A Lexicon-Constrained Character Model for Chinese Morphological Analysis

Yao Meng, Hao Yu, and Fumihito Nishino

Fujitsu R&D Center Co., Ltd, Room B1003, Eagle Run Plaza, No. 26 Xiaoyun Road, Chaoyang District, Bejing, 100016, P. R. China
{Mengyao, Yu, Nishino}@frdc.fujitsu.com

Abstract. This paper proposes a lexicon-constrained character model that combines both word and character features to solve complicated issues in Chinese morphological analysis. A Chinese character-based model constrained by a lexicon is built to acquire word building rules. Each character in a Chinese sentence is assigned a tag by the proposed model. The word segmentation and part-of-speech tagging results are then generated based on the character tags. The proposed method solves such problems as unknown word identification, data sparseness, and estimation bias in an integrated, unified framework. Preliminary experiments indicate that the proposed method outperforms the best SIGHAN word segmentation systems in the open track on 3 out of the 4 test corpora. Additionally, our method can be conveniently integrated with any other Chinese morphological systems as a post-processing module leading to significant improvement in performance.

1 Introduction

Chinese morphological analysis is a fundamental problem that has been studied extensively [1], [2], [3], [4], [5], [6], [7], [8]. Researchers make use of word or character features to cope with this problem. However, neither of them seems completely satisfactory.

In general, a simple word-based approach can achieve about 90% accuracy for segmentation with a medium-size dictionary. However, since no dictionary includes every Chinese word, the unknown word (or Out Of Vocabulary, OOV) problem [9], [10] can severely affect the performance of word-based approaches. Furthermore, word-based models have an estimation bias when faced with segmentation candidates with different numbers of words. For example, in the standard hidden Markov model, the best result, $T^* = \arg\max_T p(T|W) = \arg\max_T \prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_1..t_{i-1})$, is related to the number of the words in the segmentation candidates. As such, a candidate with fewer words is preferred over those with more words in the selection process. Therefore, most word-based models are likely to fail when a combinational ambiguity[1] sequence is separated into multiple words.

[1] A typical segmentation ambiguity, it refers to a situation in which the same Chinese sequence may be one word or several words in different contexts.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 542 – 552, 2005.

Compared with Chinese words, Chinese characters are relatively less unambiguous. The Chinese character set is very limited. Therefore, unknown characters occur rarely in a sentence. The grammatical advantages of characters have inspired researchers to adopt character features in Chinese morphology and parsing [5], [6], [11], [12]. However, it is difficult to incorporate necessary word features, such as the form of a Chinese word and its fixed part-of-speech tags, in most character-based approaches. For this reason, character-based approaches have not achieved satisfactory performance in large-scale open tests.

In this paper, we propose a lexicon-constrained character model to combine the merits of both approaches. We explore how to capture the Chinese word building rules using a statistical method, which reflects the regularities in the word formation process. First, a character hidden Markov method assigns the candidate tags to each character. Next, a large-size word list combined with linguistic information is used to filter out erroneous candidates. Finally, segmentation and part-of-speech tagging for the sentence are provided based on the character tags.

The proposed model solves the problems of unknown word detection, word segmentation and part-of-speech tagging using both word and character features. Additionally, our module is a post-processing module, which can be coupled to any existing Chinese morphological system; and it can readily recall some of the unknown words omitted by the system, and as a result, significantly improves the overall performance. Evaluations of the proposed system on SIGHAN open test sets indicate that our method outperforms the best bakeoff results on 3 test sets, and ranks 2nd in the 4th test set [9].

2 A Lexicon-Constrained Character Model for Chinese Morphology

2.1 An Elementary Model to Describe Chinese Word Building Rules

It is recognized that there are some regularities in the process of forming words from Chinese characters. This in general can be captured by word building rules. In this paper, we explore a statistical model to acquire such rules. The following are some definitions used in the proposed model.

[Def. 1] character position feature

We use four notations to denote the position of a character in a Chinese word. 'F' means the first character of the word, 'L' the last character, 'M' is a character within it and 'S' the word itself.

[Def. 2] character tag set

It is the product of the set of character position features and the set of part-of-speech tags.

Character tag set $=\{xy|\, x \in word\ \text{POS}\ set, y \in \{S, F, M, L\}\}$, where, x denotes one part-of-speech (POS) tag and y a character position feature. Together they are used to define the rules of Chinese word formation.

[Def. 3] character tagging

Given a Chinese sentence; character tagging is the process for assigning a character tag to each character in the sentence.

Word building rules are acquired based on the relation between the character and the corresponding character tag. Word segmentation and part-of-speech tagging can be achieved easily based on the result of character tagging. For example, a character with '*x*S' is a single character word with the part-of-speech tag '*x*'; a character sequence starting with '*x*F' and ending with '*x*L' is a multiple character word with the part-of-speech tag '*x*'.

The elementary model adopts the character bi-gram hidden Markov model. In hidden Markov model, given the sentence, $s: c_1 c_2 ... c_{n-1} c_n$, and character tagging result $t: xy_1 xy_2 ... xy_{n-1} xy_n$, the probability of result *t* of *s* is estimated as:

$$p(t \mid s) = \prod_{i=1,n} p(xy_i \mid xy_{i-2}\ xy_{i-1}) \times p(c_i \mid xy_i) \tag{1}$$

The best character tagging result for the sentences is given by equation (2):

$$t^* = \arg\max_t \prod_{i=1,n} p(xy_i \mid xy_{i-2} xy_{i-1}) \times p(c_i \mid xy_i) \tag{2}$$

We used the People's Daily Corpus of 1998 [13] to train this model. Also we adopted a 100,000-word dictionary listing all valid part-of-speech tags for each Chinese word in the training phase to solve the data sparseness problem. The training data are converted into character tagging data through the following steps: a single character word with '*x*' is converted into the character marked with tag '*x*S'; a two-character word with '*x*' is converted into a first character with '*x*F' and a second character with '*x*L'; a word with more than two characters with '*x*' are converted into a first character with '*x*F', middle characters with '*x*M' and last character with '*x*L'. We adopt the POS tag set from the People's Daily Corpus, which consists of 46 tags. Taking into account of the four position features, the final character tag set is comprised of 184 tags.

The emitted probability and transition probability of the model are estimated by the maximum likelihood method. The emitted probability is counted by the training Corpus and the dictionary, where the Chinese words in the dictionary are counted one time. The transition probability is trained from the training Corpus only.

2.2 An Improved Character-Based Model Using Lexicon Constraints

We tested the above model based on the SIGHAN open test set [9]. The average precision for word segmentation was more than 88%. This means that most of the word building rules in Chinese have been obtained by the elementary model. However, the performance was relatively inferior to other word segmentation systems. It indicated that the model needed more features to learn word building rules. In error analysis, we found that the elementary model was so flexible that it produced many pseudo-words and invalid part-of-speech tags. In practice, a Chinese word is a stable sequence of Chinese characters, whose formation and part-of-speech tags are fixed by long-term usage. It seemed that only character position and meaning cannot describe a word building rule effectively.

We also observed that word segmentation systems based on a simple dictionary matching algorithm and a few linguistic rules could achieve about 90% accuracy [14]. This suggested that a lexicon may have contribution to word building rules. Thus, we tried to incorporate a lexicon to the model to improve the performance.

The major errors in the elementary model were pseudo words and invalid part-of-speech (POS) tags. We proposed two constraints based on the lexicon to deal with these errors:

1. If a possible word produced from the elementary model is in the word-dictionary, the character tag of the characters forming this word should be consistent with the part-of-speech tag of the word in the dictionary.
2. If a possible word produced is not in the dictionary, it must include one or more single characters, and none of which may be subsumed by any word in the dictionary in the current context.

The first constraint eliminates invalid character tags. For example, the character '*明*' has six character tags: '*aF*' (first in adjective) , '*dF*' (first in adverb), '*nF*' (first in noun), '*nrF*' (first in person name), '*tF*' (first in time), and '*vF*' (first in verb). The character '*天* has five character tags: '*dL*', '*nL', 'nrL', 'tL*', and '*vL*'. The combination of the two characters produces the possible word '*明天*', which includes five possible word part-of-speech tags: '*d*', '*n', 'nr', 't*', and '*v*' based on these character tags. But '*明天*' is a word in the dictionary, which only has two valid part-of-speech tags, namely, 'time' and 'person name'. Obviously, the part-of-speech tags: '*d', 'n*' and '*v*' of '*明天*' are invalid. Accordingly, the tags '*aF', 'dF', 'nF' , 'vF*' on '*明*' and the tags '*dL', 'nL', 'vL*' on '*天*' are also invalid. So they should be pruned from the candidates of the character tagging.

The second constraint prunes pseudo words in the elementary model. Many studies in dictionary-based segmentation treat unknown words as sequences of single characters [1], [14]. The second constraint ensures that the new word produced by the elementary model must have one or more 'unattached' single characters (not subsumed by any other words). For example, the sequence '*程序错误*' (program error) will combine the pseudo word '*序错*' because of the tag '*nF*' on '*序*' and the tag '*nL*' on '*错*'. The second constraint will prune '*序错*' since '*程序*' (program) and '*错误*' (error) are already in the dictionary and there is no "unattached" single character in it. Accordingly, the tag '*nF*' on '*序*' and the tag '*nL*' on '*错*' will be deleted from the candidates of character tagging.

The following experiments show the lexicon-based constraints are very effective in eliminating error cases. The elementary model faces an average of 9.3 character tags for each character. The constraints will prune 70% of these error tags from it. As a result, the performance of character tagging is improved.

It is worth noting that the lexicon in the elementary model cannot distort the probability of the character tagging results in the model. The pruned cases are invalid cases which cannot occur in the training data because all the words and POS tags in the training data are valid. Thus, the model built from the training data is not affected by the pruning process.

2.3 Case Study

In this subsection, we illustrate the advantages of the proposed method for Chinese morphology with an example.

Example: 小明明天将就程序错误进行分析

(Xiaoming will analyze the program errors tomorrow).

Where, '*小明*' is an unknown word (person name), and the sequence '*将就*' is a combinational ambiguity (either '***将就***' (put up with) or '*将*'+ '*就*' (will)). Here is how our approach works.

Step 1: List all the character tags for each character. Figure 1 shows the character tags in the sequence '*小明明天*' .

小	aF	dF	nF	nrF	nM	nrM	nsM	qM	vM	aL	dL	vL	aS
明	aF	dF	nF	nrF	vF	tF	nM	lM	tM	aL	dL	nrL	aS
明	aF	dF	nF	nrF	vF	tF	nM	lM	tM	aL	dL	nrL	aS
天	nF	tF	nrM	dL	nL	nrL	tL	vL					

Fig. 1. Candidates for the sequence '**小明明天**'

In this step we are able to find possible unknown words based on character position features. For example, the character tags in '*小明明天*' combine four possible unknown words: '*小明*', '*小明明*', '***明明天***' , and '*小明明天*'.

Step 2: Prune the invalid candidates using constraints.

The first constraint prunes some invalid character tags. For example, '*明明*' can be either an adverb (d) or a personal name (nr); '*明天*' is a time (t) word. The other part-of-speech tags of these two words will be deleted. With the second constraint, we can delete '*明明天*' because '*明明*' and '*明天*' are words in the dictionary. However, '*小明*' , '*小明明*', and '*小明明天*' will be kept because '*小*' is a "unattached" single character. The remaining candidates are shown in figure 2.

小	aF	dF	nF	nrF	nM	nrM	nsM	qM	vM	aL	dL	vL	aS
明		dF		nrF			nM	lM	tM				aS
明						tF	nM	lM	tM		dL	nrL	aS
天	nF	tF	nrM			nrL	tL						

Fig. 2. Remaining Candidates for the sequence '**小明明天**'

Step 3: Choose the best character tagging result based on the proposed character hidden Markov model.

The best character tagging result is chosen using equation 2 in Section 2.1. The ambiguities in segmentation and word POS tagging are solved in the character tagging process.

Consider the combinational ambiguity '*将就*' in the following 2 candidates:

Candidate 1: '*小明/nr 明天/t 将/d 就/d 程序/n 错误/n 进行/v 分析/v*'

Candidate 2: '*小明/nr 明天/t 将就/v 程序/n 错误/n 进行/v 分析/v*'

In word-based linear model, the erroneous candidate 2 will be prior to the correct candidate 1 since the model counts 9 nodes in candidate 1 but 8 nodes in candidate 2. However, there is no such bias in the character model because the number of characters does not change. The combinational ambiguity '*将就*' will be denoted as '*将/dS 就/dS*' or '*将/vF 就/vL*'. The number of nodes in all candidates of character tagging is the same.

At last, the correct result '小/*nr*F 明/*nr*L 明/*t*F 天/*t*L 将/*d*S 就/*d*S程/*n*F 序/*n*L 错/*n*F 误/*n*L 进/*v*F 行/*v*L 分/*v*F 析/*v*L' is selected, and the corresponding morphological result is: '小明/*nr* 明天/*t* 将/*d* 就/*d* 程序/*n* 错误/*n* 进行/*v* 分析/*v* '.

The above steps show the proposed approach solves the various issues related to Chinese morphology by a concise character tagging process where word building is revealed.

3 Experiments and Discussion

We evaluated the proposed character method using the SIGHAN Backoff data, i.e. the one-month People's Daily Corpus of 1998, and the first version of Penn Chinese Treebank [15]. We compared our approach against two state-of-the-art systems: one is based on a bi-gram word segmentation model [7], and the other based on a word-based hidden Markov model [3]. For simplicity, we only considered three kinds of unknown words (personal name, location name, and organization name) in the all methods.

The same corpus and word-dictionary were used to train the above three systems. The training data set was the 5-month People's Daily Corpus of 1998, which contained approximately 6,300,000 words and 46 word part-of-speech tags. The system dictionary contained 100,000 words and the valid part-of-speech tag(s) of each word. On average, there were 1.3 part-of-speech tags for a word in the dictionary.

In the following, chr-HMM refers to the proposed elementary model; chr-HMM+Dic refers to the character model improved by integrating linguistic information. W-Bigram is the word-based bi-gram system, and W-HMM is the word-based hidden Markov system.

3.1 Morphological Experimental Results

We examined the performance of our model in comparison against W-Bigram and W-HMM. Table 1 compares the segmentation performance of our model against that of other models. Table 2 shows the accuracy in unknown word identification. Table 3 illustrates the performance of the part-of-speech tagging. The experiments in Table 1 and Table 2 were examined using the SIGHAN open test corpora. The experiments in Table 3 were performed again on the one-month People's Daily Corpus (PD corpus) and 4,000 sentences in the Penn Chinese Treebank (Penn CTB). We only examined 4 major word categories in the Penn Chinese Treebank due to inconsistency in the part-of-speech tag sets between the two corpora. The 4 major word categories were: noun (shown as NN, NR in Penn CTB; n, nr, ns, nz in PD corpus), verb (VV in Penn CTB; v, vd, vn in PD corpus), adjective (JJ in Penn CTB; a, ad, an in PD corpus) and adverb (AD in Penn CTB; d in PD corpus).

Segmentation and word POS tagging performance is measured in precision (P%), recall (R%) and F-score (F). Unknown words (NW) are those words not found in our word-dictionary, which include named entities and other new words. The unknown word rate (NW-Rate), the precision on unknown words (NW-Precision) and recall on total unknown words (NW-Recall) are given by:

$$\text{NW-Rate}=\frac{\#\text{ of unknown words}}{\text{total }\#\text{ of NW identified}} \qquad \text{NW-Precision}=\frac{\#\text{of valid unknown words}}{\text{total }\#\text{of NW identified}}$$

$$\text{NW-Recall}=\frac{\#\text{of valid unknown word}}{\text{total }\#\text{of NW in testing data}}$$

Table 1 shows that the above three systems achieve similar performances on the PK testing corpus. All of them were trained by the People's Daily corpus. For this reason, their performances were similar when the testing data had similar styles. But for other texts, the proposed character model performed much better than the word-based models in both recall and precision. This indicated that our approach performed better for unseen data.

Table 2 shows that our method for unknown word identification also outperforms the word-based method. We notice that word-based approaches and character-based approaches have similar precision on unknown word identification, however word-based approaches have much lower recall than character-based ones. The main reason for this is that word-based systems focus only on unknown words with proper word structures, but cannot recognize newly generated words, rare words, and other new words unlisted in the dictionary. A very high proportion of these types of unknown word in the SIGHAN testing data affects the recall of the word-based methods on unknown words. The experiments reveal that our method could effectively identify all kinds of new words. This is because our model has defined word building rules for all kinds of words.

Without a widely recognized testing standard, it is very hard to evaluate the performance on part-of-speech tagging. The results in Penn Chinese Treebank was better than that in the People's Daily Corpus since we examined all 42 POS tags in the People's Daily Corpus, but we only tested four major POS tags in Penn Chinese Treebank. Our approach is better than the word-based method for two test data sets. However, we could not conclude that our method was superior to the word-based method because of the limited testing approaches and testing data. A thorough empirical comparison among different approaches should be investigated in the future.

Table 1. Comparison of word segmentation based on SIGHAN open test sets

	PK		CTB		HK		AS	
	R%/ P%	F	R%/ P%	F	R%/ P%	F	R%/ P%	F
Chr-HMM	91.9/91.8	91.8	86.9/87.3	87.1	87.7/86.7	87.2	89.9/89.1	89.5
Chr-HMM+Dic	**95.9/96.7**	**96.3**	**92.7/93.5**	**93.1**	**91.1/91.9**	**91.5**	**92.3/93.9**	**93.1**
W-Bigram	94.7/95.4	95.1	87.4/86.8	87.1	88.7/83.7	86.3	87.9/85.1	86.5
W-HMM	94.6/95.1	94.9	88.6/89.2	88.9	90.7/89.1	89.9	90.7/87.2	89.0
Rank 1 in SIG	96.3/95.6	96.0	91.6/90.7	91.2	95.8/95.4	95.6	91.5/89.4	90.5
Rank 2 in SIG	96.3/94.3	95.3	91.1/89.1	90.1	90.9/86.3	88.6	89.2/85.3	87.3

Table 2. Accuracy of unknown word identification for SIGHAN open test sets

	PK			CTB			HK			AS		
Chr-HMM	UWR%	P%	R%	UWR%	P%	R%	UWR%	P%	R%	UWR%	P%	R%
Chr-HMM+Dic	2.3	56.2	54.8	10.4	68.8	64.4	9.7	61.4	58.4	8	65.4	62.9
W-Bigram	2.3	54.7	53.6	10.4	53.9	23.8	9.7	53.0	29.6	8	64.6	35.3
W-HMM	2.3	58.1	51.3	10.4	68.3	37.2	9.7	62.3	40.7	8	68.4	41.1

Table 3. Comparison of word part-of-speech tagging

	People Daily			Penn CTB		
	P%	R%	F-score	P%	R%	F-score
Chr-HMM	82.4%	82.5%	82.5	89.7%	88.5%	89.1
Chr-HMM+Dic	89.3	87.8	88.6	92.5	91.5	92.0
W-HMM	86.2%	85.4%	85.7	91.1%	90.8%	91.0

From Table 1 and Table 3, we notice that chr-HMM achieved 88% accuracy in word segmentation and 80% in part-of-speech tagging without a word-dictionary. Chr-HMM is a state-of-the-art Chinese morphology system without a word-dictionary. Its performance is comparable to some dictionary-based approaches (e.g., forward-maximum). This result indicates that our model has effectively captured most of the Chinese word building rules.

The results also show that chr-HMM+Dic outperformed the best SIGHAN word segmentation system on 3 out of the 4 SIGHAN open track test corpora, and achieved top 2 in the case of HK testing corpus.

3.2 Incorporation with Other Systems

The advantage of the proposed model is proficiency in describing word building rules and since many existing NLP application systems are weak in identifying new words, it is intuitive to integrate our model to existing systems and serves as a post-processing subsystem. In this subsection, we show how existing word segmentation systems could be improved using chr-HMM.

Given a segmentation result, we assume that unidentified new words may be a sequence of unattached characters. That is, all multiple-character words in the given result are considered correct, while single words, which might include unidentified new words will be rechecked by the chr-HMM. The entire process involves 3 steps:

1. Only character tags that are consistent with the position of the character in the word are listed for multi-character words.
2. The unattached characters are tagged with all possible character tags. In this way, the original segmentation result is converted into a group of character tagging candidates.
3. We then input these character tagging candidates into the chr-HMM to select the best one.

Consider an original result:

乔丹 [昨 日 从 谷 底] 强力 [反 弹] (Jordan bounced back strongly from the bottom yesterday)

The parts in brackets are the sequence of single characters where the new words may appear. The chr-HMM will list all possible character tags for these "unattached" characters. The parts outside the brackets are multiple-character words identified by the original system. They are assumed correct and maintain also positional information. Only the character tags, which are consistent with the positions of the character in the word are listed. The character tagging candidates for the above sample is given in Figure 3:

			…		…	…				
			vL	…	nL	nsL			…	
			tL	vL	nM	nrL			vM	
		…	nL	lM	nsF	nL			nM	…
nrF	nrL	tL	nM	vF	nrF	nM	vF	vL	vF	vL
nF	nL	nL	aM	pS	nF	vF	nF	nL	aF	nL
aF	aL	tS	tS	dS	nS	nS	aF	aL	aS	vS
乔	丹	昨	日	从	谷	底	强	力	反	弹

Fig. 3. Character tagging candidates for rechecking

Chr-HMM is then applied to the character tagging candidates and the best character tagging selected based on the probability of the candidates is output as the result. In this example, the result is: '*乔丹 (Jordan)* ***昨日*** *(yesterday)* ***从*** *(from)* ***谷底*** *(earth) 强力 (strongly)* ***反弹*** *(bound)*'. The three missing new words in the original system are identified by this post-processing subsystem.

We re-assigned the word segmentation results for all participants who have given permission to release data from the SIGHAN site (available for download from http://www.sighan.org/bakeoff2003). Table 4 enlists the performance of SIGHAN open test with and without chr-HMM. The participant numbers correspond to the sites listed in [9].

Table 4. Comparison of results with and without chr-HMM

Corpus	Site		R%	P%	F
AS	03	Before	89.2	85.3	87.2
		After	90.8	92.0	91.4
CTB	01	Before	88.7	87.6	88.1
		After	90.1	91.8	90.9
	03	Before	85.3	80.6	82.9
		After	86.4	87.8	87.1
	10	Before	91.1	89.1	90.1
		After	91.0	93.5	92.3
HK	03	Before	90.9	86.3	88.6
		After	89.4	91.0	90.2
PK	03	Before	94.1	91.1	92.5
		After	94.4	95.3	94.9
	10	Before	96.3	95.6	95.9
		After	95.6	97.7	96.7

From Table 4, it is obvious that word segmentation precision increases significantly, and at the same time, the corresponding recall remains the same or slightly declined. This implies that the chr-HMM retains the correct words by the original system and concurrently decreases significantly its errors.

4 Related Work

Although character features are very important in Chinese morphology, research in character-based approach is unpopular. Chooi-Ling Goh et al. [16], Jianfeng Gao et al. [8] and Huaping Zhang [3] adopted character information to handle unknown words; X. Luo [11], Yao Meng [12] and Shengfen Luo [17] each presented character-based parsing models for Chinese parsing or new-word extraction. T. Nakagawa used word-level information and character-level information for word segmentation [6]. Hwee Tou Ng et al. [5] investigated word-based and character-based approaches and proposed a maximum entropy character-based POS analyzer. Although the character tags proposed in this paper are essentially similar to some of the previous work mentioned above, here our focus is to integrate various word features with the character-based model in such a way that the probability of the model is undistorted. The proposed model is effective in acquiring word building rules. To our knowledge, our work is the first character-based approach, which outperforms the word-based approaches for SIGHAN open test. Also, our approach is versatile and can be easily integrated with existing morphological systems to achieve improved performance.

5 Conclusion and Future Works

A lexicon-constrained character model is proposed to capture word building rules using word features and character features. The combination of word and character features improves the performance of word segmentation and part-of-speech tagging. The proposed model can solve complicated issues in Chinese morphological analysis. The Chinese morphological analysis is generalized into a process of specific character tagging and word filtering. A lexicon supervises the character-based model to eliminate invalid character tagging candidates.

Our system outperformed the best SIGHAN word segmentation system in 3 out of the 4 SIGHAN open test sets. To our knowledge, our work is the first character-based approach, which performs better than word-based approaches for SIGHAN open test. In addition, the proposed method is versatile and can be easily integrated to any existing Chinese morphological system as a post-processing subsystem leading to enhanced performance.

In this paper, we focused on word features in character-based mode, and adopted HMM as the statistical model to identify the rules. Other statistical models, such as maximum entropy, boosting, support vector machine, etc., may also be suitable for this application. They are worth investigating. The data sparseness problem is practically non-existent in the character-based model for the Chinese character set is limited. However, odd characters are occasionally found in Chinese personal or place names. Some rules using named entity identification technique may help smoothen

this. In a broader view, the word building rules proposed in our model is simple enough for linguistic studies to better understand for example formation of Chinese words or even the Chinese language itself.

References

1. Andi Wu. Chinese Word Segmentation in MSR-NLP. In Proc. of SIGHAN Workshop, Sapporo, Japan, (2003) 127-175
2. GuoDong Zhou and Jian Su. A Chinese Efficient Analyzer Integrating Word Segmentation, Part-Of-Speech Tagging, Partial Parsing and Full Parsing. In Proc. Of SIGHAN Workshop, Sapporo, Japan, (2003) 78-83
3. Huaping Zhang, Hong-Kui Yu et al.. HHMM-based Chinese Lexical Analyzer ICTCLAS. In Proc. Of SIGHAN Workshop, Sapporo, Japan, (2003) 184-187
4. Nianwen Xue and Libin Shen. Chinese Word Segmentation as LMR Tagging. In Proc. Of SIGHAN Workshop, Sapporo, Japan, (2003) 176-179
5. Hwee Tou Ng, Low, Jin Kiat. *Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based?* In Proc. of EMNLP, Barcelona, Spain, (2004) 277-284
6. Tetsuji Nakagawa. Chinese and Japanese Word Segmentation Using Word-level and Character-level Information, In Proc. of the 20th COLING, Geneva, Switzerland, (2004) 466-472
7. Guohong Fu and Kang-Kwong Luke. *A Two-stage Statistical Word Segmentation System for Chinese*. In Proc. Of SIGHAN Workshop, Sapporo, Japan, (2003) 156-157
8. Jianfeng Gao, Andi Wu, Chang-Ning Huang et al. *Adaptive Chinese Word Segmentation*. In Proc. of 42nd ACL. Barcelona, Spain, (2004) 462-469
9. Richard Sproat and Thomas Emerson. *The First International Chinese Word Segmentation Bakeoff.* In Proc. Of SIGHAN Workshop, Sapporo, Japan, (2003) 133-143
10. X. Luo. A Maximum Entropy Chinese Character-based Parser. In Proc. of EMNLP. Sapporo, Japan, (2003) 192-199
11. Honglan Jin, Kam-Fai Wong, "A Chinese Dictionary Construction Algorithm for Information Retrieval", ACM Transactions on Asian Language Information Processing, 1(4):281-296, Dec. 2002.
12. Yao Meng, Hao Yu and Fumihito Nishino. 2004. *Chinese New Word Identification Based on Character Parsing Model*. In Proc. of 1st IJCNLP, Hainan, China, (2004) 489-496
13. Shiwen Yu, Huiming Duan, etal. 北京大学现代汉语语料库基本加工规范. 中文信息学报v(5), (2002) 49-64, 58-65
14. Maosong Sun and Benjamin K. T' Sou. *Ambiguity Resolution in Chinese Word Segmentation.* In Proc. of 10th Pacific Asia Conference on Language, Information & Computation, (1995) 121-126
15. Nianwen Xue, Fu-Dong Chiou and Martha Palmer. *Building a Large-scale Annotated Chinese Corpus*. In Proc. of the 19th COLING. Taibei, Taiwan, (2002)
16. Chooi-Ling GOH, Masayuki Asahara, Yuji Matsumoto. *Chinese Unknown Word Identification Using Character-based Tagging and Chunking*. In Proc. of the 41st ACL, Interactive Poster/Demo Sessions, Sapporo, Japan, (2003) 197-200
17. Shengfen Luo, Maosong Sun. 2003, *Two-character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measure*, In Proc. of the 2nd SIGHAN Workshop, Sapporo, Japan, (2003) 20-30