Chinese Morphological Analysis with Character-level POS Tagging

Mo Shen[†], **Hongxiao Liu**[‡], **Daisuke Kawahara**[†], and **Sadao Kurohashi**[†]
[†]Graduate School of Informatics, Kyoto University, Japan
[‡]School of Computer Science, Fudan University, China
shen@nlp.ist.i.kyoto-u.ac.jp {dk,kuro}@i.kyoto-u.ac.jp
12210240027@fudan.edu.cn

Abstract

The focus of recent studies on Chinese word segmentation, part-of-speech (POS) tagging and parsing has been shifting from words to characters. However, existing methods have not yet fully utilized the potentials of Chinese characters. In this paper, we investigate the usefulness of character-level part-of-speech in the task of Chinese morphological analysis. We propose the first tagset designed for the task of character-level POS tagging. We propose a method that performs character-level POS tagging jointly with word segmentation and word-level POS tagging. Through experiments, we demonstrate that by introducing character-level POS information, the performance of a baseline morphological analyzer can be significantly improved.

1 Introduction

In recent years, the focus of research on Chinese word segmentation, part-of-speech (POS) tagging and parsing has been shifting from words toward characters. Character-based methods have shown superior performance in these tasks compared to traditional word-based methods (Ng and Low, 2004; Nakagawa, 2004; Zhao et al., 2006; Kruengkrai et al., 2009; Xue, 2003; Sun, 2010). Studies investigating the morphological-level and character-level internal structures of words, which treat character as the true atom of morphological and syntactic processing, have demonstrated encouraging results (Li, 2011; Li and Zhou, 2012; Zhang et al., 2013). This line of research has provided great insight in revealing the roles of characters in word formation and syntax of Chinese language.

However, existing methods have not yet fully utilized the potentials of Chinese characters. While Li (2011) pointed out that some characters can productively form new words by attaching to existing words, these characters consist only a portion of all Chinese characters and appear in 35% of the words in Chinese Treebank 5.0 (CTB5) (Xue et al., 2005). Zhang (2013) took one step further by investigating the character-level structures of words; however, the machine learning of inferring these internal structures relies on the character forms, which still suffers from data sparseness.

Character-level Part-of-Speech	Examples of Verb
verb + noun	投资 (invest : throw + wealth)
noun + verb	心疼 (feel sorry : heart + hurt)
verb + adjective	认清 (realize : recognize + clear)
adjective + verb	痛恨 (hate : pain + hate)
verb + verb	审查 (inspect : examine + review)

Table 1. Character-level POS sequence as a more specified version of word-level POS: an example of verb.

In our view, since each Chinese character is in fact created as a word in origin with complete and independent meaning, it should be treated as the actual minimal morphological unit in Chinese language, and therefore should carry specific part-of-speech. For example, the character "打" (beat) is a verb and the character "破" (broken) is an adjective. A word on the other hand, is either single-character, or a compound formed by single-character words. For example, the verb "打破" (break) can be seen as a compound formed by the two single-character words with the construction "verb + adjective".

Under this treatment, we observe that words with the same construction in terms of character-level POS tend to also have similar syntactic roles. For example, the words having the con-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 253–258,
Baltimore, Maryland, USA, June 23-25 2014.

struction “verb + adjective” are typically verbs, and those having the construction “adjective + noun” are typically nouns, as shown in the following examples:

(a) verb : verb + adjective
“打破”(break) : “打”(beat) + “破”(broken)
“更新”(update) : “更”(replace) + “新”(new)
“漂白”(bleach) : “漂”(wash) + “白”(white)

(b) noun : adjective + noun
“主题”(theme) : “主”(main) + “题”(topic)
“新人”(newcomer) : “新”(new) + “人”(person)
“快车”(express) : “快”(fast) + “车”(car)

This suggests that character-level POS can be used as cues in predicting the part-of-speech of unknown words.

Another advantage of character-level POS is that, the sequence of character-level POS in a word can be seen as a more fine-grained version of word-level POS. An example is shown in Table 1. The five words in this table are very likely to be tagged with the same word-level POS as verb in any available annotated corpora, while it can be commonly agreed among native speakers of Chinese that the syntactic behaviors of these words are different from each other, due to their distinctions in word constructions. For example, verbs having the construction “verb + noun” (e.g. 投资) or “verb + verb” (e.g. 审查) can also be nouns in some context, while others cannot; And verbs having the constructions “verb + adjective” (e.g. 认清) require exact one object argument, while others generally do not. Therefore, compared to word-level POS, the character-level POS can produce information for more expressive features during the learning process of a morphological analyzer.

In this paper, we investigate the usefulness of character-level POS in the task of Chinese morphological analysis. We propose the first tagset designed for the task of character-level POS tagging, based on which we manually annotate the entire CTB5. We propose a method that performs character-level POS tagging jointly with word segmentation and word-level POS tagging. Through experiments, we demonstrate that by introducing character-level POS information, the performance of a baseline morphological analyzer can be significantly improved.

Tag	Part-of-Speech	Example
n	noun	法案/NN (bill)
v	verb	发布/VV (publish)
j	adj./adv.	广阔/VA (vast)
t	numerical	三点一四/CD (3.14)
m	quantifier	一/CD 件/M (a piece of)
d	date	九五年/NT (1995)
k	proper noun	中美/NR (sino-US)
b	prefix	副市长/NN (vice mayor)
e	suffix	建筑业/NN (construction inductry)
r	transliteration	阿尔帕德/NR (Árpád)
u	punctuation	查尔斯·狄更斯/NR (Charles Dickens)
f	foreign chars	X 射线/NN (X-ray)
o	onomatopoeia	隆隆/AD (rumble)
s	surname	王新民/NR (Wang Xinmin)
p	pronoun	他们/PN (they)
c	other functional	用于/VV (be used for)

Table 2. Tagset for character-level part-of-speech tagging. The underlined characters in the examples correspond to the tags on the left-most column. The CTB-style word-level POS are also shown for the examples.

2 Character-level POS Tagset

We propose a tagset for the task of character-level POS tagging. This tagset contains 16 tags, as illustrated in Table 2. The tagset is designed by treating each Chinese character as a single-character word, and each (multi-character) word as a phrase of single-character words. Some of these tags are directly derived from the commonly accepted word-level part-of-speech, such as noun, verb, adjective and adverb. It should be noted that, for single-character words, the difference between adjective and adverb can almost be ignored, because for any of such words that can be used as an adjective, it usually can also be used as an adverb. Therefore, we have merged these two tags into one.

On the other hand, some other tags are designed specifically for characters, such as transliteration, surname, prefix and suffix. Unlike some Asian languages such as Japanese, there is no explicit character set in Chinese that are used exclusively for expressing names of foreign persons, places or organizations. However, some characters are used much more frequently than others in these situations. For example, in the person’s name “阿尔帕德” (Árpád), all the four characters can be frequently observed in words

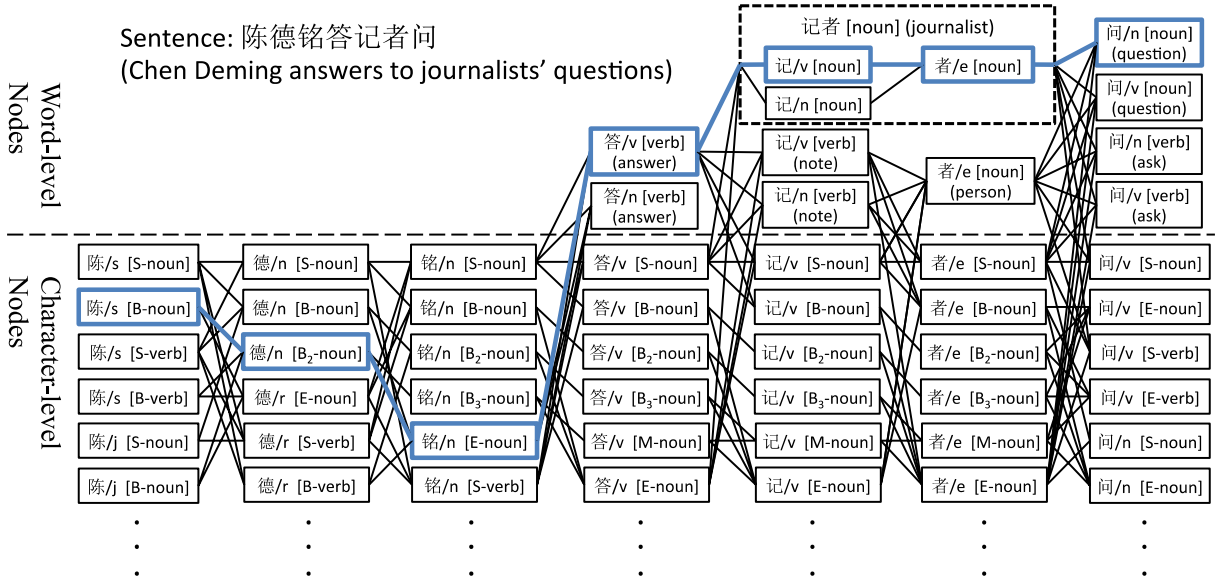

Figure 1. A Word-character hybrid lattice of a Chinese sentence. Correct path is represented by blue bold lines.

Word Length	1	2	3	4	5	6	7 or more
Tags	*S*	*BE*	BB_2E	BB_2B_3E	BB_2B_3ME	BB_2B_3MME	$BB_2B_3M...ME$

Table 3. Word representation with a 6-tag tagset: S, B, B_2, B_3, M, E

of transliterations. Similarly, surnames in Chinese are also drawn from a set of limited number of characters. We therefore assign specific tags for this kind of character sets. The tags for prefixes and suffixes are motivated by the previous studies (Li, 2011; Li and Zhou, 2012).

We have annotated character-level POS for all words in CTB5 [1]. Fortunately, character-level POS in most words are independent of context, which means it is sufficient to annotate word forms unless there is an ambiguity. The annotation was conducted by two persons, where each one of them was responsible for about 70% of the documents in the corpus. The redundancy was set for the purposes of style unification and quality control, on which we find that the inter-annotator agreement is 96.2%. Although the annotation also includes the test set, we blind this portion in all the experiments.

3 Chinese Morphological Analysis with Character-level POS

3.1 System Description

Previous studies have shown that jointly processing word segmentation and POS tagging is preferable to pipeline processing, which can propagate errors (Nakagawa and Uchimoto, 2007; Kruengkrai et al., 2009). Based on these studies, we propose a word-character hybrid model which can also utilize the character-level POS information. This hybrid model constructs a lattice that consists of word-level and character-level nodes from a given input sentence. Word-level nodes correspond to words found in the system's lexicon, which has been compiled from training data. Character-level nodes have special tags called position-of-character (POC) that indicate the word-internal position (Asahara, 2003; Nakagawa, 2004). We have adopted the 6-tag tagset, which (Zhao et al., 2006) reported to be optimal. This tagset is illustrated in Table 3.

Figure 2 shows an example of a lattice for the Chinese sentence: "陈德铭答记者问" (Chen Deming answers to journalists' questions). The correct path is marked with blue bold lines. The

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?CharPosCN

Category	Template	Condition
Baseline-unigram	$\langle w_0 \rangle \langle p_0 \rangle \langle w_0, p_0 \rangle \langle l_0, p_0 \rangle \langle \text{begin}(w_0), p_0 \rangle \langle \text{end}(w_0), p_0 \rangle$	W_0
	$\langle \text{begin}(w_0), \text{end}(w_0), p_0 \rangle$	
	$\langle c_{-2}, p_0 \rangle \langle c_{-1}, p_0 \rangle \langle c_0, p_0 \rangle \langle c_1, p_0 \rangle \langle c_2, p_0 \rangle$	C_0
	$\langle c_{-2}, c_{-1}, p_0 \rangle \langle c_{-1}, c_0, p_0 \rangle \langle c_0, c_1, p_0 \rangle \langle c_1, c_2, p_0 \rangle \langle c_{-1}, c_1, p_0 \rangle$	
Baseline-bigram	$\langle w_{-1}, w_0 \rangle \langle p_{-1}, p_0 \rangle \langle w_{-1}, p_0 \rangle \langle p_{-1}, w_0 \rangle \langle w_{-1}, p_{-1}, w_0 \rangle \langle w_{-1}, w_0, p_0 \rangle$	$W_{-1} \times W_0$
	$\langle w_{-1}, p_{-1}, p_0 \rangle \langle p_{-1}, w_0, p_0 \rangle \langle w_{-1}, p_{-1}, w_0, p_0 \rangle \langle l_{-1}, p_{-1}, l_0 \rangle \langle l_{-1}, l_0, p_0 \rangle$	
	$\langle l_{-1}, p_{-1}, p_0 \rangle \langle p_{-1}, l_0, p_0 \rangle \langle l_{-1}, p_{-1}, l_0, p_0 \rangle \langle \text{end}(w_{-1}), p_0 \rangle$	
	$\langle p_{-1}, \text{begin}(w_0) \rangle \langle \text{end}(w_{-1}), p_{-1}, p_0 \rangle \langle p_{-1}, \text{begin}(w_0), p_0 \rangle$	
	$\langle c_{-1}, c_0 \rangle \langle p_{-1}, p_0 \rangle \langle c_{-1}, p_{-1}, c_0 \rangle \langle c_{-1}, c_0, p_0 \rangle$	$C_{-1} \times C_0$
	$\langle c_{-1}, p_{-1}, p_0 \rangle \langle p_{-1}, c_0, p_0 \rangle \langle c_{-1}, p_{-1}, c_0, p_0 \rangle$	
	$\langle p_{-1}, p_0 \rangle$	Otherwise
Proposed-unigram	$\langle \text{CP}(c_0), p_0 \rangle$	C_0
Proposed-bigram	$\langle \text{CP}_{pair}(w_{-1}), p_0 \rangle \langle \text{CP}_{pair}(w_{-1}), p_{-1}, p_0 \rangle$	$W_{-1} \times N_0$
	$\langle \text{CP}_{all}(w_{-1}), p_0 \rangle \langle \text{CP}_{all}(w_{-1}), p_{-1}, p_0 \rangle$	
	$\langle \text{CP}_{pair}(w_{-1}), p_{-1}, \text{CP}(c_0) \rangle \langle \text{CP}_{pair}(w_{-1}), \text{CP}(c_0), p_0 \rangle$	$W_{-1} \times C_0$
	$\langle \text{CP}_{all}(w_{-1}), p_{-1}, \text{CP}(c_0) \rangle \langle \text{CP}_{all}(w_{-1}), \text{CP}(c_0), p_0 \rangle$	
	$\langle p_{-1}, \text{CP}(c_0) \rangle \langle p_{-1}, \text{CP}(c_0), p_0 \rangle$	$N_{-1} \times C_0$
	$\langle \text{CP}(c_{-1}), p_0 \rangle \langle \text{CP}(c_{-1}), p_{-1}, p_0 \rangle$	$C_{-1} \times N_0$
	$\langle \text{CP}(c_{-1}), p_{-1}, \text{CP}(c_0) \rangle \langle \text{CP}(c_{-1}), \text{CP}(c_0), p_0 \rangle \langle \text{CP}(c_{-1}), p_{-1}, \text{CP}(c_0), p_0 \rangle$	$C_{-1} \times C_0$

Table 4. Feature templates. The "Condition" column describes when to apply the templates: W_{-1} and W_0 denote the previous and the current word-level node; C_{-1} and C_0 denote the previous and the current character-level node; N_{-1} and N_0 denote the previous and the current node of any types. Word-level nodes represent known words that can be found in the system's lexicon.

upper part of the lattice (word-level nodes) represents known words, where each node carries information such as character form, character-level POS , and word-level POS. A word that contains multiple characters is represented by a sub-lattice (the dashed rectangle in the figure), where a path stands for a possible sequence of character-level POS for this word. For example, the word "记者" (journalist) has two possible paths of character-level POS: "verb + suffix" and "noun + suffix". Nodes that are inside a sub-lattice cannot be linked to nodes that are outside, except from the boundaries. The lower part of the lattice (character-level nodes) represents unknown words, where each node carries a position-of-character tag, in addition to other types of information that can also be found on a word-level node. A sequence of character-level nodes are considered as an unknown word if and only if the sequence of POC tags forms one of the cases listed in Table 3. This table also illustrates the permitted transitions between adjacent character-level nodes. We use the standard dynamic programming technique to search for the best path in the lattice. We use the averaged perceptron (Collins, 2002), an efficient online learning algorithm, to train the model.

3.2 Features

We show the feature templates of our model in Table 4. The features consist of two categories: baseline features, which are modified from the templates proposed in (Kruengkrai et al., 2009); and proposed features, which encode character-level POS information.

Baseline features: For word-level nodes that represent known words, we use the symbols w, p and l to denote the word form, POS tag and length of the word, respectively. The functions $\text{begin}(w)$ and $\text{end}(w)$ return the first and last character of w. If w has only one character, we omit the templates that contain $\text{begin}(w)$ or $\text{end}(w)$. We use the subscript indices 0 and -1 to indicate the current node and the previous node during a Viterbi search, respectively. For character-level nodes, c denotes the surface character, and p denotes the combination of POS and POC (position-of-character) tags.

Proposed features: For word-level nodes, the function $\text{CP}_{pair}(w)$ returns the pair of the character-level POS tags of the first and last characters of w, and $\text{CP}_{all}(w)$ returns the sequence of character-level POS tags of w. If either the pair or the sequence of character-level POS is ambiguous, which means there are multiple paths in the sub-lattice of the word-level node, then the values on the current best path (with local context) during the Viterbi search will be returned. If w has only one character, we omit the templates that contain $\text{CP}_{pair}(w)$. For character-level nodes, the function $\text{CP}(c)$ returns its character-level POS. The subscript indices 0 and -1 as well as

other symbols stand for the same meaning as they are in the baseline features.

4 Evaluation

4.1 Settings

To evaluate our proposed method, we have conducted two sets of experiments on CTB5: word segmentation, and joint word segmentation and word-level POS tagging. We have adopted the same data division as in (Jiang et al., 2008a; Jiang et al., 2008b; Kruengkrai et al., 2009; Zhang and Clark, 2010; Sun, 2011): the training set, dev set and test set have 18,089, 350 and 348 sentences, respectively. The models applied on all test sets are those that result in the best performance on the CTB5 dev set.

We have annotated character-level POS information for all 508,768 word tokens in CTB5. As mentioned in section 2, we blind the annotation in the test set in all the experiments. To learn the characteristics of unknown words, we built the system's lexicon using only the words in the training data that appear at least 3 times. We applied a similar strategy in building the lexicon for character-level POS, where the threshold we choose is 2. These thresholds were tuned using the development data.

We have used precision, recall and the F-score to measure the performance of the systems. Precision (P) is defined as the percentage of output tokens that are consistent with the gold standard test data, and recall (R) is the percentage of tokens in the gold standard test data that are recognized in the output. The balanced F-score (F) is defined as $\frac{2 \cdot P \cdot R}{P+R}$.

4.2 Experimental Results

We compare the performance between a baseline model and our proposed approach. The results of the word segmentation experiment and the joint experiment of segmentation and POS tagging are shown in Table 5(a) and Table 5(b), respectively. Each row in these tables shows the performance of the corresponding system. "CharPos" stands for our proposed model which has been described in section 3. "Baseline" stands for the same model except it only enables features from the baseline templates.

The results show that, while the differences between the baseline model and the proposed model in word segmentation accuracies are small, the proposed model achieves significant improvement in the experiment of joint segmentation and POS tagging[2]. This suggests that our proposed method is particularly effective in predicting the word-level POS, which is consistent with our observations mentioned in section 1.

(a) Word Segmentation Results

System	P	R	F
Baseline	97.48	98.44	97.96
CharPOS	97.55	98.51	98.03

(b) Joint Segmentation and POS Tagging Results

System	P	R	F
Baseline	93.01	93.95	93.48
CharPOS	93.42	94.18	93.80

Table 5. Experimental results on CTB5.

System	Segmentation	Joint
Baseline	97.96	93.48
CharPOS	98.03	93.80
Jiang2008a	97.85	93.41
Jiang2008b	97.74	93.37
Kruengkrai2009	97.87	93.67
Zhang2010	97.78	93.67
Sun2011	98.17	94.02

Table 6. Comparison with previous studies on CTB5.

In Table 6 we compare our approach with morphological analyzers in previous studies. The accuracies of the systems in previous work are directly taken from the original paper. As the results show, despite the fact that the performance of our baseline model is relatively weak in the joint segmentation and POS tagging task, our proposed model achieves the second-best performance in both segmentation and joint tasks.

5 Conclusion

We believe that by treating characters as the true atoms of Chinese morphological and syntactic analysis, it is possible to address the out-of-vocabulary problem that word-based methods have been long suffered from. In our error analysis, we believe that by exploring the character-level POS and the internal word structure (Zhang et al., 2013) at the same time, it is possible to further improve the performance of morphological analysis and parsing. We will address these issues in our future work.

[2] $p < 0.05$ in McNemar's test.

Reference

Masayuki Asahara. 2003. Corpus-based Japanese Morphological Analysis. Nara Institute of Science and Technology, Doctor's Thesis.

Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In Proceedings of EMNLP, pages 1–8.

Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lü. 2008a. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-speech Tagging. In Proceedings of ACL.

Wenbin Jiang, Haitao Mi, and Qun Liu. 2008b. Word Lattice Reranking for Chinese Word Segmentation and Part-of-speech Tagging. In Proceedings of COLING.

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, YiouWang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An Error-Driven Word-Character Hybird Model for Joint Chinese Word Segmentation and POS Tagging. In Proceedings of ACL-IJCNLP, pages 513-521.

Zhongguo Li. 2011. Parsing the Internal Structure of Words: A New Paradigm for Chinese Word Segmentation. In Proceedings of ACL-HLT, pages 1405–1414.

Zhongguo Li and Guodong Zhou. 2012. Unified Dependency Parsing of Chinese Morphological and Syntactic Structures. In Proceedings of EMNLP, pages 1445–1454.

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, pages 277–284.

Tetsuji Nakagawa. 2004. Chinese and japanese word segmentation using word-level and character-level information. In Proceedings of COLING, pages 466–472.

Tetsuji Nakagawa and Kiyotaka Uchimoto. 2007. Hybrid Approach to Word Segmentation and Pos Tagging. In Proceedings of ACL Demo and Poster Sessions, pages 217-220.

Weiwei Sun. 2010. Word-based and Character-based Word Segmentation Models: Comparison and Combination. In Proceedings of COLING Poster Sessions, pages 1211–1219.

Weiwei Sun. 2011. A Stacked Sub-word Model for Joint Chinese Word Segmentation and Part-of-speech Tagging. In Proceedings of ACL-HLT, pages 1385–1394.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. In International Journal of Computational Linguistics and Chinese Language Processing.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11(2):207–238.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In Proceedings of PACLIC, pages 87-94.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese Parsing Exploiting Characters. In Proceedings of ACL, page 125-134.

Yue Zhang and Stephen Clark. 2010. A Fast Decoder for Joint Word Segmentation and POS-tagging Using a Single Discriminative Model. In Proceedings of EMNLP, pages 843–852.