Which is More Suitable for Chinese Word Segmentation, the Generative Model or the Discriminative One? *

Kun Wang[a], Chengqing Zong[a], and Keh-Yih Su[b]

[a]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,
Room 1010, No. 95, Zhongguancun East Road, Haidian District, Beijing 100190, China
{kunwang, cqzong}@nlpr.ia.ac.cn
[b]Behavior Design Corporation,
2F, No. 5, Industry East Road IV, Science-Based Industrial Park, Hsinchu, Taiwan, ROC
kysu@bdc.com.tw

Abstract. Since the traditional word-based *n*-gram model, a generative approach, cannot handle those *out-of-vocabulary* (OOV) words in the testing-set, the character-based discriminative approach has been widely adopted recently. However, this discriminative model, though is more robust to OOV words, fails to deliver satisfactory performance for those *in-vocabulary* (IV) words that have been observed before. Having analyzed the word-based approach, its capability to handle the dependency between adjacent characters within a word, which is believed that the human adopts for doing segmentation, is found to account for its excellent performance for those IV words. To incorporate the intra-word characters dependency, a character-based approach with a generative model is thus proposed in this paper. The experiments conducted on the second SIGHAN Bakeoffs have shown that the proposed model not only achieves a good balance between those IV words and OOV words, but also outperforms the above-mentioned well-known approaches under the similar conditions.

Keywords: Chinese Word Segmentation, Generative Model, Discriminative Model.

1 Generative Model Versus Discriminative Model

Unlike English and other western languages, there is no space delimiter between adjacent Chinese words. Therefore, for most Chinese NLP applications, Chinese word segmentation (CWS) is the first task, which aims to find the corresponding word sequence from the given Chinese character sequence. Among various approaches for CWS, statistical methods have been increasingly applied in the past two decades.

According to the basic unit adopted to extract features, statistical approaches could be classified as either a *word-based approach* or a *character-based approach*. Besides, the word segmentation problem could also be formulated as either a *generative model* or a *discriminative model*. In terms of the above classification, the time-honored word-based model (Zhang et al., 2003; Gao et al., 2003) will be called as the *word-based generative approach*, while the well-known character-based tagging model (Xue, 2003; Ng and Low, 2004; Tseng et al., 2005) will be named as the *character-based discriminative approach*. Also, the word "model" will be loosely exchanged with the word "approach" when there is no confusion.

* The research work has been partially funded by the Natural Science Foundation of China under grant No.60736014, 60723005 and 90820303, the National Key Technology R&D Program under grant No. 2006BAH03B02, the Hi-Tech Research and Development Program (863 Program) of China under grant No. 2006AA010108-4, and also supported by the China-Singapore Institute of Digital Media as well. The authors thank Behavior Design Corporation for using their Generic-Beam-Search code according to the agreement.

The above two different kinds of classification are orthogonal to each other. However, in the literature papers that we have checked, almost all the word-based approaches adopt the *generative model*[1], and all the character-based approaches adopt the *discriminative model*. Before we argue why the proposed approach would be a better combination, a detailed discussion for the merits and drawbacks of both the word-based generative model and the character-based discriminative model is first given in the following, which would help to illustrate our motivation.

1.1 Word-Based Generative Model

The word-based generative model is formulated as follows.

$$WSeq^* = \arg\max_{WSeq} P(WSeq|c_1^n) \tag{1}$$

Where $WSeq \equiv w_1^m = [w_1, w_2, ...w_m]$ indicates a specific word sequence with *m* words, and c_1^n denotes the given sentence with *n* characters. The classical word-trigram model, expressed as $P(w_i|w_{i-2}, w_{i-1})$, is first formulated in the following.

$$P(w_1^m|c_1^n) = P(c_1^n|w_1^m) \times P(w_1^m)/P(c_1^n) \tag{2}$$

Since $P(c_1^n|w_1^m) = 1$ and $P(c_1^n)$ is the same for various $WSeq$ candidates, only $P(w_1^m)$ should be considered, and it could be further simplified with the second order Markov Chain assumption shown as below.

$$P(w_1^m) = \prod_{i=1}^{m} P(w_i|w_1^{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-2}^{i-1}) \tag{3}$$

In equation (3), dependency between characters within a word is implicitly taken care by regarding them as a joint event (treated as a single unit). This model works well when there are no *out-of-vocabulary* (OOV) words. However, this condition cannot be met in real applications. For example, named entities and numerical expressions are two kinds of OOV words which are frequently encountered. Since the associated candidates of those multi-character OOV words can not be generated during the searching process without OOV detection, it is impossible to identify them in the word-based approach. Most OOV words thus will be segmented into their corresponding sequences of uni-character-words. High *recall of IV words* (**R_{IV}**) and low *recall of OOV words* (**R_{OOV}**) are then obtained (see Table 2). In other words, the word-based models are vulnerable to those OOV words. Meanwhile, the overall precision rate would be also low, as those OOV words are forced to be segmented into more words with smaller size.

1.2 Character-Based Discriminative Model

The character-based discriminative model (Xue, 2003) treats segmentation as a tagging problem, which assigns a corresponding tag to each Chinese character and is formulated as follows.

$$P(t_1^n|c_1^n) = \prod_{k=1}^{n} P(t_k|t_1^{k-1}, c_1^n) \approx \prod_{k=1}^{n} P(t_k|t_{k-1}, c_{k-2}^{k+2}) \tag{4}$$

Where t_k indicates the corresponding position of character c_k in its associated word, and is a member of {***S**ingle, **B**egin, **M**iddle, **E**nd*} (abbreviated as ***S, B, M*** and ***E*** in the following) in our work. For example, the word "北京市 (Beijing City)" will be assigned with the corresponding tags as: "北/B (North) 京/M (Capital) 市/E (City)".

Compared with the word-based generative model, this approach is tolerable with OOV words. Since each multi-character OOV word will be converted into its corresponding sequence of character-tag-pairs (and the vocabulary size of those possible character-tag-pairs is limited), it

[1] According to Wikipedia (http://en.wikipedia.org/wiki/Generative_model), "Generative models contrast with discriminative models, in that a generative model is a full probability model of all variables, whereas a discriminative model provides a model only of the target variable(s) conditional on the observed variables"

is possible to correctly identify those OOV words. Therefore, this approach is robust to OOV words, and possesses a high R_{OOV}. However, lower R_{IV} is usually accompanied, as the dependency between adjacent characters within a word is no longer directly modeled (to be further explained in Section 2.1). Therefore, compared with the character-based discriminative approach, even the word-based unigram models possess a much higher R_{IV} (see Table 2).

2 Let the Character-Based Approach Adopt the Generative Model?

From the analysis above, it is clear that we must regard characters as basic-units to get high R_{OOV}. The remaining problem is how to further raise R_{IV} within the character-based framework. To explore the possible direction, we will first inspect what kinds of character-related clues that the human usually adopts to do word segmentation, and then integrate this clue into the character-based framework.

2.1 Adhesion and Dependency Between Adjacent Characters

Humans are known to use the adhesion between two adjacent characters (also know as character-bigram) as an important clue to do word segmentation. High adhesion usually implies that we will not put a word-break between these two characters, while low adhesion frequently indicates that we will have a word-break between them. With this observation, *Mutual Information*[2] (MI), a statistical measure closely related to the degree of adhesion, has been adopted in Sproat and Shin (1990) for judging whether a character-bigram is a bi-character word. MI and other measures are used to perform word segmentation in Sun et. al (1998) (in which 91.75% precision rate is reported).

To give a sense about how Character-Bigram MI within words distributes differently from that between words, various character-bigrams are collected from all corpora of SIGHAN Bakeoff 2005 (Emerson, 2005), and the associated MI is evaluated for each of them. Figure 1 gives the distributions of MI for the class of character-bigrams within words (shown by black bars) and the class of character-bigrams between words (shown by white bars). As indicated in the figure, the MI value for the character-bigram within words tends to be higher, which shows that it is a useful clue to judge whether the character-bigram should be segmented or not.

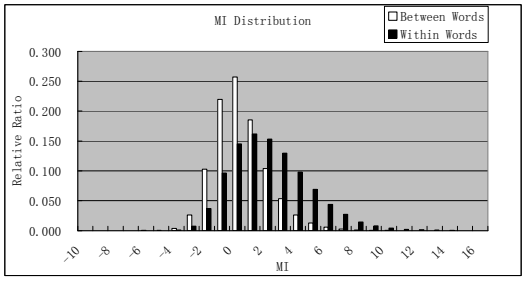

Figure 1: The distributions of MI.

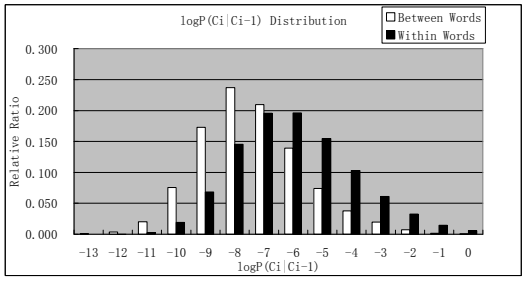

Figure 2: The distributions of $\log P(c_i|c_{i-1})$

Moreover, to study how the dependency between characters within a word, which is implicitly handled in Equation (3), makes the word-based approach possess significantly higher R_{IV}, $\log P(c_i|c_{i-1})$ is also calculated for various character-bigrams collected above. Figure 2 gives the distributions of $\log P(c_i|c_{i-1})$ for the class of character-bigrams within words (shown by black bars) and the class of character-bigrams between words (shown by white bars). As indicated in the figure, $\log P(c_i|c_{i-1})$ for the class of character-bigram within words also tends to have higher value, which explains why those IV words are more likely to be selected and

2 $MI(x,y) = \log \frac{P(x,y)}{P(x)\times P(y)}$, where $P(x,y)$ only counts the event that x precedes y (i.e., excluding the event that y precedes x).

high R_{IV} is thus resulted in. In fact, these two measures are considerably correlated to each other (especially for those character-bigrams within words), as shown in Table 1.

Table 1: The correlation coefficients between MI and $\log P(c_i|c_{i-1})$ of character-bigrams for within-words class and between-words class under various corpora.

Classes	AS	CITYU	MSR	PKU	Overall
Between Words	0.468	0.497	0.498	0.503	0.492
Within Words	0.678	0.709	0.712	0.718	0.699

2.2 Proposed Character-Based Generative Model

As explained in Section 1.1, the word-based approach is vulnerable to those OOV words. To overcome the OOV problem, the character-based approach must be adopted. However, the generative model should be also applied to handle the dependency within the character-bigram for each class (within-words and between-words), which has been shown to be useful in the last section. To take the advantage from both approaches mentioned above, w_i is first replaced with its corresponding sequence of $[character, tag]$ (denoted as $[c,t]_i$), where tag is the same as that adopted in the above character-based discriminative model. With this new representation, $P(w_1^m|c_1^n)$ could be re-derived based on the character as follows.

$$P(w_1^m|c_1^n) \equiv P([c,t]_1^n|c_1^n) = P(c_1^n|[c,t]_1^n) \times P([c,t]_1^n)/P(c_1^n) \tag{5}$$

Similar with Equation (2), only $P([c,t]_1^n)$ should be handled and could be further simplified to:

$$P([c,t]_1^n) \approx \prod_{k=1}^{n} P([c,t]_i|[c,t]_{i-k}^{i-1}) \tag{6}$$

As shown in the last section, $P(c_i|c_{i-1})$ within words inline to have a higher value than that between words. Therefore, for a bi-character-word (similarly for other multi-character-words), when $[c_{i-1}, c_i]$ is an IV word, $P([c,t]_i|[c,t]_{i-1})$ for $[t_{i-1} = B; t_i = E, M]$ is frequently higher than that for $[t_{i-1} = E, S; t_i = B, S]$ which correspond to those between-words character-bigrams. In other words, those IV words are more likely to be selected, and high R_{IV} is thus expected.

Unlike the word-based model specified above, this new approach regards the character as a unit. It is possible to correctly identify those multi-character OOV words, as their corresponding candidates could now be generated during the searching process. Besides, the capability to handle the dependency between adjacent characters under different classes (within-words and between-words), which has been shown to be important for getting high R_{IV} in the word-based approach, is still inherited in this model with the adopted generative form. Furthermore, as the basic unit in this proposed model is character, the vocabulary size of this model is much smaller than that of the word-based approach. Thus the data sparseness problem will not be severe.

Moreover, compared with the character-based discriminative approach, the proposed model still keeps the capability to handle OOV words, because it also regards the character as a unit. Also, since the generative form is adopted, the dependency between adjacent characters is now directly (and separately) modeled for each class (within-words and between-words), which will give sharper preference when the history of assignment is given. In contrast, the adhesion between adjacent characters is not explicitly modeled in the character-based discriminative approach, and is thus not used to assign tags.

3 Experiments and Results

We carried out our experiments on the data provided by SIGHAN Bakeoff 2005 (Emerson, 2005). To make a comparison with the baseline and previous work, only the closed tests[3] are

[3] According to Sighan Bakeoff 2005 regulation, the closed test could only use the training data directly provided. Any other data or information is forbidden, including the knowledge of characters set, punctuation and so on.

conducted. The metrics *Precision* (**P**), *Recall* (**R**), *F-measure* (**F**), *Recall of OOV* ($\mathbf{R_{OOV}}$) and *Recall of IV* ($\mathbf{R_{IV}}$) are used to evaluate the segmentation results. The balanced *F-measure* is F=2PR/(P+R).

3.1 Word-Based Generative Model and Character-Based Discriminative Model

We first extract a word list from the training-set as the vocabulary for the word-based generative approaches, and use SRI Language Modeling Toolkit[4] (SRILM) (Stolcke, 2002) to train various word *n*-gram models with modified Kneser-Ney smoothing (Chen and Goodman, 1998). Afterwards, a beam search decoder is applied to find out the best word sequence.

The segmentation results of word-based generative model are shown in Table 2. As expected, it shows that all those word-based *n*-gram models have high R_{IV} and very low R_{OOV} (even its unigram model outperforms the character-based discriminative approach in R_{IV}). After further analyzing the testing-set errors generated in the trigram model, we find that among total 16,781 error-patterns, 11,546 of them (69%) are to segment an OOV into a sequence of IV words, which clearly illustrates its drawback in handling OOV words and accounts for its low R_{OOV}.

For character-based discriminative approaches, various machine learning methods have been successfully applied. For example, Xue (2003) and Ng et. al (2004) use Maximum Entropy Model (ME) , Peng et. al (2004) and Tseng et. al (2005) use Conditional Random Fields (CRF) (Lafferty et al., 2001). Among them, CRF has been reported to give better performance (Zhang et al., 2006). Therefore, the package CRF++[5] is used to conduct the experiments for the character-based discriminative model, and the feature templates used are given by Ng and Low (2004), which have been widely adopted and reported in many papers, but excluding the ones forbidden by the closed test regulation. Those feature templates are listed as below:

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

Table 2 shows that the character-based discriminative model outperforms the word-trigram model on F-measure and R_{OOV}, but the latter gets higher R_{IV}. The low R_{IV} for the character-based discriminative model clearly shows the disadvantage without utilizing the dependency characteristic between adjacent characters within multi-character words. Among those 10,493 error-patterns resulted from the testing-sets, it is observed that 6,396 of them (61%) are to incorrectly segment an IV word-sequence, which clearly illustrates its weakness in handling IV words and accounts for its low R_{IV}.

3.2 Character-Based Generative Model

The proposed character-based generative model is also trained by using SRILM Toolkit with the same setting utilized by the word-based model. Table 3 shows the results of the proposed character-based generative model for various character *n*-gram-sizes ranging from *n*=2 to *n*=5. It illustrates that the character-trigram model significantly outperforms the character-bigram model over all four corpora, but almost no improvement could be observed if we keep increasing the *n*-gram size (only 4-gram improves a little on MSR corpus, as it has the largest average-word-length). This strongly suggests that our training data are inadequate to support more complex models other than trigram because of the data sparseness problem.

From the results, it can be seen that the proposed character-trigram generative model significantly exceeds the word-trigram generative model, and slightly outperforms the character-based discriminative model. Compared with the word-trigram approach, the proposed character-trigram model has dramatically raised the overall R_{OOV} from 0.047 to 0.541, with the cost of slightly degrading the overall R_{IV} from 0.987 to 0.977, which clearly shows that the handicap of the word-based model in handling OOV has been fixed.

In addition, unlike the character-based discriminative approach, the proposed trigram model is able to increase the overall R_{IV} from 0.963 to 0.977, while it pays the cost for degrading the overall R_{OOV} from 0.703 to 0.541. Also, the overall precision rate of the proposed trigram

[4] http://www.speech.sri.com/projects/srilm/
[5] http://crfpp.sourceforge.net/

model (0.950) is lower than that of the discriminative model (0.954). This implies that the proposed model tends to segment those OOV words into more words than the discriminative model does. However, the higher recall indicates that the proposed model segment more right words.

Table 2: Segmentation results of the word-based model (Word-unigram/bigram/trigram), the character-based discriminative model (Discriminative) and various proposed generative *n*-gram models.

AS	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$	**CITYU**	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$
Word-unigram	0.933	0.878	0.905	0.014	0.975	Word-unigram	0.924	0.851	0.886	0.162	0.984
Word-bigram	0.942	0.877	0.908	0.014	0.984	Word-bigram	0.928	0.851	0.888	0.162	0.990
Word-trigram	0.941	0.877	0.908	0.014	0.983	Word-trigram	0.929	0.852	0.889	0.162	0.990
Discriminative	0.956	0.946	**0.951**	0.704	0.967	Discriminative	0.940	0.945	0.943	0.701	0.960
New (Bigram)	0.954	0.934	0.944	0.509	0.975	New (Bigram)	0.949	0.932	0.941	0.603	0.976
New (Trigram)	0.958	0.938	0.948	0.518	0.978	New (Trigram)	0.951	0.937	**0.944**	0.609	0.978
New (4-gram)	0.958	0.938	0.948	0.518	0.978	New (4-gram)	0.951	0.938	0.944	0.610	0.978
New (5-gram)	0.957	0.938	0.948	0.518	0.977	New (5-gram)	0.951	0.938	0.944	0.610	0.978
MSR	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$	**PKU**	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$
Word-unigram	0.965	0.925	0.945	0.025	0.990	Word-unigram	0.939	0.909	0.924	0.016	0.972
Word-bigram	0.969	0.926	0.947	0.025	0.995	Word-bigram	0.949	0.913	0.931	0.016	0.982
Word-trigram	0.969	0.926	0.947	0.025	0.995	Word-trigram	0.949	0.913	0.930	0.016	0.982
Discriminative	0.963	0.967	0.965	0.723	0.969	Discriminative	0.943	0.954	0.948	0.689	0.952
New (Bigram)	0.965	0.955	0.960	0.522	0.977	New (Bigram)	0.949	0.946	0.948	0.494	0.965
New (Trigram)	0.974	0.967	0.970	0.561	0.985	New (Trigram)	0.952	0.951	**0.952**	0.503	0.968
New (4-gram)	0.974	0.967	**0.971**	0.568	0.985	New (4-gram)	0.952	0.952	0.952	0.511	0.967
New (5-gram)	0.974	0.967	0.971	0.568	0.985	New (5-gram)	0.952	0.952	0.952	0.510	0.968

Overall	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$
Word-unigram	0.943	0.897	0.919	0.047	0.980
Word-bigram	0.950	0.898	0.923	0.047	0.987
Word-trigram	0.950	0.898	0.923	0.047	0.987
Discriminative	0.953	0.954	0.954	0.703	0.963
New (Bigram)	0.955	0.943	0.949	0.527	0.973
New (Trigram)	0.960	0.950	**0.955**	0.541	0.977
New (4-gram)	0.960	0.950	0.955	0.541	0.977
New (5-gram)	0.960	0.950	0.955	0.545	0.977

3.3 Statistics of Remaining Errors

Having inspected those remaining testing-set tagging-errors (associated with characters) resulted from the character-trigram generative model and the character-based discriminative model, we divide them into two classes: (1) *Wrongly Broken*: Two adjacent characters should be joined while the model breaks them; (2) *Wrongly Jointed*: Two adjacent characters should be broken but the model joins them. Table 3 shows that both models share about 50% of their tagging-errors (e.g., the last row, labeled as *Overall*, shows that 7,068 errors among total 13,093 ones from the generative model are shared). To illustrate that the dependency between adjacent characters does affect segmentation performance, those tagging-errors are classified into two classes according to their MI values. Since MI of those unseen character-bigrams could not be reliably estimated, it could only affect the performance of those seen character-bigrams. Low MI and High MI are classified by the MI value that two probability distribution curves in Figure 1 cross each other, which is about 1.5.

In the last row (Overall) under Table 3, it could be observed that the proposed generative model generates less percentage of wrongly broken errors than the character-based model does when those seen-bigrams are classified as "High MI" (0.165 versus 0.171, shown in bold, in the column "Wrongly Broken"); and it also generates less percentage of wrongly joined errors when those seen-bigrams are classified as "Low MI" (0.107 versus 0.149, shown in bold, in the

column “Wrongly Joined”). The conclusion that the proposed model mimics the human behavior more closely thus could be drawn.

Table 3: Statistics for tagging-errors under the discriminative model and the generative model of character-based approaches.

Corpus	Model	Total Errors	Shared Errors	Unseen Bigram errors		Seen Bigram errors			
						Wrongly Broken		Wrongly Joined	
				Wrongly Broken	Wrongly Jointed	Low MI	High MI	Low MI	High MI
AS	Discriminative	4803	2650	0.177	0.147	0.200	0.253	0.129	0.093
	Generative	4860		0.296	0.072	0.212	0.260	0.087	0.073
CITYU	Discriminative	1985	772	0.150	0.364	0.141	0.158	0.093	0.094
	Generative	1767		0.389	0.166	0.171	0.109	0.079	0.086
MSR	Discriminative	2923	1359	0.127	0.282	0.173	0.182	0.196	0.112
	Generative	2605		0.298	0.154	0.242	0.109	0.119	0.078
PKU	Discriminative	4207	2287	0.099	0.285	0.135	0.126	0.167	0.188
	Generative	3861		0.252	0.170	0.150	0.109	0.139	0.181
Overall	Discriminative	13918	7068	0.139	0.248	0.167	**0.171**	**0.149**	0.126
	Generative	13093		0.296	0.130	0.194	**0.165**	**0.107**	0.108

4 Related Works

Since the character-based discriminative approach was first proposed by Xue (2003), it has been widely adopted and further developed by various researchers. For example, Asahara et al. (2005) use the character-based approach to first identify the OOV candidates and then integrate them into the system. Their system achieves the best result in the AS corpus in Sighan Bakeoff 2005 contest. Tseng et al. (2005) add the information of word-prefixes and word-suffixes to overcome the drawbacks of character-based approaches, and they get the best results in the remaining three corpora in that contest. Afterwards, Zhang et al. (2006) use a sub-word tagging approach to utilize the sub-word information. All of them adopt the character-based discriminative approaches. The only state-of-the-art word-based model proposed recently is Zhang and Clark (2007), which uses Perceptron, a discriminative method. The comparison between those models mentioned above is given in Table 4. It shows that the proposed model achieves a good balance between those IV words and OOV words, and also competitive results.

Table 4: Segmentation results of different Models

AS	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$	**CITYU**	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$
Asahara	0.952	0.951	0.952	0.696	0.963	Tseng	0.941	0.946	0.943	0.698	0.961
Zhang (CRF)	0.956	0.947	0.951	0.649	0.969	Zhang (CRF)	0.952	0.949	0.951	0.741	0.969
Our model	0.958	0.938	0.948	0.518	0.978	Our model	0.951	0.938	0.944	0.610	0.978
Zhang & Clark	N/A	N/A	0.946	N/A	N/A	Zhang & Clark	N/A	N/A	0.951	N/A	N/A
MSR	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$	**PKU**	**R**	**P**	**F**	$\mathbf{R_{OOV}}$	$\mathbf{R_{IV}}$
Tseng	0.962	0.966	0.964	0.717	0.968	Tseng	0.953	0.946	0.950	0.636	0.972
Zhang (CRF)	0.972	0.969	0.971	0.712	0.976	Zhang (CRF)	0.947	0.955	0.951	0.748	0.959
Our model	0.974	0.967	0.971	0.568	0.985	Our model	0.952	0.952	0.952	0.511	0.967
Zhang & Clark	N/A	N/A	0.972	N/A	N/A	Zhang & Clark	N/A	N/A	0.945	N/A	N/A

5 Conclusion

Since the traditional word-trigram generative model cannot handle those OOV words, it has very poor performance when OOV words are encountered. In addition, the popular character-based discriminative approach does not utilize the adhesion and dependency between adjacent characters. Therefore, it gives unsatisfactory performance for those IV words. To combine the strengths of these two camps, the character-based generative model is thus proposed in this

paper to let the character-based approach adopt the generative form. The experiments have shown that this new approach has achieved good balance between those IV words and OOV words. Furthermore, the statistics of remaining errors have shown that the proposed model mimics the human behavior more closely than the classic character-based discriminative model.

Moreover, the learning process of this character *n*-gram generative approach is found to be ten times faster than that of the CRF discriminative model. That gives additional advantage to the proposed approach when huge training data is at hand.

Reference

Asahara, Masayuki, Kenta Fukuoka, Ai Azuma, Chooi-Ling Goh, Yotaro Watanabe, Yuji Matsumoto and Takashi Tsuzuki, 2005. Combination of machine learning methods for optimum chinese word segmentation. In *the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 134–137, Jeju, Korea.

Chen, Stanley F. 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*.

Emerson, Thomas, 2005. The Second International Chinese Word Segmentation Bakeoff. In *the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 123-133, Jeju, Korea.

Gao, Jianfeng, Mu Li and Chang-Ning Huang, 2003. Improved Source-Channel Models for Chinese Word Segmentation. In *Proc. of ACL*, pages 272-279.

Lafferty, John, Andrew McCallum and Fernando Pereira, 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the 18th International Conference on Machine Learning*, pages 282-289.

Ng, Hwee Tou and Jin Kiat Low, 2004. Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. In *Proc. of EMNLP*, pages 277-284.

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

Sproat, Richard and Chilin Shih, 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4).pages 336-351.

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

Sun, Maosong, Dayang Shen and Benjamin K Tsou, 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proc. of COLING/ACL*, pages 1265-1271.

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

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

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

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

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