A Study of the Effectiveness of Suffixes for Chinese Word Segmentation

Xiaoqing Li † **Chengqing Zong**† **Keh-Yih Su** ‡
†National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
‡Behavior Design Corporation, Taiwan
{xqli, cqzong}@nlpr.ia.ac.cn,
kysu@bdc.com.tw

Abstract

We investigate whether suffix related features can significantly improve the performance of character-based approaches for Chinese word segmentation (CWS). Since suffixes are quite productive in forming new words, and OOV is the main error source for CWS, many researchers expect that suffix information can further improve the performance. With this belief, we tried several suffix related features in both generative and discriminative approaches. However, our experiment results have shown that significant improvement can hardly be achieved by incorporating suffix related features into those widely adopted surface features, which is against the commonly believed supposition. Error analysis reveals that the main problem behind this surprising finding is the conflict between the degree of reliability and the coverage rate of suffix related features.

1 Introduction

As words are the basic units for text analysis, Chinese word segmentation (CWS) is critical for many Chinese NLP tasks such as parsing and machine translation. Although steady improvements have been observed in previous CWS researches (Xue, 2003; Zhang and Clark, 2007; Wang et al., 2012; Sun et al., 2012), their performances are only acceptable for in-vocabulary (IV) words and are still far from satisfactory for those out-of-vocabulary (OOV) words. According to the Zipf's law (Zipf, 1949), which states that the frequency of a word is inversely proportional to its rank in the frequency table for a given corpus, it is unlikely to cover all the words of a language in the training corpus. OOV words are thus inevitable in real applications.

To further improve the performance for OOV words, various approaches have been proposed. Most of them aim to add additional resources, such as external dictionaries (Low et al., 2005; Zhao et al., 2010; Li et al., 2012) or unlabeled data (Zhao and Kit, 2008; Sun and Xu, 2011). However, additional resources are not always available and their coverage for OOV words is still limited. Researchers, especially linguists (Dong et al., 2010), thus seek to further improve the performance of OOV words by characterizing the word formation process (Li, 2011).

According to the internal structures of OOV words, they can be divided into three categories: (1) character-type related OOV, which consists of Arabic digits and foreign characters, and usually denotes time, date, number, English word, URL, etc. This kind of OOV can be well handled by rules or character-type features if the character-type information can be utilized (Low et al., 2005; Wang et al., 2012); (2) morpheme related OOV, which mainly refers to a compound word with prefix/suffix or reduplication (e.g. “高高兴兴” (happily)). According to (Wang et al., 2012), the errors related with suffix are the major type (more than 80%) within this category; (3) others (such as named entities, idioms, terminology, abbreviations, new words, etc.), which are usually irregular in structure and are difficult to handle without additional resources. Since extra knowledge about character-type and additional resources are forbidden in the

27th Pacific Asia Conference on Language, Information, and Computation pages 118－125

SIGHAN closed test (Emerson, 2005), which is widely adopted for performance comparison, we will focus on the second category to investigate how to use suffix related features in this paper.

Generally speaking, Chinese suffixes are very productive and many words can be formed in this way. For example, the word "旅行者" (traveler) is composed of a stem ("旅行", travel) and a suffix ("者", -er). Although the character and character co-occurrence features (adopted in most current approaches) are able to partially characterize the internal structure of words (Sun, 2010), and some OOV words are indeed correctly handled when compared to pure word-based approaches (Zhang et al., 2003; Gao et al., 2005), suffix related errors still remain as an important type of errors. Therefore, it is natural to expect that suffixes can be explicitly utilized to provide further help.

Furthermore, prefix/suffix related features were claimed to be useful for CWS in some previous works (Tseng et al., 2005; Zhang et al., 2006). However, in their works, the prefix/suffix features are just a part of adopted features. The performances before and after adopting prefix/suffix features are never directly compared. So we could not know how much improvement actually results from those prefix/suffix related features. Besides, those features have only been adopted under discriminative approaches (Xue, 2003; Peng, 2004). We would also like to know whether the suffix related features would be effective for the generative approach (Wang et al., 2009; Wang et al., 2010).

In comparison with the discriminative model, the generative model has the drawback that it cannot utilize trailing context in selecting the position tag (i.e. **B**eginning, **M**iddle, **E**nd and **S**ingle) (Xue, 2003) of the current character. Therefore, incorporating suffix information of the next character is supposed to be a promising supplement for the generative approach. So the real benefit of using suffixes is checked for the generative model first.

To make use of the suffix information more completely, a novel quantitative tagging bias feature is first proposed to replace the context-independent suffix list feature adopted in the literature. Compared with the original suffix-list feature, the proposed tagging bias feature takes the context into consideration and results less modeling error. A new generative model is then derived to incorporate the suffix related feature.

However, experimental results have shown that the performance cannot be considerably improved by adding suffix information, as what we expected. Furthermore, no improvement can be achieved with the suffix list when we re-implemented the discriminative approach of (Tseng et al., 2005; Zhang et al., 2006). This negative conclusion casts significant doubt on the above commonly believed supposition that suffix information can further improve the performance of CWS via incorporating it into surface features. The reasons for this surprising finding are thus studied and presented in this paper.

2 Extracting suffix information

In linguistic definition[1], a suffix is a morpheme that can be placed after a stem to form a new word. Also, a suffix cannot stand alone as a word. According to this definition, only a few characters can be regarded as suffixes, such as '者' (-er), '化' (-ize), '率' (rate), etc. However, the character '湖' (lake) in the words "昆明湖" (Kunming Lake) and "未名湖" (Weiming Lake) can help recognize those OOV words, although it can also appear as an independent word in the phrase "在/湖/中间" (in the middle of the lake). We thus loosen the constraint that a suffix cannot stand alone as a word in this paper to cover more such characters. That is, if a character tends to locate at the end of various words, it is regarded as if it plays the role of a suffix in those words. In this way, many named entities (such as the two location names mentioned above) will be also classified as suffix related words.

2.1 Difficulties in recognizing suffixes

Nonetheless, we cannot distinguish suffixes from those non-suffixes by just checking each character because whether a character is a suffix highly depends on the context. For example, the character '化' is a suffix in the word "初始化" (initial-ize). However, it becomes a prefix when it comes to the word "化纤" (chemical-fibre). Also, whether a character is a suffix varies with different annotation standards adopted by various corpora. For example, the character '厂' (factory) is a suffix in words such as "服装厂" (clothing-factory) in the PKU corpus provided by the SIGHAN 2005 Bakeoff (Emerson, 2005). Nevertheless, it is regarded as a single-character

[1] http://zh.wikipedia.org/wiki/%E8%A9%9E%E7%B6%B4

word in similar occasions in the MSR corpus. For these two reasons, suffixes cannot be directly recognized by simply locating some pre-specified characters prepared by the linguist.

2.2 Extracting a suffix-like list

Due to the difficulty in recognizing real suffixes, previous works (Tseng et al., 2005; Zhang et al., 2006) extract a suffix-like list beforehand from each corpus in context-free manner. Specifically, Tseng et al. (2005) considers characters that frequently appear at the end of those rare words as potential suffixes. In their approach, words that the numbers of occurrences in the training set are less than a given threshold are selected first, and then their ending characters are sorted according to their occurrences in those rare words. Afterwards, the suffix-like list is formed with those high-frequency characters. Zhang et al. (2006) constructs the list in a similar way, but without pre-extracting rare words.

In order to reduce the number of suffix errors resulted from the above primitive extraction procedure, we propose to obtain and use the suffix-list in a more prudent manner as follows:

- Having considered that suffix is supposed to be combined with different stems to form new words, we propose to use the *suffix productivity* as the criteria for extracting suffix list, which is defined as the size of the set $\{w \mid w \in IV, [w+sc] \in IV\}$, where *w* is a word in the training set, *sc* is a specific character to be decided if it should be extracted as a *suffix character*, and IV denotes in-vocabulary words. The cardinality of this set counts how many different IV words can be formed by concatenating the given suffix character to an IV word. Therefore, larger suffix productivity means that the given suffix character can be combined with more different stems to form new words, and is thus more likely to be a suffix.
- According to our investigation, most OOV with suffix are composed of a multi-character IV and a suffix, such as "旅行者" (i.e., "旅行" + "者"). So we set the suffix status for a given character to be true only when that character is in the suffix list and its previous character is the end of a multi-character IV word. In this way we can avoid many over-generalized errors (thus improve the precision for OOV with suffixes) and it only has little harm for the recall.

2.3 Adopting tagging bias information

There are two drawbacks to adopt the above suffix-like list: (1) The associated context that is required to decide whether a character should be regarded as a suffix is either completely not taken into account (in previous approaches) or treated too coarsely (in the above proposed approach). (2) The probability value (a finer information) that a given character acts as a suffix is not utilized; only a hard-decision flag (in or outside the list) is assigned to each character.

To overcome these two drawbacks, we introduce the *context-dependent tagging bias level*, which reflects the likelihood that the next character tends to be the beginning of a new word (or be a single-character word) based on the local context. This is motivated by the following observation: if the trailing character is biased towards 'S' or 'B', then the current character will prefer to be tagged as 'S' or 'E'; on the contrary, if the trailing character is biased towards 'M' or 'E', then the current character will prefer to be tagged as 'B' or 'M'.

Having considered that the surrounding context might be unseen for the testing instances, we introduce four different kinds of tagging bias probabilities as follows (and they will be trained in parallel for each character in the training-set):

- *Context-free tagging bias level* (qf_i): which is the quantized value of $P(t_{i+1} \in \{\text{E,M}\} \mid c_{i+1})$ that is estimated from the training corpus. In our experiments, we quantize $P(t_{i+1} \in \{\text{E,M}\} \mid c_{i+1})$ into five different intervals: [0.0-0.2], [0.2-0.4], [0.4-0.6], [0.6-0.8] and [0.8-1.0]; therefore, qf_i is a corresponding member of {-2, -1, 0, 1, 2}.
- *Left-context-dependent tagging bias level* (ql_i): Compared with qf_i, $P(t_{i+1} \in \{\text{E,M}\} \mid c_i^{i+1})$ is used instead of $P(t_{i+1} \in \{\text{E,M}\} \mid c_{i+1})$. The quantization procedure is the same.
- *Right-context-dependent tagging bias level* (qr_i): Compared with qf_i, $P(t_{i+1} \in \{\text{E,M}\} \mid c_{i+1}^{i+2})$ is used instead of $P(t_{i+1} \in \{\text{E,M}\} \mid c_{i+1})$. The quantization procedure is the same.
- *Surrounding-context-dependent tagging bias level* (qs_i): Compared with qf_i, $P(t_{i+1} \in \{\text{E,M}\} \mid c_i^{i+2})$ is used instead of $P(t_{i+1} \in \{\text{E,M}\} \mid c_{i+1})$. Quantization is the same.

3 Incorporating Suffix Information

3.1 For the generative model

Wang et al. (2009) proposed a character-based generative model for CWS as follows:

$$\overline{t}_1^n \equiv \arg\max_{t_1^n} \prod_{i=1}^{n} P([c,t]_i \,|\, [c,t]_{i-2}^{i-1}) \quad (1)$$

where $[c,t]_1^n$ is the associated character-tag-pair sequence for the given character sequence c_1^n.

To overcome the drawback that it cannot utilize trailing context, we propose to incorporate the suffix information of the *next* character (denoted by q_i), which can be either the suffix-list binary indicator or the above tagging bias level, into the model and reformulate it as follows:

$$\widehat{t}_1^n = \arg\max_{t_1^n} P(t_1^n \,|\, c_1^n, q_1^n) = \arg\max_{t_1^n} P(t_1^n, c_1^n, q_1^n)$$

$P(t_1^n, c_1^n, q_1^n)$ is then approximated by $\prod_{i=1}^{n} P([t,c,q]_i \,|\, [t,c,q]_{i-2}^{i-1})$, and its associated factor is further derived as below:

$$\begin{aligned} & P([t,c,q]_i \,|\, [t,c,q]_{i-2}^{i-1}) \\ =& P(q_i \,|\, [t,c]_{\mathrm{i}}, [t,c,q]_{i-2}^{i-1}) \times P([t,c]_i \,|\, [t,c,q]_{i-2}^{i-1}) \\ \approx& P(q_i \,|\, t_{i-1}^i, c_{i-2}^i) \times P([t,c]_i \,|\, [t,c]_{i-2}^{i-1}) \\ \approx& P_{tq[i]}(m_i \,|\, t_{\mathrm{i-1}}, c_{i-2}^i) \times P([t,c]_i \,|\, [t,c]_{i-2}^{i-1}) \end{aligned} \quad (2)$$

where m_i indicates whether t_{i} matches the suffix information of c_{i+1} or not, and $tq[i]$ specifies the corresponding type of probability factor to be adopted (i.e., qf_i, ql_i, qr_i, qs_i). For those three different suffix features (previous suffix-list, proposed suffix-list, and proposed tagging bias), m_i will be decided as follows:

- For the previous suffix-list feature, m_i will be a member of {Match, Violate, Neutral}. If c_{i+1} is in the suffix-list, when t_{i} is assigned with the position tag 'B' or 'M', m_i will be 'Match'; otherwise m_i will be 'Violate'. If c_{i+1} is not in the suffix-list, m_i will always be 'Neutral', no matter what position tag is assigned to t_{i}.
- For the proposed suffix-list feature, m_i will also be a member of {Match, Violate, Neutral}. If c_{i+1} is in the suffix list and c_i is the end of a multi-character IV word, when t_{i} is assigned position tag 'M', m_i will be 'Match'; otherwise m_i will be 'Violate'. If c_{i+1} is not in the suffix list or c_i is not the end of a multi-character IV word, m_i will always be 'Neutral'.
- For the proposed tagging bias feature, m_i will be a member of {Match[q_i], Violate[q_i], Neutral}, where q_i is a member of { qs_i, ql_i, qr_i, qf_i } and is selected according to whether the context c_i^{i+2} in the testing sentence is seen in the training corpus or not. Specifically, if c_i^{i+2} is seen in the training corpus, then q_i will be qs_i; else if c_i^{i+1} is seen, then q_i will be ql_i; else if c_{i+1}^{i+2} is seen, then q_i will be qr_i; otherwise, q_i will be qf_i. When $q_i > 0$ (i.e., $c_{\mathrm{i+1}}$ tends to be the beginning of a new word), if t_{i} is assigned 'S' or 'E', then m_i will be $\mathrm{Match}[q_i]$; otherwise, m_i will be $\mathrm{Violate}[q_i]$. On the contrary, when $q_i < 0$ (i.e., $c_{\mathrm{i+1}}$ tends not to be the beginning of a new word), if t_{i} is 'B' or 'M', then m_i will be $\mathrm{Match}[q_i]$, otherwise, m_i will be $\mathrm{Violate}[q_i]$. For example, if $q_i = 2$ and $t_{\mathrm{i}} = E$, then m_i will be 'Match[2]'. On the contrary, if $q_i = -2$ and $t_{\mathrm{i}} = E$, then m_i will be 'Violate[-2]'. Also, we will have four different $P_{tq[i]}(m_i \,|\, t_{\mathrm{i-1}}, c_{i-2}^i)$ (associated with {*qs*, *ql*, *qr*, *qf*}, respectively), and *tq*[*i*] indicates which one of them should be adopted at c_{i}. Afterwards, according to the context of each testing instance, a specific $P_{tq[i]}(m_i \,|\, t_{\mathrm{i-1}}, c_{i-2}^i)$ will be adopted.

It is reasonable to expect that the two factors in Equation 2 should be weighted differently in different cases. Besides, the second character-tag trigram factor is expected to be more reliable when c_{i-1}^i is seen in the training corpus. Therefore, these two factors are combined via log-linear interpolation. For the suffix-list feature, the scoring function will be:

$$\begin{aligned} Score(t_i) =& \alpha_k \times \log P([c,t]_i \,|\, [c,t]_{i-2}^{i-1}) \\ & + (1-\alpha_k) \log P(m_i \,|\, t_{\mathrm{i-1}}, c_{i-2}^i); \;\; 1 \le k \le 2 \end{aligned} \quad (3)$$

where α_k is selected according to whether c_{i-1}^i is seen. The values of α_k will be automatically decided in the development set via MERT (Och, 2003) procedure.

For the tagging bias feature, the scoring function will be:

$$Score(t_i) = \alpha_{tq,k} \times \log P([c,t]_i \mid [c,t]_{i-2}^{i-1}) + (1-\alpha_{tq,k}) \log P(m_i \mid t_{i-1}, c_{i-2}^{i}); 1 \le tq \le 4, 1 \le k \le 2 \quad (4)$$

where $\alpha_{tq,k}$ is selected according to which tagging bias probability factor is used and whether c_{i-1}^{i} is seen. Therefore, we will have eight different $\alpha_{tq,k}$ in this case.

3.2 For the discriminative model

We adopt the following feature templates under the maximum entropy approach that are widely adopted in previous works (Xue, 2003; Low et al., 2005):

$(a)\ C_n (n = -2, -1, 0, 1, 2);$
$(b)\ C_n C_{n+1} (n = -2, -1, 0, 1);$
$(c)\ C_{-1} C_1$

where C represents a character, and n denotes the relative position to the current character of concern.

To further utilize the suffix information, (Tseng et al., 2005) proposed a suffix-like list based feature as below.

$(d)\ s_0$, which is a binary feature indicating whether the current character of concern is in the list. In our modified approach, the suffix status will be true when the character c_0 is in the suffix-list and also c_{-1} is the end of a multi-character IV word.

Besides the above feature, (Zhang, 2006) also utilized some combinational features as follows:

$(e)\ c_0 s_{-1}, c_0 s_1, c_{-1} s_0, c_{-2} s_0$, where c denotes a character, s denotes the above suffix-like list feature.

In addition, we also tested the case of context-free tagging bias (proposed in Section 2.3), under this discriminative framework, by adding the following template.

$(f)\ qf$, which is the context-free tagging bias level. Please note that qs (also ql and qr) is not adopted because it will always be qs in the training-set (and thus will be over-fitted). Therefore, only qf is adopted to make the training and testing conditions consistent.

4 Experiments and Discussions

4.1 Setting

All the experiments are conducted on the corpora provided by SIGHAN Bakeoff 2005 (Emerson, 2005), which include Academia Sinica (AS), City University of Hong Kong (CITYU), Peking University (PKU) and Microsoft Research (MSR). For tuning the weights in Equation 3 and Equation 4, we randomly select 1% of the sentences from the training corpus as the development set.

For the generative approaches, the SRI Language Model Toolkit (Stolcke, 2002) is used to train $P([c,t]_i \mid [c,t]_{i-2}^{i-1})$ with the modified Kneser-Ney smoothing method (Chen and Goodman, 1996). The Factored Language Model in SRILM is adopted to train $P(m_i \mid t_{i-1}, c_{i-2}^{i})$, and it will sequentially back-off to $P(m_i \mid t_{i-1})$. For the discriminative approach, the ME Package provided by Zhang Le[2] is adopted to train the model. And trainings are conducted with Gaussian prior 1.0 and 300 iterations. In addition, the size of the suffix-like list in all approaches is set to 100[3], and the occurrences threshold for rare words in (Tseng et al., 2005) is set to 7. Typical F-score is adopted as the metric to evaluate the results.

4.2 Results of generative approaches

The segmentation results of using different generative models proposed in Section 3.1 are shown in Table 1. "Baseline" in the table denotes the basic generative model corresponding to Equation 1; "With Suffix-Like List" denotes the model that adopts the suffix-like list related features, corresponding to Equation 3; each sub-row right to it indicates the method used to extract the list. "With Tagging Bias" denotes the model that adopts tagging bias related features, corresponding to Equation 4. Bold entries indicate that they are statistically significantly different from their corresponding entries of the baseline model.

Table 1 shows that the improvement brought by the tagging bias approach is statistically significant[4] from the original model for three out of four corpora; however, the difference is not much. Also, for the suffix-like list approaches, the performance can only be slightly improved when the suffix-list is extracted and used in our

[2] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[3] This size is not explicitly given in their papers; so we tried several different values and find that it only makes little difference on the results. So is the threshold for rare words.

[4] The statistical significance test is done by the bootstrapping technique (Zhang et al., 2004), with sampling size of 2000 and confidence interval of 95%.

		PKU	AS	CITYU	MSR
Baseline		0.951	0.948	0.945	0.970
With Suffix-Like List	Tseng	0.951	0.948	0.946	0.970
	Zhang	0.951	0.948	0.946	0.970
	Proposed	0.952	0.949	**0.947**	0.970
With Tagging Bias		**0.953**	**0.950**	**0.947**	0.970

Table 1: Segmentation results for generative approaches in F-score

	PKU	AS	CITYU	MSR
Baseline	0.946	0.951	0.943	0.960
Tseng	0.946	0.949	0.942	0.961
Tseng+	0.946	0.949	0.942	0.960
Zhang	0.946	0.949	0.941	0.959
Zhang+	0.945	0.949	0.941	0.960
With *qf*	0.946	0.950	0.941	0.960

Table 2: Segmentation results for discriminative approaches in F-score

proposed way. To inspect if the quality of the suffix-list will affect the performance, we manually remove those characters which should not be regarded as suffixes in each list (such as Arabic numbers, and characters like "斯", "尔", which always appear at the end of transliteration). However, the performances are almost the same even with those cleaned lists (thus not shown in the table). The reasons will be found out and explained in Section 5.

4.3 Results of discriminative approaches

Table 2 shows the segmentation results for various discriminative approaches. 'Baseline' in the table denotes the discriminative model that adopts features (a)-(c) described in Section 3.2; 'Tseng' denotes the model with additional feature (d); and 'Tseng+' adopts the same feature set as 'Tseng', but the suffix-like list is obtained and used in our proposed way; similarly, the same interpretation goes for 'Zhang' and 'Zhang+'. Last, 'with *qf* ' denotes the model with additional feature (f), instead of features (d) and (e). Please note that *qs* (also *ql* and *qr*) is not adopted (explained above in Section 3.2).

The results in Table 2 show that neither the suffix-like list related feature nor the context-free tagging bias feature can provide any help for the discriminative approach. Similar to the generative approach, no significant benefit can be brought in even if the list is further cleaned by the human. This seems contradictory to the claims given at (Tseng et al., 2005; Zhang et al., 2006) and will be studied in the next section.

5 Problems Investigation

5.1 Suffix information is unreliable when associated context is not seen

Whether a character can act as a suffix is highly context dependent. Although context has been taken into consideration in our proposed suffix-list approach and tagging bias approach, the preference implied by the suffix list or tagging bias level becomes unreliable when the context is unfamiliar. Table 3 shows the percentage that the preference of different tagging bias factors matches the real tag in the training set. It can be seen that the matching rate (or the influence power) is higher with broader seen context. When no context is available (the last column; the suffix-list approach), it drops dramatically. As a result, many over-generalized words are produced when *qf* must be adopted. For example, two single-character words "该/局" (this bureau) are wrongly merged into a pseudo OOV "该局". As another example, the first three characters in the sequence "冠军/奖碟" (championship award-tray) are wrongly merged into a pseudo OOV "冠军奖" (championship-award). Because the related context "奖碟" is never seen for the character '奖', it is thus regarded as a suffix in this case (as it is indeed a suffix in many other cases such as "医学奖" (medicine-prize) and "一等奖" (first-prize)).

Corpus	*qs*	*ql*	*qr*	*qf*
PKU	0.996	0.977	0.923	0.686
AS	0.993	0.970	0.899	0.662
CITYU	0.997	0.976	0.919	0.653
MSR	0.992	0.970	0.898	0.662

Table 3: The matching rates of various tagging bias factors in the training set

Corpus	*qs*	*ql*	*qr*	*qf*
PKU	0.457	0.135	0.135	0.002
AS	0.374	0.083	0.082	0.004
CITYU	0.515	0.148	0.149	0.008
MSR	0.299	0.060	0.060	0.0003

Table 4: Unseen ratios for *qs* , *ql* , *qr* and *qf* in the testing set

5.2 Required context is frequently unobserved for testing instances

However, according to the empirical study of Zhao et al., (2010), the OOV rate can be linearly reduced only with an exponential increasing of

corpus size, roughly due to Zipf's law; and n-gram is expected to also follow this pattern (Marco, 2009). Therefore, the sparseness problem gets more serious for the n-gram with a larger "n" (i.e., with wider context) because its number of possible distinct types would become much greater. As a consequence, there will be much more unseen bigrams than unseen unigrams in the testing set (Of course, unseen trigrams will be even more). Table 4 shows the unseen ratios for *qs*, *ql*, *qr* and *qf* in the testing set. It is observed that the unseen ratio for *qs* is much larger than that for *qf*. However, according to the discussion in the previous subsection, the preference of tagging bias level is not reliable for *qf*. Therefore, more reliable a suffix-feature is, less likely it can be utilized in the testing-set. As the result, no significant improvement can be brought in by using suffix related features.

6 Conclusion

Since suffixes are quite productive in forming new words, and OOV is the main error source for all state-of-the-art CWS approaches, it is intuitive to expect that utilizing suffix information will further improve the performance. Some papers even claim that suffix-like list is useful in their discriminative models, though without presenting direct evidence. Against the above intuition, the empirical study of this paper reveals that when suffix related features are incorporated into those widely adopted surface features, they cannot considerably improve the performance of character-based generative and discriminative models, even if the context is taken into consideration. Error analysis reveals that the main problem behind this surprising finding is the conflict between the reliability and the coverage of those suffix related features. This conclusion is valuable for those relevant researchers in preventing them from wasting time on similar attempts.

Last, the reason that humans can distinguish suffixes correctly is largely due to their ability in utilizing associated syntactic and semantic knowledge of the plain text. We still believe suffix information can help for CWS if such knowledge can be effectively incorporated into the model. And this will be our future work.

Acknowledgments

The research work has been partially funded by the Natural Science Foundation of China under Grant No. 61003160 and Hi-Tech Research and Development Program ("863" Program) of China under Grant No. 2012AA011101 and 2012AA011102.

References

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 311-318.

Baroni Marco. 2009. Distributions in text. In A. Lüdeling and M. Kytö (eds.), *Corpus linguistics: An international handbook*. Mouton de Gruyter, Berlin, Germany.

Franze Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160-167, Sapporo, Japan.

Fuchun Peng, Fangfang Feng and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of COLING*, pages 562–568.

George Kingsley Zipf. 1949. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wisley. Oxford, UK.

Hai Zhao, Chang-Ning Huang, Mu Li and Bao-Liang Lu. 2010. A Unified Character-Based Tagging Framework for Chinese Word Segmentation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 9 (2). pages 1-32.

Hai Zhao and Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In *Sixth SIGHAN Workshop on Chinese Language Processing*.

Hai Zhao, Yan Song and Chunyu Kit. 2010. How Large a Corpus do We Need : Statistical Method vs. Rulebased Method. In *Proceedings of LREC-2010*. Malta.

Huaping Zhang, Hongkui Yu, Deyi Xiong and Qun Liu. 2003. HMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 184-187.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168-171.

Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2005. Chinese word segmentation and

named entity recogni-tion: a pragmatic approach. *Computational Linguistics*, 31(4), pages 531-574.

Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages. 161-164, Jeju Island, Korea.

Kun Wang, Chengqing Zong and Keh-Yih Su. 2009. Which is more suitable for Chinese word segmentation, the generative model or the discriminative one? In *Proceedings of PACLIC*, pages 827-834, Hong Kong, China.

Kun Wang, Chengqing Zong and Keh-Yih Su. 2010. A Character-Based Joint Model for Chinese Word Segmentation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1173-1181, Beijing, China.

Kun Wang, Chengqing Zong and Keh-Yih Su. 2012. Integrating Generative and Discriminative Character-Based Models for Chinese Word Segmentation. *ACM Transactions on Asian Language Information Processing*, Vol.11, No.2, June 2012, pages 7:1-7:41.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*, 8 (1). pages 29-48.

Ruiqiang Zhang, Genichiro Kikui and Eiichiro Sumita. 2006. Subword-based Tagging for Confidence-dependent Chinese Word Segmentation. In *Proceedings of the COLING/ACL*, pages 961-968, Sydney, Australia.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. *Technical Report TR-10-98, Harvard University Center for Research in Computing Technology*.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 123-133.

Weiwei Sun. 2010. Word-based and character-based word segmentation models: Comparison and combination. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1211-1219, Beijing, China.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese Word Segmentation Using Unlabeled Data. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 970-979, Edinburgh, Scotland, UK.

Xiaoqing Li, Kun Wang, Chengqing Zong and Keh-Yih Su. 2012. Integrating Surface and Abstract Features for Robust Cross-Domain Chinese Word Segmentation. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Pages 1653-1669, Mumbai, India.

Xu Sun, Houfeng Wang and Wenjie Li. 2012. Fast Online Training with Frequency-Adaptive Learning Rates for Chinese Word Segmentation and New Word Detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 253-262, Jeju Island, Korea.

Ying Zhang, Stephan Vogel and Alex Waibel, 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system. In *Proceedings of LREC*, pages 2051–2054.

Yue Zhang and Stephen Clark, 2007. Chinese Segmentation with a Word-Based Perceptron Algorithm. In *Proceedings of ACL*, pages 840-847, Prague, Czech Republic.

Zhengdong Dong, Qiang Dong and Changling Hao. 2010. Word segmentation needs change - from a linguist's view. In *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing*. Beijing, China.

Zhongguo Li. 2011. Parsing the Internal Structure of Words: A New Paradigm for Chinese Word Segmentation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1405--1414, Portland, Oregon, USA.