,414 four-character idioms, plus the word lists generated from the training data provided by the second international Chinese Word Segmentation bakeoff and 80114 GKB words. We also used the training data provided by the last bakeoff for training our trigram word-based language model.

Table 1 presents the results of this track. For comparison, we also include in the table (Row 1) the results of basic system. From Row 2 to Row 11, it shows the relative contribution of each component and resource to the overall word segmentation performance. The second column shows the recall, the third column the precision, and the fourth column F-score. The last two columns present the recall of the OOV words and the recall of IV words, respectively.

(%)	R	P	F	R_{oov}	R_{iv}
1.basic system	0.971	0.958	0.964	0.590	0.984
2.1+new factoid	0.966	0.958	0.962	0.642	0.978
3.1+GKB lexicon	0.975	0.966	0.971	0.716	0.984
4.3+new factoid	0.971	0.967	0.969	0.768	0.978
5. 4+NE	0.971	0.973	0.972	0.838	0.975
6. 5+TBL	0.977	0.976	0.977	0.840	0.982
7.5+new TBL	0.980	0.978	0.979	0.839	0.985
8. 4+TBL	0.977	0.970	0.974	0.769	0.984
9.4+new TBL	0.980	0.971	0.975	0.769	0.987
10. 8+NE	0.977	0.976	0.977	0.840	0.982
11. 9+NE	0.979	0.978	0.979	0.841	0.984

Table1: Our system results on Open tracks

From Table 1 we can find that, in Row 1, the basic system participated in the last bakeoff already achieves quite good recall, but the recall of OOV is not very good because it cannot correctly identify unknown words that are not in the lexicon such as factoids and name entities (especially the nested named entity) and new words (except factoids, named entities and words abstracted from training data). In Row 2, we only rewrite the factoid rules according to the MSRA Guidelines, and the recall of OOV improves significantly while the recall of IV falls slightly. It shows that the factoid detection affects the recall of IV. As shown in Table 1, the GKB lexicon has made significant and persistent progress in all performance because the GKB lexicon is refined and the words are conformed to the MSRA standard. We also find that the NE postprocessor can improve the recall of OOV but affects slightly the recall of IV in all experiments. It shows that our named entity recognition has make improvement compared with that of last year. As shown in Table 1, TBL has made slightly but persistent progress in all steps it applies to. After TBL adaptation OOV recall stays almost unchanged, for the rules are derived from training corpus, and no OOV words would meet the condition of applying them in theory, but IV recall improves, which compensates the loss of IV recall caused by NE post-process and the factoid detection. It is interesting comparing the performance of two TBL template sets, the first template set is simple and the threshold for generating rules is 3 by default (called TBL in Table 1), and the second is more complicated with a "0" threshold (called New TBL in Table 1). The number of rules generated is 1061 and 12135 respectively. Our experiments demonstrate that more precise rule template set with low threshold always leads to better performance, for they could cover more situations, although a simple rule template set with high threshold does better in OOV word recognition.

3.1.2 MSRA Open track of NER

In the track, we used People's Daily 2000 corpus (Yu, 2003) for building our lexicon and training our model.

Considering that organization names are irregular in their forms compared with person names and location names, and there are many abbreviations and anaphora, TBL adaptation may degrade the performance of organization, we submitted two results, as shown in Table 2. 1+TBL1 means that TBL only adapt person and location results of basic system, the organization performance of basic system and 1+TBL1 would be identical. 1+TBL2 means TBL adapt all three types of NE. For comparison, we list (Column 2) the results of basic system. The Row 2 to Row 13 shows the recall, the precision, and the F-score of PN, LN, ON and total.

(%)		1.basic	1+TBL1	1+TBL2
PN	R	87.28	91.43	91.74
	P	90.63	92.56	92.77
	F	88.92	91.99	92.25
LN	R	80.18	87.39	89.74
	P	81.68	87.51	89.77
	F	80.92	87.45	89.76
ON	R	65.59	65.59	76.48
	P	73.80	73.80	75.44
	F	69.45	69.45	76.11
Total	R	79.31	83.99	87.53
	P	82.98	86.45	87.67
	F	81.10	85.20	87.60

Table 2: MSRA Open track of NER

To our surprise, performance listed in Table 2 demonstrates that applying TBL causes a dramatic improvement in all three types of NE, especially organization performance. The great similarity between training corpus and test corpus of MSRA may explain this. For the inconsistency of standard between MSRA and PKU, the recall, especially of the ONs, is not very good. We did some effort in the standard adaptation, such as constraint the length and type of candidate words in combining the named entities, but the result is not very good.

3.2 Closed tracks

The Table 3 and Table 4 present the results of MSRA and CityU closed tracks respectively.

(%)	R	P	F	R_{oov}	R_{iv}
1.basic system(2)	0.924	0.877	0.900	0.575	0.936
2.1+training lexicon	0.955	0.953	0.954	0.575	0.969
3.2+TBL	0.960	0.955	0.958	0.575	0.973
4.basic system(3)	0.919	0.880	0.899	0.602	0.930
5.4+training lexicon	0.950	0.954	0.952	0.602	0.962
6.5+TBL	0.954	0.955	0.955	0.603	0.966

Table 3: Our system results on MSRA Closed

(%)	R	P	F	R_{oov}	R_{iv}
1.basic system	0.947	0.916	0.931	0.716	0.957
2.1+training lexicon	0.959	0.960	0.959	0.716	0.969
3.2+TBL	0.969	0.964	0.967	0.716	0.980
4.1+factoid tool	0.946	0.915	0.931	0.713	0.956
5.4+training lexicon	0.958	0.959	0.959	0.713	0.968
6.5+TBL	0.969	0.964	0.966	0.712	0.980
6'	0.962	0.962	0.962	0.722	0.972

Table 4: Our system results on CityU Closed

In Table 3, the basic system (2) shows the window size of the template is 2 and the basic system (3) is 3. As is shown in the table, except the precision and the recall of OOV, the performance of window size with 2 outperforms that of window size with 3.

In Table 4, the system 6' is the one we submitted in this closed CityU track, but the system 6 is better than the system 6'. In TBL training, we made a mistake that the training data weren't processed by factoid tool and lexicon combining. We also can find that the factoid tool doesn't improve the performance. The system 6 isn't the best one (system 3).

Combining the separated words according to training lexicon improved the performance of both MSRA and CITYU closed track. In the meantime, TBL worked considerably well in all closed tracks.

4 Conclusions

The evaluation results show that the performance of NER need be improved in abbreviations recognition and anaphora resolution.

Acknowledgements

The work reported here was a team effort. We thank Yonggang Xue, Duo Ji, Haitao Luo, Nan He and Xinnian Mao for their help in the experimentation and evaluation of the system. We also thank Prof. Shiwen Yu for the People's Daily 2000 corpus (Yu 2003) and GKB (Yu 2002) lexicon.

References

Heng Li, etc. 2005. Chinese Word Segmentation in FTRD Beijing. Proceedings of the Fourth SIGHAN workshop on Chinese Language Processing. Pages:150-154

Hwee Tou Ng, Jin Kiat Low. 2004. Chiense part-of-speech tagging: One-at-a-time or all-at-once? Word-based or character-based?. Proceedings of the 2004 conference on Empirical Methods in Natural Language Processing. Pages:277-284

Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2004a. Chinese word segmentation: a pragmatic approach. Microsoft Research Technical Report, MSR-TR-2004-123.

Nianwen Xue, Libin Shen. July 2003. Chinese word segmentation as LMR tagging. Proceedings of the Second SIGHAN workshop on Chinese Language Processing. Pages:176-179.

Shiwen Yu, etc. 2003. Specification for Corpus Processing at Peking University:Word Segmentation, POS Tagging and Phonetic Notation. Journal of Chinese Language and Computing, 13(2) 121-158.

Shiwen Yu, etc. 2002. The Grammatical Knowledge-base of Contemporary Chinese --- A Complete Specification. Tsinghua University Press.