Chinese Word Segmentation and Named Entity Recognition by Character Tagging

Kun Yu[1] Sadao Kurohashi[2] Hao Liu[1] Toshiaki Nakazawa[1]

Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan, 113-8656[1]

Graduate School of Informatics, Kyoto University, Kyoto, Japan, 606-8501[2]

{kunyu, liuhao, nakazawa}@kc.t.u-tokyo.ac.jp[1]
kuro@i.kyoto-u.ac.jp[2]

Abstract

This paper describes our word segmentation system and named entity recognition (NER) system for participating in the third SIGHAN Bakeoff. Both of them are based on character tagging, but use different tag sets and different features. Evaluation results show that our word segmentation system achieved 93.3% and 94.7% F-score in UPUC and MSRA open tests, and our NER system got 70.84% and 81.32% F-score in LDC and MSRA open tests.

1 Introduction

Dealing with word segmentation as character tagging showed good results in last SIGHAN Bakeoff (J.K.Low et al.,2005). It is good at unknown word identification, but only using character-level features sometimes makes mistakes when identifying known words (T.Nakagawa, 2004). Researchers use word-level features (J.K.Low et al.,2005) to solve this problem. Based on this idea, we develop a word segmentation system based on character-tagging, which also combine character-level and word-level features. In addition, a character-based NER module and a rule-based factoid identification module are developed for post-processing.

Named entity recognition based on character-tagging has shown better accuracy than word-based methods (H.Jing et al.,2003). But the small window of text makes it difficult to recognize the named entities with many characters, such as organization names (H.Jing et al.,2003). Considering about this, we developed a NER system based on character-tagging, which combines word-level and character-level features together. In addition, in-NE probability is defined in this system to remove incorrect named entities and create new named entities as post-processing.

2 Character Tagging for Word Segmentation and NER

2.1 Basic Model

We look both word segmentation and NER as character tagging, which is to find the tag sequence T^* with the highest probability given a sequence of characters $S=c_1c_2\ldots c_n$.

$$T^* = \arg\max_{T} P(T \mid S) \tag{1}$$

Then we assume that the tagging of one character is independent of each other, and modify formula 1 as

$$\begin{aligned} T^* &= \arg\max_{T=t_1t_2\ldots t_n} P(t_1t_2\ldots t_n \mid c_1c_2\ldots c_n) \\ &= \arg\max_{T=t_1t_2\ldots t_n} \prod_{i=1}^{n} P(t_i \mid c_i) \end{aligned} \tag{2}$$

Beam search (n=3) (Ratnaparkhi,1996) is applied for tag sequence searching, but we only search the valid sequences to ensure the validity of searching result. SVM is selected as the basic classification model for tagging because of its robustness to over-fitting and high performance (Sebastiani, 2002). To simplify the calculation, the output of SVM is regarded as $P(t_i|c_i)$.

2.2 Tag Definition

Four tags 'B, I, E, S' are defined for the word segmentation system, in which 'B' means the character is the beginning of one word, 'I' means the character is inside one word, 'E' means the character is at the end of one word and 'S' means the character is one word by itself.

For the NER system, different tag sets are defined for different corpuses. Table 1 shows the

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 146–149,
Sydney, July 2006. ©2006 Association for Computational Linguistics

tag set defined for MSRA corpus. It is the product of Segment-Tag set and NE-Tag set, because not only named entities but also words are segmented in this corpus. Here NE-Tag 'O' means the character does not belong to any named entities. For LDC corpus, because there is no segmentation information, we delete NE-Tag 'O' but add tag 'NONE' to indicate the character does not belong to any named entities (Table 2).

Table 1 Tags of NER for MSRA corpus

Segment-Tag	×	NE-Tag
B, I, E, S		PER, LOC, ORG, O

Table 2 Tags of NER for LDC corpus

Segment Tag	×	NE Tag	+	NONE
B, I, E, S		PER, LOC, ORG, GPE		

2.3 Feature Definition

First, some features based on characters are defined for the two tasks, which are:

(a) C_n (n=-2,-1,0,1,2)

(b) $Pu(C_0)$

Feature C_n (n=-2,-1,0,1,2) mean the Chinese characters appearing in different positions (the current character and two characters to its left and right), and they are binary features. A character list, which contains all the characters in the lexicon introduced later, is used to identify them. Besides of that, feature $Pu(C_0)$ means whether C_0 is in a punctuation character list. It is also binary feature and all the punctuations in the punctuation character list come from Penn Chinese Treebank 5.1 (N.Xue et al.,2002).

In addition, we define some word-level features based on a lexicon to enlarge the window size of text in the two tasks, which are:

(c) W_n (n=-1,0,1)

Feature W_n (n=-1,0,1) mean the lexicon words in different positions (the word containing C_0 and one word to its left and right) and they are also binary features. We select all the possible words in the lexicon that satisfy the requirements, not like only selecting the longest one in (J.K.Low et al.,2005). To create the lexicon, we use following steps. First, a lexicon from NICT (National Institute of Information and Communications Technology, Japan) is used as the basic lexicon, which is extracted from Peking University Corpus of the second SIGHAN Bakeoff (T.Emerson, 2005), Penn Chinese Treebank 4.0 (N.Xue et al.,2002), a Chinese-to-English Wordlist[1] and part of NICT corpus (K.Uchimoto et al.,2004; Y.J.Zhang et al.,2005). Then, all the words containing digits and letters are removed from this lexicon. At last, all the punctuations in Penn Chinese Treebank 5.1 (N.Xue et al.,2002) and all the words in the training data of UPUC and MSRA corpuses are added into the lexicon.

Besides of above features, some extra features are defined only for NER task.

First, we add some character-based features to improve the accuracy of person name recognition, which are CN_n (n=-2,-1,0,1,2). They mean whether C_n (n=-2,-1,0,1,2) belong to a Chinese surname list. All of them are binary features. The Chinese surname list contains the most famous 100 Chinese surnames, such as 赵, 钱, 孙, 李 (Zhao, Qian, Sun, Li).

Then, we add some word-based features to help identify the organization name, which are $WORG_n$ (n=-1,0,1). They mean whether W_n (n= -1,0,1) belong to an organization suffix list. All of them are also binary features. The organization suffix list is created by extracting the last word from all the organization names in the training data of both MSRA and LDC corpuses.

3 Post-processing

Besides of the basic model, a NER module and a factoid identification module are developed in our word segmentation system for post-processing. In addition, we define in-NE probability to delete the incorrect named entities and identify new named entities in the post-processing phrase of our NER system.

3.1 Named Entity Recognition for Word Segmentation

In this module, if two or more segments in the outputs of basic model are recognized as one named entity, we combine them as one segment.

This module uses the same basic NER model as what we introduced in the previous section. But it only identifies person and location names, because organization names often contain more than one word. In addition, to keep the high accuracy of person name recognition, the features about organization suffixes are not used here.

3.2 Factoid Identification for Word Segmentation

Rules are used to identify the following factoids among the segments from the basic word segmentation model:

NUMBER: Integer, decimal, Chinese number
PERCENT: Percentage and fraction
DATE: Date
FOREIGN: English words

[1] http://projects.ldc.upenn.edu/Chinese/

Table 3 shows some rules defined here.

Table 3 Some Rules for Factoid Identification

Factoid	Rule
NUMBER	If previous segment ends with DIGIT and current segment starts with DIGIT, then combine them.
PERCENT	If previous segment is composed of DIGIT and current segment equals '%', then combine them.
DATE	If previous segment is composed of DIGIT and current segment is in the list of '年, 月, 日, 号 (Year, Month, Day, Day)', then combine them.
FOREIGN	Combine the consequent letters as one segment.

(DIGIT means both Arabic and Chinese numerals)

3.3 NER Deletion and Creation

In-word probability has been used in unknown word identification successfully (H.Q.Li et al., 2004). Accordingly, we define in-NE probability to help delete and create named entities (NE).

Formula 3 shows the definition of in-NE probability for character sequence $c_i c_{i+1} \ldots c_{i+n}$. Here '# of $c_i c_{i+1} \ldots c_{i+n}$ as NE' is defined as $Time_{InNE}$ and the occurrence of $c_i c_{i+1} \ldots c_{i+n}$ in different type of NE is treated differently.

$$P_{InNE}(c_i c_{i+1} .. c_{i+n}) = \frac{\#\text{ of } c_i c_{i+1} .. c_{i+n} \text{ as NE}}{\#\text{ of } c_i c_{i+1} .. c_{i+n} \text{ in testing data}} \quad (3)$$

Then, we use some criteria to delete the incorrect NE and create new possible NE, in which different thresholds are set for different tasks.

Criterion 1: If $P_{InNE}(c_i c_{i+1} \ldots c_{i+n})$ of one NE type is lower than T_{Del}, and $Time_{InNE}(c_i c_{i+1} \ldots c_{i+n})$ of the same NE type is also lower than T_{Time}, then delete this type of NE composed of $c_i c_{i+1} \ldots c_{i+n}$.

Criterion 2: If $P_{InNE}(c_i c_{i+1} \ldots c_{i+n})$ of one NE type is higher than T_{Cre}, and in other places the character sequence $c_i c_{i+1} \ldots c_{i+n}$ does not belong to any NE, then create a new NE containing $c_i c_{i+1} \ldots c_{i+n}$ with this NE type.

4 Evaluation Results and Discussion

4.1 Evaluation Setting

SVMlight (T.Joachims, 1999) was used as SVM tool. In addition, we used the MSRA training corpus of NER task in this Bakeoff to train our NER post-processing module.

4.2 Results of Word Segmentation

We attended the open track of word segmentation task for two corpuses: UPUC and MSRA. Table 4 shows the evaluation results.

Table 4 Results of Word Segmentation Task (in percentage %)

Corpus	Pre.	Rec.	F-score	Roov	Riv
UPUC	94.4	92.2	93.3	68.0	97.0
MSRA	94.0	95.3	94.7	50.3	96.9

The F-score of our word segmentation system in UPUC corpus ranked 4th (same as that of the 3rd group) among all the 8 participants. And it was only 1.1% lower than the highest one and 0.2% lower than the second one. It showed that our character-tagging approach was feasible. But the F-score of MSRA corpus was only higher than one participant in all the 10 groups (the highest one was 97.9%). Error analysis shows that there are two main reasons.

First, in MSRA corpus, they tend to segment one organization name as one word, such as 美国中国商会(China Chamber of Commerce in USA). But our basic segmentation model segmented such word into several words, e.g. 美国/中国/商会(USA/China/Chamber of Commerce), and our post-processing NER module does not consider about organization names.

Second, our factoid identification rule did not combine the consequent DATE factoids into one word, but they are combined in MSRA corpus. For example, our system segmented the word 晚上 9 时整 (9 o'clock in the evening) into three parts 晚上/9 时/整 (Evening/9 o'clock/Exact). This error can be solved by revising the rules for factoid identification.

Besides of that, we also found although our large lexicon helped identify the known word successfully, it also decreased the recall of OOV words (our Riv of UPUC corpus ranked 2nd, with only 0.6% decrease than the highest one, but Roov ranked 4th, with 8.8% decrease than the highest one). The large size of this lexicon is looked as the main reason.

Our lexicon contains 221,407 words, in which 6,400 words are single-character words. It made our system easy to segment one word into several words, for example word 经济组 (Economy Group) in UPUC corpus was segmented into 经济 (Economy) and 组(Group). Moreover, the large size of this lexicon also brought errors of combining two words into one word if the word was in the lexicon. For example, words 只 (Only) and 有 (Have) in MSRA corpus were identified as one word because there existed the word 只有 (Only) in our lexicon. We will reduce our lexicon to a reasonable size to solve these problems.

4.3 Results of NER

We also attended the open track of NER task for both LDC corpus and MSRA corpus. Table 5 and Table 6 give the evaluation results.

There were only 3 participants in the open track of LDC corpus and our group got the best F-score. In addition, among all the 11 participants for MSRA corpus, our system ranked 6th by F-score. It showed the validity of our character-tagging method for NER. But for location name (LOC) in LDC corpus, both the precision and recall of our NER system were very low. It was because there were too few location names in the training data (there were only 476 LOC in the training data, but 5648 PER, 5190 ORG and 9545 GPE in the same data set).

Table 5 Results of NER Task for LDC corpus (in percentage %)

	PER	LOC	ORG	GPE	Overall
Pre.	83.29	58.52	61.48	78.66	76.16
Rec.	66.93	18.87	45.19	79.94	66.21
F-score	74.22	28.57	52.09	79.30	70.84

Table 6 Results of NER Task for MSRA corpus (in percentage %)

	PER	LOC	ORG	Overall
Pre.	90.76	85.62	73.90	84.68
Rec.	76.13	85.41	65.74	78.22
F-score	82.80	85.52	69.58	81.32

Besides of that, error analysis shows there are four types of main errors in the NER results.

First, some organization names were very long and can be divided into several words, in which parts of them can also be looked as named entities. In such case, our system only recognized the small parts as named entities. For example, 哈佛大学费正清东亚研究中心 (Fei Zhengqing Eastern Asia Research Center of Harvard Univ.) was an organization name. But our system recognized it as 哈佛大学(Harvard Univ.)/ORG+费正清 (Fei Zheng Qing)/PER+ 东亚 (Eastern Asia)/LOC+ 研究中心(Research Center)/ORG. Adding more context features may be useful to resolve this issue.

In addition, our system was not good at recognizing foreign person names, such as 赖尔登 (Riordan), and abbreviations, such as 洛市 (Los Angeles), if they seldom or never appeared in training corpus. It is because the use of the large lexicon decreased the unknown word identification ability of our NER system simultaneously.

Third, the in-NE probability used in post-processing is helpful to identify named entities which cannot be recognized by the basic model. But it also recognized some words which can only be regarded as named entities in the local context incorrectly. For example, our system recognized 南京 (Najing) as GPE in 送到南京医治 (Send to Najing for remedy) in LDC corpus. We will consider about adding the in-NE probability as one feature into the basic model to solve this problem.

At last, in LDC corpus, they combine the attributive of one named entity (especially person and organization names) with the named entity together. But our system only recognized the named entity by itself. For example, our system only recognized 刘桂芳 (Liu Gui Fang) as PER in the reference person name 不知内情的刘桂芳 (Liu Gui Fang who does not know the inside).

5 Conclusion and Future Work

Through the participation of the third SIGHAN Bakeoff, we found that tagging characters with both character-level and word-level features was effective for both word segmentation and NER. While, this work is only our preliminary attempt and there are still many works needed to do in the future, such as the control of lexicon size, the use of extra knowledge (e.g. pos-tag), the feature definition, and so on. In addition, our word segmentation system only combined the NER module as post-processing, which resulted in that lots of information from NER module cannot be used by the basic model. We will consider about combining the NER and factoid identification modules into the basic word segmentation model by defining new tag sets in our future work.

Acknowledgement

We would like to thank Dr. Kiyotaka Uchimoto for providing the NICT lexicon.

Reference

T.Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *the 4th SIGHAN Workshop*. pp. 123-133.

H.Jing et al. 2003. HowtogetaChineseName(Entity): Segmentation and Combination Issues. In *EMNLP 2003*. pp. 200-207.

T.Joachims. 1999. Making large-scale SVM learning practical. *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

H.Q.Li et al. 2004. The Use of SVM for Chinese New Word Identification. In *IJCNLP 2004*. pp. 723-732.

J.K.Low et al. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. In *the 4th SIGHAN Workshop*. pp. 161-164.

T.Nakagawa. 2004. Chinese and Japanese Word Segmentation Using Word-level and Character-level Information. In *COLING 2004*. pp. 466-472.

A.Ratnaparkhi. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *EMNLP 1996*.

F.Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*. 34(1): 1-47.

K.Uchimoto et al. 2004. Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and its Applications. In *Proceedings of the MLR 2004*. pp. 63-70.

N.Xue et al. 2002. Building a Large-Scale Annotated Chinese Corpus. In *COLING 2002*.

Y.J.Zhang et al. 2005. Building an Annotated Japanese-Chinese Parallel Corpus – A part of NICT Multilingual Corpora. In *Proceedings of the MT SummitX*. pp. 71-78.