A Character-Based Joint Model for Chinese Word Segmentation

Kun Wang and **Chengqing Zong**
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Science
{kunwang,cqzong}@nlpr.ia.ac.cn

Keh-Yih Su
Behavior Design Corporation
Kysu@bdc.com.tw

Abstract

The character-based tagging approach is a dominant technique for Chinese word segmentation, and both discriminative and generative models can be adopted in that framework. However, generative and discriminative character-based approaches are significantly different and complement each other. A simple joint model combining the character-based generative model and the discriminative one is thus proposed in this paper to take advantage of both approaches. Experiments on the Second SIGHAN Bakeoff show that this joint approach achieves 21% relative error reduction over the discriminative model and 14% over the generative one. In addition, closed tests also show that the proposed joint model outperforms all the existing approaches reported in the literature and achieves the best F-score in four out of five corpora.

1 Introduction

Chinese word segmentation (CWS) plays an important role in most Chinese NLP applications such as machine translation, information retrieval and question answering. Many statistical methods for CWS have been proposed in the last two decades, which can be classified as either *word-based* or *character-based*. The word-based approach regards the word as the basic unit, and the desired segmentation result is the best word sequence found by the search process. On the other hand, the character-based approach treats the word segmentation task as a character tagging problem. The final segmentation result is thus indirectly generated according to the tag assigned to each associated character. Since the vocabulary size of possible character-tag-pairs is limited, the character-based models can tolerate *out-of-vocabulary* (OOV) words and have become the dominant technique for CWS in recent years.

On the other hand, statistical approaches can also be classified as either adopting a *generative model* or adopting a *discriminative model*. The generative model learns the joint probability of the given input and its associated label sequence, while the discriminative model learns the posterior probability directly. Generative models often do not perform well because they make strong independence assumptions between features and labels. However, (Toutanova, 2006) shows that generative models can also achieve very similar or better performance than the corresponding discriminative models if they have a structure that avoids unrealistic independence assumptions.

In terms of the above dimensions, methods for CWS can be classified as:

1) The word-based generative model (Gao et al., 2003; Zhang et al., 2003), which is a well-known approach and has been used in many successful applications;

2) The word-based discriminative model (Zhang and Clark, 2007), which generates word candidates with both word and character features and is the only word-based model that adopts the discriminative approach;

3) The character-based discriminative model (Xue, 2003; Peng et al., 2004; Tseng et al., 2005; Jiang et al., 2008), which has become the dominant method as it is robust on OOV words and is capable of handling a range of different features, and it has been adopted in many previous works;

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1173–1181,
Beijing, August 2010

4) The character-based generative model (Wang et al., 2009), which adopts a character-tag-pair-based *n*-gram model and achieves comparable results with the popular character-based discriminative model.

In general, character-based models are much more robust on OOV words than word-based approaches do, as the vocabulary size of characters is a closed set (versus the open set of that of words). Furthermore, among those character-based approaches, the generative model and the discriminative one complement each other in handling *in-vocabulary* (IV) words and OOV words. Therefore, a character-based joint model is proposed to combine them.

This proposed joint approach has achieved good balance between IV word recognition and OOV word identification. The experiments of closed tests on the second SIGHAN Bakeoff (Emerson, 2005) show that the joint model significantly outperforms the baseline models of both generative and discriminative approaches. Moreover, statistical significance tests also show that the joint model is significantly better than all those state-of-the-art systems reported in the literature and achieves the best F-score in four of the five corpora tested.

2 Character-Based Models for CWS

The goal of CWS is to find the corresponding word sequence for a given character sequence. Character-based model is to find out the corresponding tags for given character sequence.

2.1 Character-Based Discriminative Model

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

$$P(t_1^n \mid c_1^n) = \prod_{k=1}^{n} P(t_k \mid t_1^{k-1}, c_1^n) \approx \prod_{k=1}^{n} P(t_k \mid c_{k-2}^{k+2}) \qquad (1)$$

Where t_k is a member of {***B**egin*, ***M**iddle*, ***E**nd*, ***S**ingle*} (abbreviated as B, M, E and S from now on) to indicate the corresponding position of character c_k in its associated word. 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)".

Since this tagging approach treats characters as basic units, the vocabulary size of those possible character-tag-pairs is limited. Therefore, this method is robust to OOV words and could possess a high *recall of OOV words* (R_{OOV}). Although the dependency between adjacent tags/labels can be addressed, the dependency between adjacent characters within a word cannot be directly modeled under this framework. Lower *recall of IV words* (R_{IV}) is thus usually accompanied (Wang et al., 2009).

In this work, the character-based discriminative model is implemented by adopting the feature templates given by (Ng and Low, 2004), but excluding those ones that are forbidden by the closed test regulation of SIGHAN (e.g., $Pu(C_0)$: whether C_0 is a punctuation). Those feature templates adopted are listed below:

$(a)\, C_n (n = -2, -1, 0, 1, 2);$

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

$(c)\, C_{-1} C_1$

For example, when we consider the third character "奥" in the sequence "北京奥运会", template (a) results in the features as following: C_{-2}=北, C_{-1}=京, C_0=奥, C_1=运, C_2=会, and template (b) generates the features as: $C_{-2}C_{-1}$=北京, $C_{-1}C_0$=京奥, C_0C_1=奥运, C_1C_2=运会, and template (c) gives the feature $C_{-1}C_1$=京运.

2.2 Character-Based Generative Model

To incorporate the dependency between adjacent characters in the character-based approach, (Wang et al., 2009) proposes a character-based generative model. In this approach, word w_i is first replaced with its corresponding sequence of [character, tag] (denoted as [*c*, *t*]), where tag is the same as that adopted in the above character-based discriminative model. With this representation, this model can be expressed as:

$$\begin{aligned} &P(w_1^m \mid c_1^n) \equiv P([c,t]_1^n \mid c_1^n) \\ &= P(c_1^n \mid [c,t]_1^n) \times P([c,t]_1^n) \big/ P(c_1^n) \end{aligned} \qquad (2)$$

Since $P(c_1^n \mid [c,t]_1^n) \equiv 1$ and $P(c_1^n)$ is the same for various candidates, only $P([c,t]_1^n)$ should be considered. It can be further simplified with Markov Chain assumption as:

$$P([c,t]_1^n) \approx \prod_{i=1}^{n} P([c,t]_i \mid [c,t]_{i-k}^{i-1}). \qquad (3)$$

Compared with the character-based discriminative model, this generative model keeps the capability to handle OOV words because it also regards the character as basic unit. In addition, the dependency between adjacent

宿	Gold and Discriminative Tag: M					Generative Trigram Tag: E				
Tag probability:	B/0.0333		E/0.2236			M/0.7401		S/0.0030		
Feature / Tag	C_{-2}	C_{-1}	C_0	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	$C_{-2}C_{-1}$	$C_{-1}C_0$	C_0C_1	C_1C_2	$C_{-1}C_1$
B	-1.4375	0.1572	0.0800	0.2282	0.7709	0.2741	0.0000	0.0000	-0.6718	0.0000
E	1.3558	0.1910	0.7229	**-1.2696**	**-0.5970**	0.0049	0.0921	0.0000	0.8049	0.0000
M	1.1071	-0.5527	-0.3174	**2.9422**	**0.4636**	-0.1708	0.0000	0.0000	-0.9700	0.0000
S	-1.0254	0.2046	-0.4856	-1.9008	-0.6375	0.0000	0.0000	0.0000	0.8368	0.0000
者	Gold and Discriminative Tag: E					Generative Trigram Tag: S				
Tag probability:	B/0.0009		E/0.8138			M/0.0012		S/0.1841		
Feature / Tag	C_{-2}	C_{-1}	$\boldsymbol{C_0}$	$\boldsymbol{C_1}$	C_2	$C_{-2}C_{-1}$	$C_{-1}C_0$	C_0C_1	C_1C_2	$C_{-1}C_1$
B	0.3586	0.4175	0.0000	-0.7207	0.4626	0.0085	0.0000	0.0000	0.0000	0.0000
E	0.3666	0.0687	**4.5381**	**2.8300**	-0.0846	0.0000	0.0000	-1.0279	0.6127	0.0000
M	-0.5657	-0.4330	1.8847	0.0000	-0.0918	0.0000	0.0000	0.0000	0.0000	0.0000
S	-0.1595	-0.0532	**2.7360**	**1.8223**	-0.2862	-0.0024	0.0000	1.0494	0.7113	0.0000

Table 1: The corresponding lambda weight of features for "露宿者" in the sentence "[該] [處] [的] [露宿者] [只] [有] [數] [人]". In the Feature column and Tag row, the value is the corresponding lambda weight for the feature and tag under ME framework. The meanings of those features are explained in Section 2.1.

characters is now directly modeled. This will give sharper preference when the history of assignment is given. Therefore, this approach not only holds robust IV performance but also achieves comparable results with the discriminative model. However, the OOV performance of this approach is still lower than that of the discriminative model (see in Table 5), which would be discussed in the next section.

3 Problems with the Character-Based Generative Model

The character-based generative model can handle the dependency between adjacent characters and thus performs well on IV words. However, this generative trigram model is derived under the second order Markov Chain assumption. Future character context (i.e., C_1 and C_2) is thus not utilized in the model when the tag of the current character (i.e., t_0) is determined. Nevertheless, the future context would help to select the correct tag when the associated trigram has not been observed in the training-set, which is just the case for those OOV words. In contrast, the discriminative one could get help from the future context in this case. The example given in the next paragraph clearly shows the above situation.

At the sentence "該(that) 處(place) 的(of) 露宿者(street sleeper) 只(only) 有(have) 數(some) 人(person) (There are only some street sleepers in that place)" in the CITYU corpus, "露/B宿/M者/E(street sleeper)" is observed to be an OOV word, while "露/B 宿/E(sleep on the street)" is an IV word, where the associated tag of each character is given after the slash symbol. The character-based generative model wrongly splits "露宿者" into two words "露/B 宿/E" and "者/S (person)", as the associated trigram for "露宿者" is not seen in the training set. However, the discriminative model gives the correct result for "宿/M" and the dominant features come from its future context "者" and "只". Similarly, the future context "只" helps to give the correct tag to "者/E". Table 1 gives the corresponding lambda feature weights (under the Maximum Entropy (ME) (Ratnaparkhi, 1998) framework) for "露宿者" in the discriminative model. It shows that in the column of "C_1" below "宿", the lambda value associated with the correct tag "M" is 2.9422, which is the highest value in that column and is far greater than that of the wrong tag "E" (i.e., -1.2696) assigned by the generative model. Which indicates that the future feature "C_1" is the most useful feature for tagging "宿".

The above example shows the character-based generative model fails to handle some OOV words such as "露宿者" because this approach cannot utilize future context when it is indeed required. However, the future context for the generative model scanning from left to right is just its past context when it scans from right to left. It is thus expected that this kind of

errors will be fixed if we let the model scans from both directions, and then combine their results. Unfortunately, it is observed that these two scanning modes share over 90% of their errors. For example, in CITYU corpus, the left-to-right scan generates 1,958 wrong words and the right-to-left scan results 1,947 ones, while 1,795 of them are the same. Similar behavior can also be observed on other corpora.

To find out what are the problems, 10 errors that are similar to "露宿者" are selected to examine. Among those errors, only one of them is fixed, and "露宿者" still cannot be correctly segmented. Having analyzed the scores of the model scanning from both directions, we found that the original scores (from left-to-right scan) at the stages "者" and "宿" indeed get better if the model scans from right-to-left. However, the score at the stage "露" deteriorates because the useful feature "者" (a past non-adjacent character for "露" when scans form right-to-left) still cannot be utilized when the past context "宿者" as a whole is unseen, when the related probabilities are estimated via modified Kneser-Ney smoothing (Chen and Goodman, 1998) technique.

Two scanning modes seem not complementing each other, which is out of our original expectation. However, we found that the character-based generative model and the discriminative one complement each other much more than the two scanning modes do. It is observed that these two approaches share less than 50% of their errors. For example, in CITYU corpus, the generative approach generates 1,958 wrong words and the discriminative one results 2,338 ones, while only 835 of them are the same.

The statistics of the remaining errors resulted from the generative model and the discriminative model is shown in Table 2. As shown in the table, it can be seen that the generative model and the discriminative model complement each other on handling IV words and OOV words (In the "IV Errors" column, the number of "G+D-" is much more than the "G-D+", while the behavior is reversed in the "OOV Errors" column).

IV Errors			OOV Errors		
G+D-	**G-D+**	**G-D-**	**G+D-**	**G-D+**	**G-D-**
12,027	4,723	7,481	2,384	6,139	3,975

Table 2: Statistics for remaining errors of the character-based generative model and the discriminative one on the second SIGHAN Bakeoff ("G+D-" in the "IV Errors" column means that the generative model segments the IV words correctly but the discriminative one gives wrong results. The meanings of other abbreviations are similar with this one.).

4 Proposed Joint Model

Since the performance of both IV words and OOV words are important for real applications, we need to combine the strength from both models. Among various combining methods, log-linear interpolation combination is a simple but effective one (Bishop, 2006). Therefore, the following character-based joint model is proposed, and a parameter α is used to weight the generative model in a cross-validation set.

$$\begin{aligned} Score(t_k) = \alpha \times \log(P([c,t]_k \,|\, [c,t]_{k-2}^{k-1})) \\ + (1-\alpha) \times \log(P(t_k \,|\, c_{k-2}^{k+2})) \end{aligned} \quad (4)$$

Where t_k indicates the corresponding position of character c_k, and $\alpha\ (0.0 \le \alpha \le 1.0)$ is the weight for the generative model. $Score(t_k)$ will be used during searching the best sequence. It can be seen that these two models are integrated naturally as both are character-based.

Generally speaking, if the "G(or D)+" has a strong preference on the desired candidate, but the "D(or G)-" has a weak preference on its top-1 incorrect candidate, then this combining method would correct most "G+D- (also G-D+)" errors. On the other hand, the advantage of combining two models would vanish if the "G(or D)+" has a weak preference while the "D(or G)-" has a strong preference over their top-1 candidates. In our observation, these two models meet this requirement quite well.

5 Weigh Various Features Differently

For a given observation, intuitively each feature should be trained only once under the ME framework and its associated weight will be automatically learned from the training corpus. However, when we repeat the work of (Jiang et al., 2008), which reports to achieve the state-of-art performance in the data-sets that we adopt, it has been found that some features (e.g., C_0) are unnoticeably trained several times in their model (which are implicitly generated from different feature templates used in the paper). For example, the feature C_0 actually

Corpus	Abbrev.	Encoding	Training Size (Words/Type)	Test Size (Words/Type)	OOV Rate
Academia Sinica (Taipei)	AS	Unicode/Big5	5.45M/141K	122K/19K	0.046
City University of Hong Kong	CITYU	Unicode/Big5	1.46M/69K	41K/9K	0.074
Microsoft Research (Beijing)	MSR	Unicode/CP936	2.37M/88K	107K/13K	0.026
Peking University	PKU(ucvt.)	Unicode/CP936	1.1M/55K	104K/13K	0.058
	PKU(cvt.)	Unicode/CP936	1.1M/55K	104K/13K	0.035

Table 3: Corpus statistics for the second SIGHAN Bakeoff

appears twice, which is generated from two different templates C_n (with n=0, generates C_0) and $[C_0C_n]$ (used in (Jiang et al., 2008), with n=0, generates $[C_0C_0]$). The meanings of features are illustrated in Section 2.1. Those repetitive features also include $[C_{-1}C_0]$ and $[C_0C_1]$, which implicitly appear thrice. And it is surprising to discover that its better performance is mainly due to this implicit feature repetition but the authors do not point out this fact. As all the features adopted in (Jiang et al., 2008) possess binary values, if a binary feature is repeated n times, then it should behave like a real-valued feature with its value to be "n", at least in principle. Inspired by the above discovery, accordingly, we convert all the binary-value features into their corresponding real-valued features. After having transformed binary features into their corresponding real-valued ones, the original discriminative model is re-trained under the ME framework.

This new implementation, which would be named as the character-based discriminative-plus model, just weights various features differently before conducting ME training. Afterwards, it is further combined with the generative trigram model, and is called the character-based joint-plus model.

6 Experiments

The corpora provided by the second SIGHAN Bakeoff (Emerson, 2005) were used in our experiments. The statistics of those corpora are shown in Table 3.

Note that the PKU corpus is a little different from others. In the training set, Arabic numbers and English characters are in full-width form occupying two bytes. However, in the testing set, these characters are in half-width form occupying only one byte. Most researchers in the SIGHAN Bakeoff competition performed a conversion before segmentation (Xiong et al., 2009). In this work, we conduct the tests on both unconverted (ucvt.) case and converted (cvt.) case. After the conversion, the OOV rate of converted corpus is obviously lower than that of unconverted corpus.

To fairly compare the proposed approach with previous works, we only conduct *closed tests*[1]. The metrics *Precision* (**P**), *Recall* (**R**), *F-score* (**F**) (F=2PR/(P+R)), *Recall of OOV* ($\mathbf{R_{OOV}}$) and *Recall of IV* ($\mathbf{R_{IV}}$) are used to evaluate the results.

6.1 Character-Based Generative Model and Discriminative Model

As shown in (Wang et al., 2009), the character-based generative trigram model significantly exceeds its related bigram model and performs the same as its 4-gram model. Therefore, SRI Language Modeling Toolkit[2] (Stolcke, 2002) is used to train the trigram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998). Afterwards, a beam search decoder is applied to find out the best sequence.

For the character-based discriminative model, the ME Package[3] given by Zhang Le is used to conduct the experiments. Training was done with Gaussian prior 1.0 and 300, 150 iterations for AS and other corpora respectively. Table 5 gives the segmentation results of both the character-based generative model and the discriminative model. From the results, it can be seen that the generative model achieves comparable results with the discriminative one and they outperform each other on different corpus. However, the generative model exceeds the discriminative one on R_{IV} (0.973 vs. 0.956) but loses on R_{OOV} (0.511 vs. 0.680). It illustrates that they complement each other.

[1] According to the second Sighan Bakeoff 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 set, etc.

[2] http://www.speech.sri.com/projects/srilm/

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

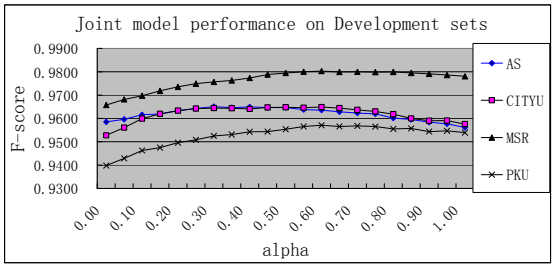

Figure 1: Development sets performance of Character-based joint model.

Corpus	Set	Words	OOV Num	OOV Rate
AS	Development	17,243	445	0.026
	Testing	122,610	5,308/5,311	0.043/0.043
MSR	Development	17,324	355	0.020
	Testing	106,873	2,829/2,833	0.026/0.027
CITYU	Development	12,075	537	0.044
	Testing	40,936	3,028/3,034	0.074/0.074
PKU	Development	13,576	532	0.039
	Testing (ucvt.)	104,372	6,006/6,054	0.058/0.058
	Testing (cvt.)	104,372	3,611/3,661	0.035/0.035

Table 4: Corpus statistics for Development sets and Testing sets. A "/" separates the OOV number (or OOV rate) with respect to the original training sets and the new training sets.

6.2 Character-Based Joint Model

For the character-based joint model, a development set is required to obtain the weight α for its associated generative model. A small portion of each original training corpus is thus extracted as the development set and the remaining data is regarded as the new training-set, which is used to train two new parameter-sets for both generative and discriminative models associated.

The last 2,000, 600, 400, and 300 sentences for AS, MSR, CITYU, and PKU are extracted from the original training corpora as their corresponding development sets. The statistics for new data sets are shown in Table 4. It can be seen that the variation of the OOV rate could be hardly noticed. The F-scores of the joint model, versus different α, evaluated on four development sets are shown in Figure 1. It can be seen that the curves are not sharp but flat near the top, which indicates that the character-based joint model is not sensitive to the α value selected. From those curves, the best suitable α for AS, CITYU, MSR and PKU are found to be 0.30, 0.60, 0.60 and 0.60, respectively. Those alpha values will then be adopted to conduct the experiments on the testing sets.

Corpus	Model	R	P	F	R_{OOV}	R_{IV}
AS	G	0.958	0.938	0.948	0.518	0.978
	D	0.955	0.946	0.951	0.707	0.967
	D-Plus	0.960	0.948	0.954	0.680	0.973
	J	0.962	0.950	**0.956**	0.679	0.975
	J-Plus	0.963	0.949	**0.956**	0.652	0.977
CITYU	G	0.951	0.937	0.944	0.609	0.978
	D	0.941	0.944	0.942	0.708	0.959
	D-Plus	0.951	0.952	0.952	0.720	0.970
	J	0.957	0.951	0.954	0.691	0.979
	J-Plus	0.959	0.952	**0.956**	0.700	0.980
MSR	G	0.974	0.967	0.970	0.561	0.985
	D	0.957	0.962	0.960	0.719	0.964
	D-Plus	0.965	0.967	0.966	0.675	0.973
	J	0.974	0.971	**0.972**	0.659	0.983
	J-Plus	0.975	0.970	**0.972**	0.632	0.984
PKU (ucvt.)	G	0.929	0.933	0.931	0.435	0.959
	D	0.922	0.941	0.932	0.620	0.941
	D-Plus	0.934	0.949	0.941	0.649	0.951
	J	0.935	0.946	0.941	0.561	0.958
	J-Plus	0.937	0.947	**0.942**	0.556	0.960
PKU (cvt.)	G	0.952	0.951	0.952	0.503	0.968
	D	0.940	0.951	0.946	0.685	0.949
	D-Plus	0.949	0.958	0.953	0.674	0.958
	J	0.954	0.958	0.956	0.616	0.966
	J-Plus	0.955	0.958	**0.957**	0.610	0.967
Overall	G	0.953	0.946	0.950	0.511	0.973
	D	0.944	0.950	0.947	0.680	0.956
	D-Plus	0.952	0.955	0.953	0.676	0.965
	J	0.957	0.955	0.956	0.633	0.971
	J-Plus	0.958	0.955	**0.957**	0.621	0.973

Table 5: Segmentation results of various character-based models on the second SIGHAN Bakeoff, the generative trigram model (G), the discriminative model (D), the discriminative-plus model (D-Plus), the joint model (J) and the joint-plus model (J-Plus).

As shown in Table 5, the joint model significantly outperforms both the character-based generative model and the discriminative one in F-score on all the testing corpora. Compared with the generative approach, the joint model increases the overall R_{OOV} from 0.510 to 0.633, with the cost of slightly degrading the overall R_{IV} from 0.973 to 0.971. This shows that the joint model holds the advantage of the generative model on IV words. Compared with the discriminative model, the proposed joint model improves the overall R_{IV} from 0.956 to 0.971, with the cost of degrading the overall R_{OOV} from 0.680 to 0.633. It clearly shows that the joint model achieves a good balance between IV words and OOV words and achieves the best F-scores obtained so far (21% relative error reduction over the discriminative model and 14% over the generative model).

6.3 Weigh Various Features Differently

Inspired by (Jiang et al., 2008), we set the real-value of C_0 to be 2.0, the value of $C_{-1}C_0$ and C_0C_1 to be 3.0, and the values of all other features to be 1.0 for the character-based discriminative-plus model. Although it seems reasonable to weight those closely relevant features more (C_0 should be the most relevant feature for assigning tag t_0), both implementations seem to be equal if their corresponding lambda-values are also updated accordingly. However, Table 5 shows that this new discriminative-plus implementation (D-Plus) significantly outperforms the original one (overall F-score is raised from 0.947 to 0.953) when both of them adopt real-valued features. It is not clear how this change makes the difference.

Similar improvements can be observed with two other ME packages. One anonymous reviewer pointed out that the duplicated features should not make difference if there is no regularization. However, we found that the duplicated features would improve the performance whether we give Gaussian penalty or not.

Afterwards, this new implementation and the generative trigram model are further combined (named as the joint-plus model). Table 5 shows that this joint-plus model also achieves better results compared with the discriminative-plus model, which illustrates that our joint approach is an effective and robust method for CWS. However, compared with the original joint model, the new joint-plus approach does not show much improvement, regardless of the significant improvement made by the discriminative-plus model, as the additional benefit generated by the discriminative-plus model has already covered by the generative approach (Among the 6,965 error words corrected by the discriminative-plus model, 6,292 (90%) of them are covered by the generative model).

7 Statistical Significance Tests

Although Table 5 has shown that the proposed joint (joint-plus) model outperforms all the baselines mentioned above, we want to know if the difference is statistically significant enough to make such a claim. Since there is only one testing set for each training corpus, the bootstrapping technique (Zhang et al., 2004) is adopted to conduct the tests: Giving an

Models		AS	CITYU	MSR	PKU (ucvt.)	PKU (cvt.)
A	B					
G	D	<	~	>	~	>
D-Plus	G	>	>	<	>	>
D-Plus	D	>	>	>	>	>
J	G	>	>	>	>	>
J	D	>	>	>	>	>
J-Plus	G	>	>	>	>	>
J-Plus	D-Plus	>	>	>	~	>
J-Plus	J	~	>	~	>	>

Table 6: Statistical significance test of F-score among various character-based models.

testing-set T_0, additional *M-1* new testing-sets $T_0,\ldots,T_{M-1}$ (each with the same size of T_0) will be generated by repeatedly re-sampling data from T_0. Then, we will have a total of M testing-sets (M=2000 in our experiments).

7.1 Comparisons with Baselines

We then follow (Zhang et al., 2004) to measure the 95% confidence interval for the discrepancy between two models. If the confidence interval does not include the origin point, we then claim that system A is significantly different from system B. Table 6 gives the results of significant tests among various models mentioned above. In this table, ">" means that system A is significantly better than B, where as "<" denotes that system A is significantly worse than B, and "~" indicates that these two systems are not significantly different.

As shown in Table 6, the proposed joint model is significantly better than the two baseline models on all corpora. Similarly, the proposed joint-plus model also significantly outperforms the generative model and the discriminative-plus model on all corpora except on the PKU(ucvt.). The comparison shows that the proposed joint (also joint-plus) model indeed exceeds each of its component models.

7.2 Comparisons with Previous Works

The above comparison mainly shows the superiority of the proposed joint model among those approaches that have been implemented. However, it would be interesting to know if the joint (and joint-plus) model also outperforms those previous state-of-the-art systems.

The systems that performed best for at least one corpus in the second SIGHAN Bakeoff are first selected for comparison. This category includes (Asahara et al., 2005) (denoted as

Asahara05) and (Tseng et al., 2005) [4] (**Tseng05**). (Asahara et al., 2005) achieves the best result in the AS corpus, and (Tseng et al., 2005) performs best in the remaining three corpora. Besides, those systems that are reported to exceed the above two systems are also selected. This category includes (Zhang et al., 2006) (**Zhang06**), (Zhang and Clark, 2007) (**Z&C07**) and (Jiang et al., 2008) (**Jiang08**). They are briefly summarized as follows. (Zhang et al., 2006) is based on sub-word tagging and uses a confidence measure method to combine the sub-word CRF (Lafferty et al., 2001) and rule-based models. (Zhang and Clark, 2007) uses perceptron (Collins, 2002) to generate word candidates with both word and character features. Last, (Jiang et al., 2008)[5] adds repeated features implicitly based on (Ng and Low, 2004). All of the above models, except (Zhang and Clark, 2007), adopt the character-based discriminative approach.

All the results of the systems mentioned above are shown in Table 7. Since the systems are not re-implemented, we cannot generate paired samples from those *M* testing-sets. Instead, we calculate the 95% confidence interval of the joint (also joint-plus) model. Afterwards, those systems can be compared with our proposed models. If the F-score of system B does not fall within the 95% confidence interval of system A (joint or joint-plus), then they are statistically significantly different.

Table 8 gives the results of significant tests for those systems mentioned in this section. It shows that both our joint-plus model and joint model exceed (or are comparable to) almost all the state-of-the-art systems across all corpora, except (Zhang and Clark, 2007) at PKU(ucvt.). In that special case, (Zhang and Clark, 2007)

Participants \ Corpus	AS	CITYU	MSR	PKU (ucvt.)	PKU (cvt.)
Asahara05	0.952	0.941	0.958	N/A	0.941
Tseng05	0.947	0.943	0.964	N/A	0.950
Zhang06	0.951	0.951	0.971	N/A	0.951
Z&C07	0.946	0.951	**0.972**	**0.945**	N/A
Jiang08	0.953	0.948	0.966	0.937	N/A
Our Joint	**0.956**	0.954	**0.972**	0.941	0.956
Our Joint-Plus	**0.956**	**0.956**	**0.972**	0.942	**0.957**

Table 7: Comparisons of F-score with previous state-of-the-art systems.

Systems A	Systems B	AS	CITYU	MSR	PKU (ucvt.)	PKU (cvt.)
J	Asahara05	>	>	>	N/A	>
	Tseng05	>	>	>	N/A	>
	Zhang06	>	~	~	N/A	>
	Z&C07	>	>	~	<	N/A
	Jiang08	>	>	>	>	N/A
J-Plus	Asahara05	>	>	>	N/A	>
	Tseng05	>	>	>	N/A	>
	Zhang06	>	>	~	N/A	>
	Z&C07	>	>	~	<	N/A
	Jiang08	~	>	>	>	N/A

Table 8: Statistical significance test of F-score for previous state-of-the-art systems.

outperforms the joint-plus model by 0.3% on F- score (0.4% for the joint model). However, our joint-plus model exceeds it more over AS and CITYU corpora by 1.0% and 0.5%, respectively (1.0% and 0.3% for the joint model). Thus, it is fair to say that both our joint model and joint-plus model are superior to the state-of-the-art systems reported in the literature.

8 Conclusion

From the error analysis of the character-based generative model and the discriminative one, we found that these two models complement each other on handling IV words and OOV words. To take advantage of these two approaches, a joint model is thus proposed to combine them. Experiments on the Second SIGHAN Bakeoff show that the joint model achieves 21% error reduction over the discriminative model (14% over the generative model). Moreover, closed tests on the second SIGHAN Bakeoff corpora show that this joint model significantly outperforms all the state-of-the-art systems reported in the literature.

Last, it is found that weighting various features differently would give better result. However, further study is required to find out the true reason for this strange but interesting phenomenon.

[4] We are not sure whether (Asahara et al., 2005) and (Tseng et al., 2005) performed a conversion before segmentation in PKU corpus. In this paper, we followed previous works, which cited and compared with them.

[5] The data for (Jiang et al., 2008) given at Table 7 are different from what were reported at their paper. In the communication with the authors, it is found that the script for evaluating performance, provided by the SIGHAN Bakeoff, does not work correctly in their platform. After the problem is fixed, the re-evaluated real performances reported here deteriorate from their original version. Please see the announcement in Jiang's homepage (http://mtgroup.ict.ac.cn/~jiangwenbin/papers/error_correction.pdf).

Acknowledgement

The authors extend sincere thanks to Wenbing Jiang for his helps with our experiments. Also, we thank Behavior Design Corporation for using their Generic-Beam-Search code and show special thanks to Ms. Nanyan Kuo for her helps with the Generic-Beam-Search code.

The research work has been partially funded by the Natural Science Foundation of China under Grant No. 60975053, 90820303 and 60736014, the National Key Technology R&D Program under Grant No. 2006BAH03B02, and also the Hi-Tech Research and Development Program ("863" Program) of China under Grant No. 2006AA010108-4 as well.

References

Masayuki Asahara, 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 *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 134–137, Jeju, Korea.

Christopher M. Bishop, 2006. *Pattern recognition and machine learning*. New York: Springer

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*.

Michael Collins, 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*, pages 1-8, Philadelphia.

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

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

Wenbin Jiang, Liang Huang, Qun Liu and Yajuan Lu, 2008. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *Proceedings of ACL*, pages 897-904.

John Lafferty, Andrew McCallum and Fernando Pereira, 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pages 282-289.

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

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.

Adwait Ratnaparkhi, 1998. Maximum entropy models for natural language ambiguity resolution. University of Pennsylvania.

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

Kristina Toutanova, 2006. Competitive generative models with structure learning for NLP classification tasks. In *Proceedings of EMNLP*, pages 576-584, Sydney, Australia.

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.

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.

Ying Xiong, Jie Zhu, Hao Huang and Haihua Xu, 2009. Minimum tag error for discriminative training of conditional random fields. *Information Sciences*, 179 (1-2). pages 169-179.

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

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

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.

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.