BUPT Systems in the SIGHAN Bakeoff 2007

Ying Qin Caixia Yuan Jiashen Sun Xiaojie Wang
Center of Intelligent Science and Technology Research
Beijing University of Posts and Telecommunications
Beijing, 100876, China
qinyingmail@163.com, yuancx@gmail.com,
b.bigart911@gmail.com, xjwang@bupt.edu.cn

Abstract

Chinese Word Segmentation(WS), Name Entity Recognition(NER) and Part-Of-Speech(POS) are three important Chinese Corpus annotation tasks. With the great improvement in these annotations on some corpus, now, the robustness, a capability of keeping good performances for a system by automatically fitting the different corpus and standards, become a focal problem. This paper introduces the work on robustness of WS and POS annotation systems from Beijing University of Posts and Telecommunications(BUPT), and two NER systems. The WS system combines a basic WS tagger with an adaptor used to fit a specific standard given. POS taggers are built for different standards under a two step frame, both steps use ME but with incremental features. A multiple knowledge source system and a less knowledge Conditional Random Field (CRF) based systems are used for NER. Experiments show that our WS and POS systems are robust.

1 Introduction

In the last SIGHAN bakeoff, there is no single system consistently outperforms the others on different test standards of Chinese WS and NER standards(Sproat and Emerson, 2003). Performances of some systems varied significantly on different corpus and different standards, this kind of systems can not satisfy demands in practical applications. The robustness, a capability of keeping good performances for a system by automatically fitting the different corpus and standard, thus become a focal problem in WS and NER, it is the same for Chinese Part-of-Speech(POS) task which is new in the SIGHAN bakeoff 2007.

It is worthy to distinguish two kinds of different robustness, one is for different corpus (from different sources or different domain and so on) under a same standard, we call it corpus robustness, and another is for different standards (for different application goals or demands and so on) for a same corpus. We call it standard robustness. The SIGHAN bakeoff series seems to focus more on later. We think corpus robustness should be received more attentions in the near future.

We participant all simplified Chinese track on WS, NER and POS task in the SIGHAN bakeoff 2007. There are more than two tracks for WS and POS. This gives us a chance to test the robustness of our systems. This paper reports our WS, NER and POS systems in the SIGHAN Bakeoff 2007, especially on the work of achieving robustness of WS and POS systems.

This paper is arranged as follows, we introduce our WS, NER and POS system separately in section 2, section 3 and section 4, experiments and results are listed in section 5, finally we draw some conclusions.

2 Word Segmentation

WS system includes three sequent steps, which are basic segmentation, disambiguation and out-of vocabulary (OOV) recognition. In each step, we construct a basic work unit first, and then have an adaptor to tune the basic unit to fit different standards.

2.1 Basic Segmentation

For constructing a basic work unit for WS, a common wordlist containing words ratified by four different segmentation standards (from SXU, NCC, PKU and CTB separately) are built. We finally get 64,000 words including about 1500 known entity words as the common wordlist. A forward-backward maximum matching algorithm with the common wordlist is employed as the common unit of our basic segmentor.

To cater for different characteristics in different segmentation standards, we construct another wordlist containing words for each specification. A wordlist based adaptor is built to implement the tuning task after basic segmentation.

2.2 Disambiguation

Disambiguation of overlapping Ambiguity (OA) is a major task in this step.

Strings with OA are also detected during basic forward-backward maximum matching in basic WS step. These strings are common OA strings for different standards. Class-based bigram model is applied to resolve the ambiguities. In class-based bigram, all named entities, all punctuation and factoids is one class respectively and each word is one class. We train the bigram transition probability based on the corpus of Chinese People's Daily 2000 newswire.

For corpus from different standards, overlapping ambiguity strings with less than 3 overlapping chain are extracted from each train corpus. We do not work on all of them but on some strings with a frequency that is bigger than a given value. A disambiguation adaptor using the highest probability segmentations is built for OA strings from each different standard.

2.3 OOV Recognition

In OOV recognition, we have a similar model which consists of a common part based on common characteristics and an individual part automatically constructed for each standard.

We divide OOV into factoid which contains non-Chinese characters like date, time, ordinal number, cardinal number, phone number, email address and non-factoid.

Factoid is recognized by an automaton. To compatible to different standards, we also built core automata and several adaptors.

Non-factoid is tackled by a unified character-based segmentation model based on CRF. We first transform the WS training dataset into character-based two columns format as the training dataset in NER task. The right column is a boundary tag of each character. The boundary tags are B I and S, which B is the tag of the first character of a word which contains more than two characters, I is the other non-initial characters in a word, S is for the single character word. Then the transformed training data is used to train the CRF model. Features in the model are current character and other three characters within the context and bigrams.

The trigger of non-factoid recognition is continual single character string excluding all the punctuations in a line after basic word matching, disambiguation and factoid incorporation. The model will tell whether these consecutive characters can form multi-character words in a given context.

At last, several rules are used to recognize some proper names separated by coordinate characters like "、", "和", "与" and symbol " • " in foreign person names.

3 Named Entity Recognition

We built two NER systems separately. One is a unified named entity model based on CRF. It used only a little knowledge include a small scale of entity dictionary, a few linguistic rules to process some special cases such as coordinate relation in corpus and some special symbols like dot among a transliteration foreign person name.

Another one is an individual model for each kind of entity based on Maximum Entropy where more rules found from corpus are used on entity boundary detection. Some details on this model can be found in Suxiang Zhang et al 2006.

4 POS Tagging

In POS, we construct POS taggers for different standards under a two steps frame, both steps use ME but with incremental features. First, we use normal features based Maximum Entropy (ME) to train a basic model, and then join some probabilistic features acquired from error analysis to training a finer model.

4.1 Normal Features for ME

In the first step of feature selection for ME tagger, we select contextual syntactic features for all words basing on a series of incremental experiments.

For shrinking the search space, the model only assigns each word a label occurred in the training data. That is, the model builds a subset of all POS tags for each token and restricts all possible labels of a word within a small candidate set, which greatly saves computing cost.

We enlarged PKU training corpus by using one month of Peking University's China Daily corpus (June in 2003) and CTB training corpus by using CTB 2.0 data which includes 325 passages.

To adapt with the given training corpus, the samples whose labels are not included in the standard training data were omitted firstly. After preprocessing, we get two sets of training samples for PKU and CTB with 1178 thousands tokens and 206 thousands tokens respectively. But the NCC test remains its original data due to we have no corpus with this standard.

4.2 Probabilistic feature for ME

By detecting the label errors when training and testing using syntactic features such as words around the current tokens and tags of previous tokens, words with multiple possible tags are obviously error-prone. We thus define some probabilistic features especially for multi-tag words.

We find labels of these tokens are most closely related to POS tag of word immediately previous to them. For instance, in corpus of Peking University, word “Report” has three different tags of “n(noun), v(verb), vn(noun verb)”. But when we taken into account its immediately previous words, we can find that when previous word's label is “q(quantifier)”, “Report” is labeled as “n” with a frequency of 91.67%, “v” with a frequency of 8.33% and “vn” with a frequency of 0.0%. We can assume that “Report” is labeled as “n” with the 91.67% probability when previous word's label is “q”, and so on.

Such probability is calculated from the whole training data and is viewed as discriminating probabilistic feature when choosing among the multiple tags for each word. But for words with only one possible tag, no matter what the label of previous word is, the label for them is always the tag occurred in the training data.

5 Experiments

We participant all simplified Chinese tracks on WS, NER and POS task in the SIGHAN bakeoff 2007. Our systems only deal with Chinese in GBK code. There are some mistakes in some results submitted to bakeoff organizer due to coding transform from GBK to UTF-16. We then use WS evaluation program in the SIGHAN bakeoff 2006 to re-evaluate WS system using same corpus, as for POS, since there is no POS evaluation in the SIGHAN bakeoff 2006, we implement a evaluation using ourselves’ program using same corpus.

Table 1 shows evaluation results of WS using evaluation programs from both the SIGHAN bakeoff 2007 and the SIGHAN bakeoff 2006. Table 2 lists evaluation results of NER using evaluation program from the SIGHAN bakeoff 2007. Table 3 gives evaluation results of POS using evaluation programs from both the SIGHAN bakeoff 2007 and ourselves(BUPT).

Track	UTF-16 (SIGHAN4)	GBK (SIGHAN 3)
CTB	0.9256	0.950
SXU	0.8741	0.969
NCC	0.9592	0.972

Table 1. WS results (F-measure)

SIGHAN 4	R	P	F
System-1	0.8452	0.872	0.8584
System-2	0.8675	0.9163	0.8912

Table 2. NER results (F-measure)

Track	UTF-16 (SIGHAN 4)	GBK (BUPT)
CTB	0.9689	0.9689
NCC	0.9096	0.9096
PKU	0.6649	0.9462

Table 3. POS Results (F-measure)

From the table 1 and Table 3, we can find our system is robust enough. WS system keeps at a relatively steady performance. Difference in POS between NCC and other two tracks is mainly due to the difference of the training corpus.

6 Conclusion

Recently, the robustness, a capability of keeping good performances for a system by automatically fitting the different corpus and standards, become a focal problem. This paper introduces our WS, NER and POS systems, especially on how they can get a robust performance.

The SIGHAN bakeoff series seems to focus more on standard robustness. We think corpus robustness should be received more attentions in the near future.

Acknowledgement

Thanks to Zhang Yan, Zhang Bichuan, Zhang Taozheng, Liu Haipeng and Jiang Huixing for all the work they done to make the WS, NER and POS systems go on wheels in a very short time.

References

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

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Republic of Korea.

NanYuan Liang. 1987 A Written Chinese Segmentation system– CDWS. *Journal of Chinese Information Processing*, Vol.2: 44-52

YaJuan Lv, Tie-jun Zhao, et al. 2001. Leveled unknown Chinese Words resolution by dynamic programming. *Journal Information Processing*, 15(1): 28-33.

Yintang Yan, XiaoQiang Zhou. 2000. Study of Segmentation Strategy on Ambiguous Phrases of Overlapping Type *Journal of The China Society For Scientific and Technical Information* Vol. 19 , №6

Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Caixia Yuan, Xiaojie Wang, Yixin Zhong. Some Improvements on Maximum Entropy Based Chinese POS Tagging. *The Journal of China Universities of Posts and Telecommunications*, Vol. 13, pp 99-103, 2006.

Suxiang Zhang, Xiaojie Wang, Juan Wen, Ying Qin, Yixin Zhong. A Probabilistic Feature Based Maximum Entropy Model for Chinese Named Entity Recognition, in *proceedings of 21st International Conference on the Computer Processing of Oriental Languages*,December 17-19, 2006, Singapore.