A Hybrid Approach to Word Segmentation and POS Tagging

Tetsuji Nakagawa
Oki Electric Industry Co., Ltd.
2−5−7 Honmachi, Chuo-ku
Osaka 541−0053, Japan
nakagawa378@oki.com

Kiyotaka Uchimoto
National Institute of Information and Communications Technology
3−5 Hikaridai, Seika-cho, Soraku-gun
Kyoto 619−0289, Japan
uchimoto@nict.go.jp

Abstract

In this paper, we present a hybrid method for word segmentation and POS tagging. The target languages are those in which word boundaries are ambiguous, such as Chinese and Japanese. In the method, word-based and character-based processing is combined, and word segmentation and POS tagging are conducted simultaneously. Experimental results on multiple corpora show that the integrated method has high accuracy.

1 Introduction

Part-of-speech (POS) tagging is an important task in natural language processing, and is often necessary for other processing such as syntactic parsing. English POS tagging can be handled as a sequential labeling problem, and has been extensively studied. However, in Chinese and Japanese, words are not separated by spaces, and word boundaries must be identified before or during POS tagging. Therefore, POS tagging cannot be conducted without word segmentation, and how to combine these two processing is an important issue. A large problem in word segmentation and POS tagging is the existence of unknown words. Unknown words are defined as words that are not in the system's word dictionary. It is difficult to determine the word boundaries and the POS tags of unknown words, and unknown words often cause errors in these processing.

In this paper, we study a hybrid method for Chinese and Japanese word segmentation and POS tagging, in which word-based and character-based processing is combined, and word segmentation and POS tagging are conducted simultaneously. In the method, word-based processing is used to handle known words, and character-based processing is used to handle unknown words. Furthermore, information of word boundaries and POS tags are used at the same time with this method. The following sections describe the hybrid method and results of experiments on Chinese and Japanese corpora.

2 Hybrid Method for Word Segmentation and POS Tagging

Many methods have been studied for Chinese and Japanese word segmentation, which include word-based methods and character-based methods. Nakagawa (2004) studied a method which combines a word-based method and a character-based method. Given an input sentence in the method, a lattice is constructed first using a word dictionary, which consists of word-level nodes for all the known words in the sentence. These nodes have POS tags. Then, character-level nodes for all the characters in the sentence are added into the lattice (Figure 1). These nodes have position-of-character (POC) tags which indicate word-internal positions of the characters (Xue, 2003). There are four POC tags, B, I, E and S, each of which respectively indicates the beginning of a word, the middle of a word, the end of a word, and a single character word. In the method, the word-level nodes are used to identify known words, and the character-level nodes are used to identify unknown words, because generally word-level information is precise and appropriate for processing known words, and character-level information is robust and appropriate for processing unknown words. Extended hidden Markov models are used to choose the best path among all the possible candidates in the lattice, and the correct path is indicated by the thick lines in Figure 1. The POS tags and the POC tags are treated equally in the method. Thus, the word-level nodes and the character-level nodes are processed uniformly, and known words and unknown words are identified simultaneously. In the method, POS tags of known words as well as word boundaries are identified, but POS tags of unknown words are not identified. Therefore, we extend the method in order to conduct unknown word POS tagging too:

Hybrid Method
The method uses subdivided POC-tags in order to identify not only the positions of characters but also the parts-of-speech of the composing words (Figure 2, A). In the method, POS tagging of unknown words is conducted at the same time as word segmentation and POS tag-

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 217–220,
Prague, June 2007.

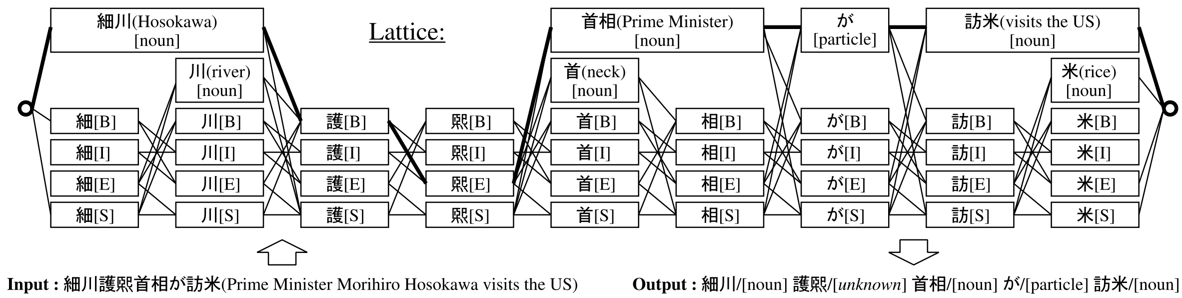

Figure 1: Word Segmentation and Known Word POS Tagging using Word and Character-based Processing

ging of known words, and information of parts-of-speech of unknown words can be used for word segmentation.

There are also two other methods capable of conducting unknown word POS tagging (Ng and Low, 2004):

Word-based Post-Processing Method
This method receives results of word segmentation and known word POS tagging, and predicts POS tags of unknown words using words as units (Figure 2, B). This approach is the same as the approach widely used in English POS tagging. In the method, the process of unknown word POS tagging is separated from word segmentation and known word POS tagging, and information of parts-of-speech of unknown words cannot be used for word segmentation. In later experiments, maximum entropy models were used deterministically to predict POS tags of unknown words. As features for predicting the POS tag of an unknown word w, we used the preceding and the succeeding two words of w and their POS tags, the prefixes and the suffixes of up to two characters of w, the character types contained in w, and the length of w.

Character-based Post-Processing Method
This method is similar to the word-based post-processing method, but in this method, POS tags of unknown words are predicted using characters as units (Figure 2, C). In the method, POS tags of unknown words are predicted using exactly the same probabilistic models as the hybrid method, but word boundaries and POS tags of known words are fixed in the post-processing step.

Ng and Low (2004) studied Chinese word segmentation and POS tagging. They compared several approaches, and showed that character-based approaches had higher accuracy than word-based approaches, and that conducting word segmentation and POS tagging all at once performed better than conducting these processing separately. Our hybrid method is similar to their character-based all-at-once approach. However, in their experiments, only word-based and character-based methods were examined. In our experiments, the combined method of word-based and character-based processing was examined. Furthermore, although their experiments were conducted with only Chinese data, we conducted experiments with Chinese and Japanese data, and confirmed that the hybrid method performed well on the Japanese data as well as the Chinese data.

3 Experiments

We used five word-segmented and POS-tagged corpora; the Penn Chinese Treebank corpus 2.0 (CTB), a part of the PFR corpus (PFR), the EDR corpus (EDR), the Kyoto University corpus version 2 (KUC) and the RWCP corpus (RWC). The first two were Chinese (C) corpora, and the rest were Japanese (J) corpora, and they were split into training and test data. The dictionary distributed with JUMAN version 3.61 (Kurohashi and Nagao, 1998) was used as a word dictionary in the experiments with the KUC corpus, and word dictionaries were constructed from all the words in the training data in the experiments with other corpora. Table 1 summarizes statistical information of the corpora: the language, the number of POS tags, the sizes of training and test data, and the splitting methods of them[1]. We used the following scoring measures to evaluate performance of word segmentation and POS tagging:

R : Recall (The ratio of the number of correctly segmented/POS-tagged words in system's output to the number of words in test data),

P : Precision (The ratio of the number of correctly segmented/POS-tagged words in system's output to the number of words in system's output),

[1]The unknown word rate for word segmentation is not equal to the unknown word rate for POS tagging in general, since the word forms of some words in the test data may exist in the word dictionary but the POS tags of them may not exist. Such words are regarded as known words in word segmentation, but as unknown words in POS tagging.

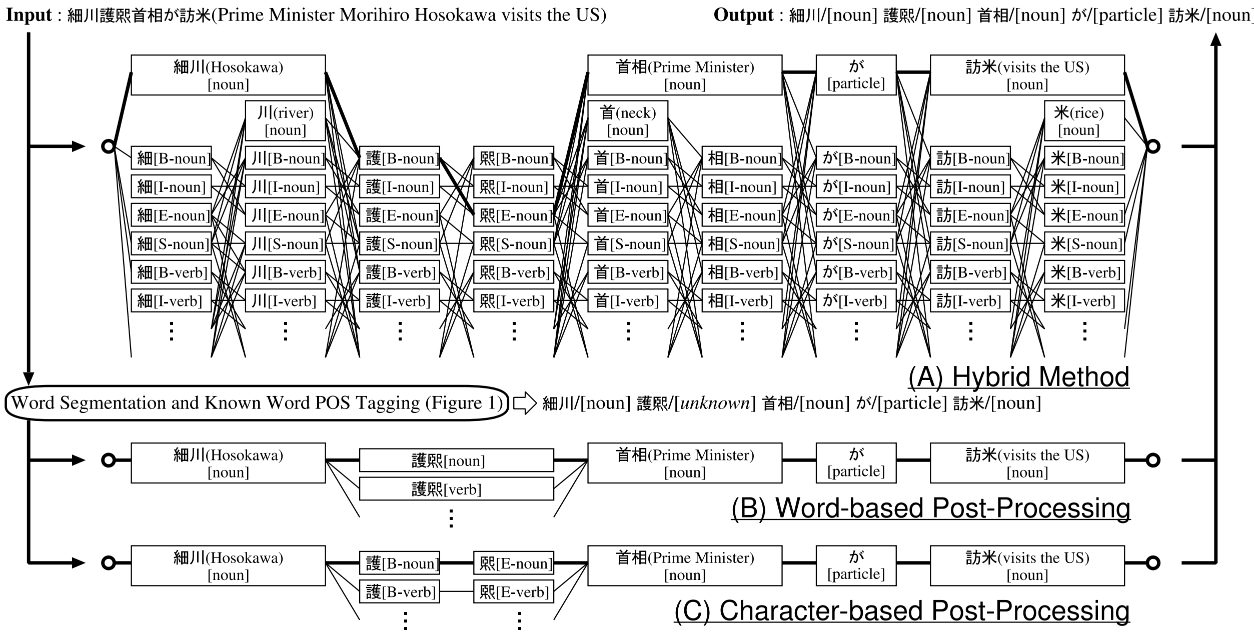

Figure 2: Three Methods for Word Segmentation and POS Tagging

F : F-measure ($F = 2 \times R \times P/(R + P)$),
$R_{unknown}$: Recall for unknown words,
R_{known} : Recall for known words.

Table 2 shows the results[2]. In the table, **Word-based Post-Proc.**, **Char.-based Post-Proc.** and **Hybrid Method** respectively indicate results obtained with the word-based post-processing method, the character-based post-processing method, and the hybrid method. Two types of performance were measured: performance of word segmentation alone, and performance of both word segmentation and POS tagging. We first compare performance of both word segmentation and POS tagging. The F-measures of the hybrid method were highest on all the corpora. This result agrees with the observation by Ng and Low (2004) that higher accuracy was obtained by conducting word segmentation and POS tagging at the same time than by conducting these processing separately. Comparing the word-based and the character-based post-processing methods, the F-measures of the latter were higher on the Chinese corpora as reported by Ng and Low (2004), but the F-measures of the former were slightly higher on the Japanese corpora. The same tendency existed in the recalls for known words; the recalls of the character-based post-processing method were highest on the Chinese corpora, but those of the word-based method were highest on the Japanese corpora, except on the EDR corpus. Thus, the character-based method was not always better than the word-based method as reported by Ng and Low (2004) when the methods were used with the word and character-based combined approach on Japanese corpora. We next compare performance of word segmentation alone. The F-measures of the hybrid method were again highest in all the corpora, and the performance of word segmentation was improved by the integrated processing of word segmentation and POS tagging. The precisions of the hybrid method were highest with statistical significance on four of the five corpora. In all the corpora, the recalls for unknown words of the hybrid method were highest, but the recalls for known words were lowest.

Comparing our results with previous work is not easy since experimental settings are not the same. It was reported that the original combined method of word-based and character-based processing had high overall accuracy (F-measures) in Chinese word segmentation, compared with the state-of-the-art methods (Nakagawa, 2004). Kudo et al. (2004) studied Japanese word segmentation and POS tagging using conditional random fields (CRFs) and rule-based unknown word processing. They conducted experiments with the KUC corpus, and achieved F-measure of 0.9896 in word segmentation, which is better than ours (0.9847). Some features we did not used, such as base forms and conjugated forms of words, and hierarchical POS tags, were used in

[2]The recalls for known words of the word-based and the character-based post-processing methods differ, though the POS tags of known words are identified in the first common step. This is because known words are sometimes identified as unknown words in the first step and their POS tags are predicted in the post-processing step.

Corpus (Lang.)	Number of POS Tags	Number of Words (Unknown Word Rate for Segmentation/Tagging) [partition in the corpus]	
		Training	Test
CTB (C)	34	84,937 [sec. 1–270]	7,980 (0.0764 / 0.0939) [sec. 271–300]
PFR (C)	41	304,125 [Jan. 1–Jan. 9]	370,627 (0.0667 / 0.0749) [Jan. 10–Jan. 19]
EDR (J)	15	2,550,532 [$id = 4n + 0, id = 4n + 1$]	1,280,057 (0.0176 / 0.0189) [$id = 4n + 2$]
KUC (J)	40	198,514 [Jan. 1–Jan. 8]	31,302 (0.0440 / 0.0517) [Jan. 9]
RWC (J)	66	487,333 [1–10,000th sentences]	190,571 (0.0513 / 0.0587) [10,001–14,000th sentences]

Table 1: Statistical Information of Corpora

Corpus (Lang.)	Scoring Measure	Word Segmentation			Word Segmentation & POS Tagging		
		Word-based Post-Proc.	Char.-based Post-Proc.	Hybrid Method	Word-based Post-Proc.	Char.-based Post-Proc.	Hybrid Method
CTB (C)	R	0.9625	0.9625	**0.9639**	0.8922	0.8935	**0.8944**
	P	0.9408	0.9408	**0.9519***	0.8721	0.8733	**0.8832**
	F	0.9516	0.9516	**0.9578**	0.8821	0.8833	**0.8887**
	$R_{unknown}$	0.6492	0.6492	**0.7148**	0.4219	0.4312	**0.4713**
	R_{known}	**0.9885**	**0.9885**	0.9845	0.9409	**0.9414**	0.9382
PFR (C)	R	0.9503	0.9503	**0.9516**	0.8967	0.8997	**0.9024***
	P	0.9419	0.9419	**0.9485***	0.8888	0.8917	**0.8996***
	F	0.9461	0.9461	**0.9500**	0.8928	0.8957	**0.9010**
	$R_{unknown}$	0.6063	0.6063	**0.6674**	0.3845	0.3980	**0.4487**
	R_{known}	**0.9749**	**0.9749**	0.9719	0.9382	**0.9403**	0.9392
EDR (J)	R	**0.9525**	**0.9525**	**0.9525**	**0.9358**	0.9356	0.9357
	P	0.9505	0.9505	**0.9513***	0.9337	0.9335	**0.9346**
	F	0.9515	0.9515	**0.9519**	0.9347	0.9345	**0.9351**
	$R_{unknown}$	0.4454	0.4454	**0.4630**	0.4186	0.4103	**0.4296**
	R_{known}	**0.9616**	**0.9616**	0.9612	**0.9457**	**0.9457**	0.9454
KUC (J)	R	**0.9857**	**0.9857**	0.9850	0.9572	0.9567	**0.9574**
	P	0.9835	0.9835	**0.9843**	0.9551	0.9546	**0.9566**
	F	0.9846	0.9846	**0.9847**	0.9562	0.9557	**0.9570**
	$R_{unknown}$	0.9237	0.9237	**0.9302**	0.6724	0.6774	**0.6879**
	R_{known}	**0.9885**	**0.9885**	0.9876	**0.9727**	0.9719	0.9721
RWC (J)	R	0.9574	0.9574	**0.9592**	0.9225	0.9220	**0.9255***
	P	0.9533	0.9533	**0.9577***	0.9186	0.9181	**0.9241***
	F	0.9553	0.9553	**0.9585**	0.9205	0.9201	**0.9248**
	$R_{unknown}$	0.6650	0.6650	**0.7214**	0.4941	0.4875	**0.5467**
	R_{known}	**0.9732**	**0.9732**	0.9720	**0.9492**	0.9491	0.9491

(Statistical significance tests were performed for R and P, and * indicates significance at $p < 0.05$)

Table 2: Performance of Word Segmentation and POS Tagging

their study, and it may be a reason of the difference. Although, in our experiments, extended hidden Markov models were used to find the best solution, the performance will be further improved by using CRFs instead, which can easily incorporate a wide variety of features.

4 Conclusion

In this paper, we studied a hybrid method in which word-based and character-based processing is combined, and word segmentation and POS tagging are conducted simultaneously. We compared its performance of word segmentation and POS tagging with other methods in which POS tagging is conducted as a separated post-processing. Experimental results on multiple corpora showed that the hybrid method had high accuracy in Chinese and Japanese.

References

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of EMNLP 2004*, pages 230–237.

Sadao Kurohashi and Makoto Nagao. 1998. *Japanese Morphological Analysis System JUMAN version 3.61*. Department of Informatics, Kyoto University. (in Japanese).

Tetsuji Nakagawa. 2004. Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information. In *Proceedings of COLING 2004*, pages 466–472.

Hwee Tou Ng and 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 EMNLP 2004*, pages 277–284.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics and Chinese*, 8(1):29–48.