Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese

†Haibo Li, ‡Masato Hagiwara, §Qi Li, §Heng Ji

†City University of New York; ‡Rakuten Institute of Technology; §Rensselaer Polytechnic Institute
†‡New York, NY USA; §Troy, NY USA
lihaibo.c@gmail.com; masato.hagiwara@mail.rakuten.com; liqiearth@gmail.com; jih@rpi.edu

Abstract

Word Segmentation is usually considered an essential step for many Chinese and Japanese Natural Language Processing tasks, such as name tagging. This paper presents several new observations and analysis on the impact of word segmentation on name tagging; (1). Due to the limitation of current state-of-the-art Chinese word segmentation performance, a character-based name tagger can outperform its word-based counterparts for Chinese but not for Japanese; (2). It is crucial to keep segmentation settings (e.g. definitions, specifications, methods) consistent between training and testing for name tagging; (3). As long as (2) is ensured, the performance of word segmentation does not have appreciable impact on Chinese and Japanese name tagging.

Keywords: Name Tagging, Word Segmentation, Information Extraction

1. Introduction

Unlike most Indo-European languages, a Chinese and Japanese sentence is represented as a sequence of characters without natural delimiters. Therefore, Word Segmentation (WS) is usually considered as an essential step for many downstream Chinese and Japanese natural language processing tasks such as name tagging.

A key problem of word-based name tagging lies on the performance of WS system performance's on out-of-vocabulary (OOV) words. Current state-of-the-art WS system can only achieve about 40% of recall on some corpora (Gao et al., 2005). However, most names are very varied and out of the vocabulary of WS system. If the boundaries between a name and its contexts are mistakenly decided, it may make the detection of this name impossible. For example, a state-of-the-art word segmentation system splits a Geographical/Political Entity (GPE) "文莱 (Brunei)" falsely in the following sentence:

- correct: 收到 主办国 文莱 的 回复 ...
- wrong: 收到 主办 国文 莱 的 回复 ...
- English: Received the host country Brunei's reply ...

Similarly, Janpanese segmenter may also split names mistakenly:

- correct: 辻 元 議員は ...
- wrong: 辻元 議員は ...
- English: Tuji ex-congressman ...

It is impossible for a word based name tagger to detect the GPE "文莱 (Brunei)" using this incorrect segmentation for Chinese. At the same time, the segmentation of Japanese example also makes the tagging of the person name "辻 (Tuji)" impossible.

In this paper we aim to investigate and compare the impact of word segmentation on name tagging for Chinese and Japanese. The new observations can be summarized as follows.

- **With or Without word segmentation**: Similar to previous work (He and Wang, 2008) and (Liu et al., 2010), we found that a character-based name tagger can outperform word-based taggers for Chinese. However, for Japanese the character-based name tagger performs poorly because Japanese names are usually longer and include more complicated internal structures.
- **Training and Testing**: We found that it is crucial to keep the segmentation settings consistent between training and testing for both Chinese and Japanese name tagging. Applying a worse segmenter consistently to both training and testing, name tagger can achieve better performance than applying different better segmenters to training and testing.
- **Propagation of segmentation performance to name tagging**: When the segmentation settings in training and testing are consistent, the performance of WS is not propagated into name tagging for both languages.

2. Related Work

Chinese word segmentation has been intensively investigated in recent years. Many methods have been evaluated by international evaluations such as the Sighan Bakeoffs (GOH et al., 2004; Xu et al., 2004; Emerson, 2005; Levow, 2006; Jin and Chen, 2008; Zhao and Liu, 2010). Segmentation performance has been improved significantly, from the earliest Maximal Match (dictionary-based) approaches to CRF approach (Chang et al., 2005). In this paper we applied the improved version of that system based on lexicon features to demonstrate the effect of word segmentation on name tagging (Chang et al., 2008).

Many Chinese NER systems have been proposed and evaluated including (Emerson, 2005; Levow, 2006; Jin and Chen, 2008). These methods systematically investigated the performance of different methods, including: Hidden Markov Model (HMM), CRF, boosting, multi-phase model and hybrid models (Feng et al., 2006; Li et al., 2006; Chen

Feature Type	Description
n-gram	Uni-gram, bi-gram and tri-gram unit (character or word) sequences in the context window of the current unit. For example, $U_n(n = -3, -2, -1, 0, 1, 2, 3)$, $U_nU_{n+1}(n = -3, -2, -1, 0, 1, 2)$ and $U_nU_{n+1}U_{n+2}(n = -3, -2, -1, 0, 1)$.
Dictionary	Various types of gazetteers[2], such as person names, organizations, countries and cities, titles and idioms are used. For example, a feature "B-Country" means the current token is the first token of an entry of our country name list.
Part-of-Speech	Part-of-Speech tags in the contexts are used. This feature is only used for word level name tagging. For example, "POS_1=N" means the first word after current word is a noun.
Conjunction	Conjunctions of various features. Similar to the n-gram feature, Part-of-Speech tags of each unit in bi-gram and tri-gram unit sequences are combined as conjunction features. For example, POS_1POS_2=N&N.

Table 1: Features for Chinese and Japanese Name Tagging.

et al., 2006; Wu et al., 2006). Specifically, for character based methods, many different methods are adopted. For example, (Zhao and Kit, 2006; He and Wang, 2008) adopted a CRF-based method; a beam search based model is applied to Chinese name tagging based on Support Vector Machines (Yu et al., 2006); (Carpenter, 2006) used a Hidden Markov model of the LingPipe toolkit to recognize Chinese names. (Zhu et al., 2003) proposed source-channel model framework for single character name tagging. (Mao et al., 2008) proposed a CRF-based two-stage architecture to exploit non-local features and alleviate class imbalanced distribution on name tagging data set. In (Klein et al., 2003), the authors proposed a character-level HMM with minimal context information, and a model using maximum-entropy conditional markov model with substantially richer context features. (Shi and Wang, 2007) presented a joint decoding method on dual-layer CRFs guarding against violations of hard-constrains. The proposed method consistently improves the baselines that do not perform joint decoding.

Although a very intense work on Chinese and Japanese word segmentation and Chinese and Japanese name tagging has been done, the way in which word segmentation affects name tagging performance is not well understood. In this paper, besides investigating the performance of character based model and word based model, we also tested the effect of different segmentation settings on name tagging results. Furthermore, the consistency of segmentation settings between training and testing was also studied.

3. Word Segmenters

To determine the effect of word segmenters on name tagging, we applied two types of segmenters: one is dictionary based and the other is CRF-based.

For the dictionary based Chinese word segmenter (Wan and Luo, 2003), a dictionary including 50,551 unique entries is used in a Maximum Matching (MM) algorithm (Liu et al., 1994). The algorithm starts from the left end of a Chinese sentence and tries to match the first longest word wherever possible. If there are unknown words, they will be segmented as single characters.

The CRF-based segmenter is built with a large number of linguistic features such as character identity and character reduplication (Chang et al., 2008). The character identity features are represented using feature functions that are the key of the identity of the character in the current, proceeding and subsequent positions.

Data set	Dic-segmenter			CRF-segmenter		
	P	R	F_1	P	R	F_1
as	72.1	91.0	80.5	95.0	94.3	**94.7**
cityu	67.0	88.3	76.2	94.1	94.6	**94.3**
msr	79.1	94.6	86.2	96.2	96.6	**96.4**
pku	80.3	94.0	86.6	94.6	95.4	**95.0**
BCCWJ	85.7	78.15	81.8	91.3	89.9	**90.6**

Table 2: Chinese Word Segmentation performance (%) on SIGHAN 2005 data set (as, cityu, msr, pku) and Japanese Word Segmentation performance (%) on BCCWJ data set. (the bold F-scores are the best for each data set).

We compare the performance of two segmenters on SIGHAN 2005 corpus (Table 2). The performance of the CRF-based segmenter is got from the original paper of this segmenter (Chang et al., 2005). It is very obvious that CRF-based model outperforms the dictionary based segmenter on all corpora dramatically.

4. Name Taggers

4.1. General Pipeline

In this paper, the name tagging task is cast as a sequential labeling problem, where each unit (a word or a character) is assigned a label from a predefined tag set. More formally, let $x = (x_1, \cdots, x_T)$ be the input sentence, the output is a sequence of labels $y = (y_1, \cdots, y_T)$, where y_t is label for the unit x_t. We apply linear-chain Conditional Random Field (CRF) to address this problem. In the framework of linear-chain CRF, given an input sequence $\mathbf{x}$, the conditional distribution of the output label sequence $\mathbf{y}$ is defined as:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \cdot exp \sum_{j=1}^{T} \sum_{k=1}^{K} \theta_k \cdot f_k(y_j, y_{j-1}, \mathbf{x}, j) \quad (1)$$

where f_k is a feature function, θ_k is its weight, and $Z(\mathbf{x})$ is the normalization factor.

4.2. Features

Given the CRF-based framework, the remaining challenge is to design features for both character based and word based methods. In general we adopted four types of features for the CRF-based model, which are described in the table 1.

Among these features, the dictionary-based feature is a bridge between string matching based method and statistical method, which not only finds clues for named

Methods		Named Entity Types			
		GPE	PER	ORG	ALL
Dic-based Training CRF-based Testing	P	86.6	90.2	71.9	84.1
	R	95.7	92.0	79.7	91.2
	F_1	90.9	**91.1**	75.6	85.5
CRF-based Training Dic-based Testing	P	78.7	89.0	70.8	79.3
	R	92.3	89.6	84.5	89.9
	F_1	85.0	89.3	77.0	84.3
Dic-based Training Dic-based Testing	P	85.7	89.5	72.3	83.6
	R	96.1	91.2	84.1	92.2
	F_1	90.6	90.4	77.8	87.6
CRF-based Training CRF-based Testing	P	86.5	90.1	71.2	83.7
	R	96.2	91.3	83.7	92.1
	F_1	**91.1**	90.7	76.9	87.7
Character-based Method	P	86.2	91.0	75.5	84.8
	R	95.5	89.6	85.2	91.7
	F_1	90.6	90.2	**80.1**	**88.1**

Table 3: Performance (%) on ACE 2005 Chinese data set (the bold F_1-scores are the best for each type).

entities from dictionary lookup, but also assigns a real-valued weight to each matching through the statistical classifier. If the current token matches one entry in a given dictionary, then a feature representing the type of this dictionary is introduced to the token.

However, In character-based name tagging, the unit of tagging is a character, while the minimal unit of gazetteers is a word. This difference makes it difficult to perform dictionary lookup directly in character-based system. We efficiently addressed this problem by using prefix tree (a.k.a tire tree).

1. Match the first token in a sentence to the first level in the prefix tree.
2. If a match is found, then repeat step 1 to match the next token in next level until a leaf.
3. If the above procedure fails, go to next token in the sentence.
4. If a path from the root to a leaf is found, then an entry is matched. Repeat step 1 at the token next to the matched sub-string.

Finally, if a sub-sequence of the sentence matches an entry in dictionary D, following the "$BILUO$" tagging schema, we use "B-D" as a feature for the first character, and "I-D" for the following characters, and "L-D" for the last character. For example, if the "D" is "GPE", and the string "北京" matched an entry in D, there will be a feature "B-GPE" for "北" and "L-GPE" for "京", respectively.

4.3. Word based and Character based Models

In order to evaluate the impact of the basic unit granularity on name tagging, we developed two CRF-based models with different unit granularities: word level and character level. These two models used the same feature templates as shown in table 1, except that part-of-speech based features are only used for word-based models. During the training of word-based models, we merged word segmentation results and gold-standard name tagging results by giving higher priority to the latter.

Methods		Named Entity Types			
		GPE	PER	ORG	ALL
Dic-based Training CRF-based Testing	P	86.6	90.2	71.9	84.1
	R	95.7	92.0	79.7	91.2
	F_1	90.9	**91.1**	75.6	85.5
CRF-based Training Dic-based Testing	P	78.7	89.0	70.8	79.3
	R	92.3	89.6	84.5	89.9
	F_1	85.0	89.3	77.0	84.3
Dic-based Training Dic-based Testing	P	85.7	89.5	72.3	83.6
	R	96.1	91.2	84.1	92.2
	F_1	90.6	90.4	77.8	87.6
CRF-based Training CRF-based Testing	P	86.5	90.1	71.2	83.7
	R	96.2	91.3	83.7	92.1
	F_1	**91.1**	90.7	76.9	87.7
Character-based Method	P	86.2	91.0	75.5	84.8
	R	95.5	89.6	85.2	91.7
	F_1	90.6	90.2	**80.1**	**88.1**

Table 4: Performance (%) on ACE 2005 Chinese data set (the bold F_1-scores are the best for each type).

5. Experiment

5.1. Chinese Name Tagging

Table 4 presents the name tagging performance of various methods on the Automatic Content Extraction[1] (ACE) 2005 Chinese data set.

Our first focus is investigating the effect of WS specifications on Chinese name tagging. The last row gives the overall F_1 scores obtained by each WS specification. If we keep the segmentation setting consistent during training and test phrases, the effect of WS on name tagging is not significant. CRF segmentation based name tagger outperformed dictionary segmenter based name tagger only 0.1% on F_1 score. However, using CRF-based segmenter in training and dictionary segmenter in testing produced the worst name tagging performance: 84.3%.

In terms of the F_1 metric, the character based method outperforms word based method on organization and overall scores. Especially, compared to the best score of word based methods, the character based method achieved 2.3% improvement on organization names. Furthermore, the consistent settings outperformed inconsistent settings on average 2.75% overall performance.

5.2. Japanese Name Tagging

We then compare our findings in Chinese with Japanese. We tested different Japanese segmenters for Japanese name tagging on the BCCWJ CORE corpus which has 1982 documents and 2,370,832 characters (Maekawa, 2008).

We adopted the MeCab toolkit to construct our CRF-based Japanese segmenter, which is a statistical Japanese morphological analyzer tool based on semi-markov CRFs. IPADic dictionary is used as word dictionary by the CRF-based segmenter (Kudo et al., 2004). We appled the JUMAN 7.0 as our dictionary base segmenter (Kurohashi and Nagao, 1994). The segmentation F_1 score of CRF-based segmenter and dictionary based segmenter are 90.57% and 81.73% respectively.

For the Japanese character based model, we use the same set of features as Chinese, except the character-type fea-

[1] http://www.itl.nist.gov/iad/mig//tests/ace/

Methods		Named Entity Types			
		GPE	PER	ORG	ALL
Dic-based Training Dic-based Testing	P	87.7	89.9	85.2	88.3
	R	81.1	75.1	57.5	73.2
	F_1	84.2	81.8	68.7	80.1
CRFs-based Training CRFs-based Testing	P	87.6	89.6	85.2	88.1
	R	82.8	77.6	57.9	75.0
	F_1	**85.1**	**83.2**	**69.0**	**81.1**
Character-based Method	P	88.4	92.1	82.4	88.9
	R	76.0	72.3	59.4	70.7
	F_1	81.7	81.0	**69.0**	78.8

Table 5: Performance (%) on BCCWJ CORE Japanese corpus (the bold F_1-scores are the best for each type).

	Chinese			Japanese		
	PER	GPE	ORG	PER	GPE	ORG
Median	4	5.5	7.5	8.5	7.5	34

Table 6: Name Length of Chinese and Japanese: the number of characters.

tures for each word/character. Kanji (Chinese characters), Hiragana, Katakana, upper/lower Roman alphabets, Sino-numbers, Arabic numbers, and others, are distinguished.

The experiment results are shown in Table 5. We used the same segmentation setting in the training and the test. The same findings are in the Chinese data set, although the CRF-based segmenter outperforms dictionary based segmenter with 8.84% F_1 score, the name tagger based on CRF segmenter achieves only 1% improvement of F_1 score over dictionary based segmenter. However, the character based Japanese name tagger does not performs well. We found that the main reason is that Japanese names are much longer than Chinese names and include more complicated internal structures, and thus more sensitive to word boundaries. Table 6 shows the length median of each name type.

6. Conclusions

We investigated the effect of word segmentation on name tagging for two languages, Chinese and Japanese. We find that a character-based Chinese name tagger can outperform its word-based counterparts; and the performance of word segmentation does not have appreciable impact on Chinese and Japanese name tagging, if the training and testing segmentation settings are consistent.

7. Acknowledgements

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), U.S. NSF CAREER Award under Grant IIS-0953149, U.S. DARPA Award No. FA8750-13-2-0041 in the "Deep Exploration and Filtering of Text" (DEFT) Program, IBM Faculty award and RPI faculty start-up grant. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

8. References

Bob Carpenter. 2006. Character language models for chinese word segmentation and named entity recognition. In *Proc. of SIGHAN-5 Workshop on Chinese Language Processing*, pages 169–172.

Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Proc. of SIGHAN-4 Workshop on Chinese Language Processing*.

Pichuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proc. of WMT*.

Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. 2006. Chinese named entity recognition with conditional random fields. In *Proc. of SIGHAN-5 Workshop on Chinese Language Processing*, pages 118–121.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proc. of SIGHAN-4 Workshop on Chinese Language Processing*, volume 133.

Yuanyong Feng, Le Sun, and Yuanhua Lv. 2006. Chinese word segmentation and named entity recognition based on conditional random fields models. In *Proc. of SIGHAN-5 Workshop on Chinese Language Processing*, pages 181–184.

Jianfeng Gao, Mu Li, Andi Wu, and Chang N. Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. *Comput. Linguist.*, 31:531–574.

Chooi-Ling GOH, Masayuki Asahara, and Yuji Matsumoto. 2004. Chinese word segmentation by classification of characters. In *Proc. of SIGHAN Workshop 2004*, pages 57–64.

Jingzhou He and Houfeng Wang. 2008. Chinese named entity recognition and word segmentation based on character. In *Proc. of SIGHAN-6 Workshop on Chinese Language Processing*.

Guangjin Jin and Xiao Chen. 2008. The fourth international chinese language processing bakeoff: Chinese word segmentation named entity recognition and chinese pos tagging. In *Proc. of SIGHAN-6 Workshop on Chinese Language Processing*.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proc. of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, pages 180–183.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proc. of EMNLP*, pages 230–237.

Sadao Kurohashi and Makoto Nagao. 1994. Improvements of japanese morphological analyzer juman. In *Proc. of the International Workshop on Sharable Natural Language Resources*, pages 22–38.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proc. of SIGHAN-5*

Workshop on Chinese Language Processing, pages 108–117.

Lishuang Li, Tingting Mao, Degen Huang, and Yuansheng Yang. 2006. Hybrid models for chinese named entity recognition. In *Proc. of SIGHAN-5 Workshop on Chinese Language Processing*, pages 72–78.

Ynan Liu, Qiang Tan, and Kun Xu Shen. 1994. In *The Word Segmentation Rules and Automatic Word Segmentation Methods for Chinese Information Processing (in Chinese)*, page 36.

Zhangxun Liu, Conghui Zhu, and Tiejun Zhao. 2010. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? In *Proceedings of ICIC'10*, pages 634–640.

Kikuo Maekawa. 2008. Compilation of the kotonoha-bccwj corpus (in japanese). *Nihongo no kenkyu (Studies in Japanese)*, 4(1):82–95.

Xinnian Mao, Yuan Dong, Saike He, Sencheng Bao, and Haila Wang. 2008. Chinese word segmentation and named entity recognition based on conditional random fields. In *Proc. of SIGHAN-6 Workshop on Chinese Language Processing*.

Yanxin Shi and Mengqiu Wang. 2007. A dual-layer crfs based joint decoding method for cascaded segmentation and labeling tasks. In *Proceedings of Twentieth International Joint Conference on Artificial Intelligence*, pages 1707–1712.

Min Wan and Zhensheng Luo. 2003. Study on topic segmentation method in automatic abstracting system. In *Proc. of the Natural Language Processing and Knowledge Engineering*.

Chia-Wei Wu, Shyh-Yi Jan, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2006. On using ensemble methods for chinese named entity recognition. In *Proc. of SIGHAN-5 Workshop on Chinese Language Processing*, pages 142–145.

Jia Xu, Richard Zens, and Hermann Ney. 2004. Do we need chinese word segmentation for statistical machine translation? In *Proc. of SIGHAN Workshop 2004*, pages 122–128.

Kun Yu, Sadao Kurohashi, Hao Liu, and Toshiaki Nakazawa. 2006. Chinese word segmentation and named entity recognition by character tagging. In *Proc. of SIGHAN-5 Workshop on Chinese Language Processing*.

Hai Zhao and Chunyu Kit. 2006. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *SIGHAN-5*, pages 162–165.

Hongmei Zhao and Qun Liu. 2010. The cips-sighan clp 2010 chinese word segmentation bakeoff. In *Proc. of the Joint Conference on Chinese Language Processing*, pages 199–209.

Xiaodan Zhu, Mu Li, Jianfeng Gao, and Chang-Ning Huang. 2003. Single character chinese named entity recognition. In *Proc. of SIGHAN-2 Workshop on Chinese Language Processing*, pages 125–132.