Integrating Surface and Abstract Features for Robust Cross-Domain Chinese Word Segmentation

LI Xiaoqing[1], WANG Kun[1], ZONG Chengqing[1] and SU Keh-Yih[2]
(1) National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
(2) Behavior Design Corporation, Taiwan
{xqli, kunwang, cqzong}@nlpr.ia.ac.cn, kysu@bdc.com.tw

ABSTRACT

Current character-based approaches are not robust for cross domain Chinese word segmentation. In this paper, we alleviate this problem by deriving a novel enhanced character-based generative model with a new abstract aggregate candidate-feature, which indicates if the given candidate prefers the corresponding position-tag of the longest dictionary matching word. Since the distribution of the proposed feature is invariant across domains, our model thus possesses better generalization ability. Open tests on CIPS-SIGHAN-2010 show that the enhanced generative model achieves robust cross-domain performance for various OOV coverage rates and obtains the best performance on three out of four domains. The enhanced generative model is then further integrated with a discriminative model which also utilizes dictionary information. This integrated model is shown to be either superior or comparable to all other models reported in the literature on every domain of this task.

KEYWORDS : Chinese Word Segmentation, Cross-Domain, Robust Feature, Utilize Dictionary

Proceedings of COLING 2012: Technical Papers, pages 1653–1670,
COLING 2012, Mumbai, December 2012.

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. However, current state-of-the-art character-based approaches fail to give robust performance for cross-domain tests (Gao and Vogel, 2010; Huang et al., 2010; Jiang and Dong, 2010; Wang et al., 2012, Sun et al., 2012), despite their acceptable performance for in-domain tests. For example, with the same PKU-News training corpus, the best system in SIGHAN-2005 (Emerson, 2005) achieved a high in-domain performance (96.9% in F-score) in open[1] tests, while the best cross-domain performance for the medicine domain only achieved 93.8%[2] in CIPS-SIGHAN-2010 (Zhao and Liu, 2010) open tests.

Their poor performances result not only from the fact that out-domains have higher out-of-vocabulary (OOV) rates (please refer Table 1 in Section 3.1), but also from the fact that various domains frequently possess different tag distributions for the same character, especially for those out-domain OOV words (which are also technical terms most of times). For example, "酸" (acid) is a typical suffix of medical terms, such as "尿酸" (uric acid) and "氨基酸" (amino acid), and is frequently tagged as "E" in the medicine domain. However, it is usually a single-character word which means 'sour' in general text, and should be tagged as "S". As a result, those surface features (such as character n-grams) adopted in the approaches mentioned above have more difficulty in identifying the correct position-tag, which is a member of {**B**eginning, **M**iddle, **E**nd, **S**ingleton}(Xue, 2003), of a character within an OOV word in this case. Since many technical OOV words appear in the out-domain text, the performance thus degrades sharply. For instance, when we test the main stream character-based discriminative approach for medicine domain in CIPS-SIGHAN-2010 contest, 27.7% of the wrong segmented words are OOV terms.

On the other hand, unlike surface features, the distribution of the proposed abstract feature, which checks if the given candidate prefers the corresponding position-tag of the longest dictionary matching word, is almost invariant across different domains. Enhanced discriminative approaches (Low et al., 2005; Zhao et al., 2010) which adopt this abstract feature thus show better generalization capability in our cross-domain tests.

To adopt the above matching-longest-word feature, a dictionary is required. It is well known that the domain dictionary provides direct and reliable hints for deciding the position-tags of characters within a covered OOV word. Furthermore, unlike named entity, technical terms of a specific domain usually can be considerably covered by a domain dictionary. Since domain dictionaries are frequently available for NLP related projects (e.g., technical manual translation), they can thus provide big help in real applications.

However, the above dictionary related feature utilized in discriminative approaches (Low et al., 2005; Zhao et al., 2010) cannot be directly adopted by a generative model[3]. Therefore, we propose a new tag-matching-status feature for checking if the selected position-tag matches the longest dictionary-matching-word, and derive a novel enhanced character-based generative model. The proposed feature not only induces an additional probabilistic factor but also possesses richer information in comparison with the one adopted in previous works.

[1] Unlike close test, the open test can use any language resource, not restricted to training data only.

[2] 93.8% only corresponds to 23.7% sentence accuracy rate (average 23.3 words per sentence in the medicine corpus), evaluated from its associated 94.0% recall-rate, while 96.9% corresponds to 46.9% sentence accuracy.

[3] It is required by the integrated approach given at Section 4.2.

Several factors that might affect the performance of the new model are studied in this paper: including the context information, the OOV coverage rate of the dictionary, and the weight of the new factor in the model. We evaluated our final system on the CIPS-SIGHAN-2010 Bakeoff data. The obtained results not only convincingly demonstrate the effectiveness of the proposed model for cross-domain CWS, but also achieve the best performance on 3 out of 4 domains in the open test. Afterwards, the proposed enhanced generative model is integrated with another enhanced discriminative model (Low et al., 2005) to further improve the performance, and achieves the best performance on all the tested corpora.

The remainder of this paper is organized as: Section 2 discusses how to incorporate dictionary information and section 3 describes the proposed models. Empirical results and error analysis are presented in section 4 and 5. Section 6 reviews the related work.

2 Dictionary related features

2.1 Word-ID or Word-Matching-Indicator?

Given a dictionary, there are two kinds of features that can be utilized: word-ID, which are binary features that fire only when the word matches one specific word entry, and Word-Matching-Indicator (e.g. TM defined in Section 2.3), which checks the relationship between the assigned position tag of the current character and the dictionary words within local context. Since the statistics of OOV words can never be learnt from the training corpus, the approaches that adopt word-ID as features (Zhang and Clark, 2007; Sun, 2010; Zhang and Clark, 2011) cannot really utilize the information of the OOV words kept in the dictionary. On the contrary, the word-matching-indicator is applicable for both IV and OOV words kept in the dictionary. This feature thus provides valuable information for those OOV words covered by the dictionary. Therefore, based on the positions of those dictionary matching words, two dictionary-related features (i.e., Dictionary Coverage Status and Tag Matching Status, to be specified later) are proposed in this paper, and they will be incorporated into the character-based generative model.

2.2 Dictionary Coverage Status

Let c_i be the i-th character in a given sentence. To check whether there are ambiguities with those dictionary matching words at c_i (and what kind of ambiguities it has), we propose the Dictionary Coverage Status feature, which is a member of {No-Dictionary-Word, No-Ambiguity, Crossed-Ambiguity[4], Included-Ambiguity, Mixed-Ambiguity} that are defined below. This status depends only on the given sentence and the dictionary, and is irrelevant to the position tag assigned to the character. Let D be the given dictionary which only contains multi-character words, and $c_{[i:j]}$ denotes the string from c_i to c_j (including c_j), then the conditions for "Included-Ambiguity" and "Crossed-Ambiguity" are defined below.

(A) Conditions for Included-Ambiguity (IA):

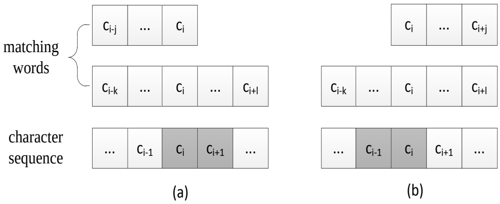
(1) Both c_i and c_{i+1} will be assigned "IA" if they meet the following condition (Figure 1(a)):

$$\exists j, l > 0, k \geq j : \{c_{[i-j:i]}, c_{[i-k:i+l]}\} \subseteq D \,;$$

(2) Both c_{i-1} and c_i will be assigned "IA" if they meet the following condition (Figure 1(b)):

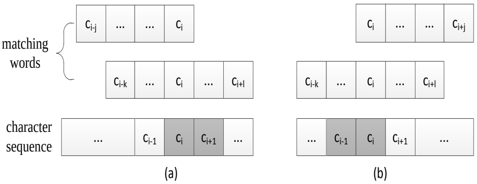
$$\exists j, k > 0, l \geq j : \{c_{[i:i+j]}, c_{[i-k:i+l]}\} \subseteq D \,.$$

[4] Please note that ambiguity status is traditionally defined on words, but ours is defined on characters.



matching words
character sequence
c_{i-j} … c_i
c_{i-k} … c_i … c_{i+l}
… c_{i-1} c_i c_{i+1} …
(a)
c_i … c_{i+j}
c_{i-k} … c_i … c_{i+l}
… c_{i-1} c_i c_{i+1} …
(b)

FIGURE 1 – Cases for Included-Ambiguous characters (marked in grey)



matching words
character sequence
c_{i-j} … … c_i
c_{i-k} … c_i … c_{i+l}
… c_{i-1} c_i c_{i+1} …
(a)
c_i … … c_{i+j}
c_{i-k} … c_i … c_{i+l}
… c_{i-1} c_i c_{i+1} …
(b)

FIGURE 2 – Cases for Crossed-Ambiguous characters (marked in grey)

(B) Conditions for Crossed-Ambiguity (CA):

(1) Both c_i and c_{i+1} will be assigned "CA" if they meet the following condition (see Figure 2(a)):

$$\exists j,l>0, 0\le k<j:\{c_{[i-j:i]}, c_{[i-k:i+l]}\}\subseteq D;$$

(2) Both c_{i-1} and c_i will be assigned "CA" if they meet the following condition (Figure 2(b)):

$$\exists j,k>0, 0\le l<j:\{c_{[i:i+j]}, c_{[i-k:i+l]}\}\subseteq D.$$

Then Dictionary Coverage Status at c_i (denoted by DC_i) can be decided as follows:

$$DC_i=\begin{cases}\text{No-Dictionary-Word, if no matching word is found;}\\ \text{Included-Ambiguity, if only (A) is satisfied;}\\ \text{Crossed-Ambiguity, if only (B) is satisfied;}\\ \text{Mixed-Ambiguity, if both (A) and (B) are satisfied;}\\ \text{No-Ambiguity, otherwise.}\end{cases}$$

The above definition implicitly implies that a character which possesses the same position-tag for all associated dictionary matching words will be assigned "No-Ambiguity".

For example, given a character sequence "大学生物" (university biology) and a set of dictionary-matching-words {"大学" (university), "大学生" (undergraduate)}, for characters '学' and "生", condition (A.1) is satisfied , but condition (B) is not; therefore, DC_2 and DC_3 should be set to "Included-Ambiguity". On the other hand, if the dictionary-matching-words are {"大学生", "生物" (biology)}, then condition (B.1) is satisfied, but condition (A) is not; DC_2 and DC_3 thus should be set to “Crossed-Ambiguity”. However, if we have all the three matching words {“大学”, “大学生”, “生物”}, then both condition (A.1) and condition (B.1) are satisfied; therefore, DC_2 and DC_3 should be set to “Mixed-Ambiguity” in this case. Furthermore, if the matching words are {“大学”, “生物”}, then DC_2 and DC_3 would be “No-Ambiguity”. Last, DC_1 and DC_4 are always “No-Ambiguity” for all above three different cases.

2.3 Tag Matching Status

To indicate the relationship between the tag assigned to c_i and those dictionary matching words which cover c_i, we introduce the Tag Matching Status feature (the abstract aggregate candidate-feature, and is abbreviated as **TM** from now on), which is a member of {Following-Longest-Word, Only-Following-Shorter-Word, Not-Following-Any-Word, Inapplicable} that are defined below. Denote the set of dictionary matching words that begin with c_i as $D_B = \{c_{[i:j]} \mid c_{[i:j]} \in D\}$, the set of dictionary matching words that enclose c_i as $D_M = \{c_{[j:k]} \mid c_{[j:k]} \in D, j < i < k\}$, and the set of dictionary matching words that end with c_i as $D_E = \{c_{[j:i]} \mid c_{[j:i]} \in D\}$. If c_i is tagged as t_i, then TM_i can be decided as follows:

(1) If $D_B \cup D_M \cup D_E = \varnothing$, which indicates that this character is not covered by any dictionary word, then TM_i is set to “Inapplicable”.

(2) If $D_{t_i} = \varnothing$, and $D_B \cup D_M \cup D_E \neq \varnothing$ (where D_{t_i} is the set of dictionary matching words corresponding to t_i; for example, D_{t_i} will be D_B, if $t_i = B$. Please note that $D_S = \varnothing$, since the adopted dictionary only contains multi-character words), which indicates that the assigned tag does not follow any dictionary matching word, then TM_i is set to “Not-Following-Any-Word”;

(3) If $\forall w \in (D_B \cup D_M \cup D_E), \exists w' \in D_{t_i} : \text{len}(w') \geq \text{len}(w)$, then TM_i is set to “Following-Longest-Word”. It indicates that the assigned tag matches the corresponding position-tag of the longest dictionary matching word at that character;

(4) Otherwise, TM_i is set to “Only-Following-Shorter-Word”. It indicates that the assigned tag does not match the corresponding position-tag of the longest dictionary matching word at that character, but matches that of some shorter words.

For example, when we consider the second character ‘学’ in the sequence “大学生” and assume that the dictionary matching words are {“大学”, “大学生”}: if the tag assigned to ‘学’ is “M”, then TM_2 will be “Following-Longest-Word”; if it is “E”, then TM_2 will be “Only-Following-Shorter-Word”; if it is “B” or “S”, TM_2 would be “Not-Following-Any-Word”. Therefore, this candidate-feature is associated with each candidate of the position-tag. However, if no dictionary word covers this character, then TM_2 will be set to “Inapplicable” regardless of which tag is assigned to ‘学’ (i.e., we do not want to disturb the original model in this case).

3 Proposed models

3.1 Enhanced generative model

Wang et al. (2009) proposed a character-based generative model for CWS, which is able to handle the dependency of character-bigrams within words and thus give a good balance for the performance of IV words and OOV words. Their approach adopts the character-tag-pair trigram model, and obtains the desired position-tag sequence $\bar{t}_1^n$ as follows:

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

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

To alleviate the data sparseness problem, we pre-convert the given character string into its corresponding unit string before segmentation, where the unit denotes a FN-String (which is a string mixed with foreign/Arabic/selected-punctuations [5] characters, such as "HTML5", "www.google.com", etc.), a CN-String (which is a string of **C**hinese **N**umbers consisting of Chinese digit-characters such as "六五七七二" (65772), etc.), or a single Chinese character. Therefore, a unit is either a Chinese character or a foreign/numerical expression as defined above, and is represented by either a Chinese-Character-ID or a Meta-Type-ID (i.e., FN-type or CN-type) in the model.

After the character string has been pre-converted into the unit string, we incorporate the dictionary related features proposed in Section 2 into the generative model and re-formulate it as follows:

$$\begin{aligned}\bar{t}_1^n &\equiv \arg\max_{t_1^n} P(t_1^n, TM_1^n \,|\, u_1^n, DW_1^n) \\ &= \arg\max_{t_1^n} \left[\frac{P([u,t,TM]_1^n \,|\, DW_1^n)}{P(u_1^n \,|\, DW_1^n)}\right] \\ &= \arg\max_{t_1^n} P([u,t,TM]_1^n \,|\, DW_1^n)\end{aligned} \tag{2}$$

where DW_i denotes the set of dictionary words that cover u_i (i-th unit). $P([u,t,TM]_1^n \,|\, DW_1^n)$ is then approximated by $\prod_{i=1}^{n} P([u,t,TM]_i \mid [u,t,TM]_{i-2}^{i-1}, DW_i)$ and its associated factor is further derived as follows:

$$\begin{aligned}&P([u,t,TM]_i \mid [u,t,TM]_{i-2}^{i-1}, DW_i) \\ &= P(TM_i \mid [u,t]_{i-2}^{i}, TM_{i-2}^{i-1}, DW_i) \times P([u,t]_i \mid [u,t,TM]_{i-2}^{i-1}, DW_i) \\ &\approx P(TM_i \mid MWL_i, DC_i, u_{i-2}^{i}) \times P([u,t]_i \mid [u,t]_{i-2}^{i-1})\end{aligned} \tag{3}$$

where MWL_i denotes the Maximum Word Length of the words that cover u_i, DC_i (Dictionary Coverage Status) and TM_i (Tag Matching Status) are defined at Section 2.2 and Section 2.3, respectively. The first term $P(TM_i \mid MWL_i, DC_i, u_{i-2}^{i})$ is the tag matching factor, which is mainly introduced to give guidance in the case that the second term $P([u,t]_i \mid [u,t]_{i-2}^{i-1})$ (the original generative model) cannot give reliable prediction when the associated character-tag bigram is unseen in the training corpus.

Equation (3) weighs the tag matching factor and the character-tag trigram factor equally. However, it is reasonable to expect that they should be weighted differently according to their contribution. We thus combine these two factors via log-linear interpolation, which is shown as follows:

[5] Selected punctuations include those members from {'+', '-', '*', '/', '.', '%', '@'}.

$$\begin{aligned}\mathrm{Score}(t_i) = {} & \alpha \times \log P([u,t]_i \mid [u,t]_{i-2}^{i-1}) \\ & + (1-\alpha)\log P(TM_i \mid MWL_i, DC_i, u_{i-2}^{i})\end{aligned} \tag{4}$$

Where α is the weighting coefficient to be decided from the development set, and $0 < \alpha < 1.0$.

3.2 Enhanced integrated model

To incorporate the dictionary information into the discriminative approach, Low et al. (2005) added two additional features (which are features (d) and (e) in the following list) to the widely adopted primitive templates described in (Ng and Low, 2004), and used them to enhance the original discriminative model (Xue, 2003):

(a) C_n $(n = -2, -1, 0, 1, 2)$; (d) MWL_0, t'_0;

(b) $C_n C_{n+1}$ $(n = -2, -1, 0, 1)$; (e) $C_n t'_0$ $(n = -1, 0, 1)$.

(c) $C_{-1} C_1$;

Let W denote the longest dictionary word that covers c_0, then MWL_0 denotes the length of W, and t'_0 denotes the corresponding tag of c_0 in W.

Since the enhanced generative model cannot utilize the features from future context, which is a common drawback of generative approaches (Wang et al., 2010), following the approach of (Wang et al., 2011), we further integrate the enhanced generative model with the above enhanced discriminative model via log-linear interpolation, shown as follows:

$$\begin{aligned}\mathrm{Score}(t_i) = {} & \beta \times [\alpha \times \log P([u,t]_i \mid [u,t]_{i-2}^{i-1}) \\ & + (1-\alpha)\log P(TM_i \mid MWL_i, DC_i, u_{i-2}^{i})] \\ & + (1-\beta) \times \log(P(t_i \mid u_{i-2}^{i+2}, MWL_i, t'_i)\end{aligned} \tag{5}$$

Where α and β are two weighting coefficients to be decided from the development set, and $0 < \alpha, \beta < 1.0$.

4 Experiments

All experiments are conducted on the corpora provided by SIGHAN-2005 (Emerson, 2005) and CIPS-SIGHAN-2010[6] (Zhao and Liu, 2010). Both the in-domain test and cross-domain tests are trained on PKU-News[7] from CIPS-SIGHAN-2010. The PKU-News testing corpus of SIGHAN-2005 is adopted for in-domain test, while the corpora of CIPS-SIGHAN-2010 are used for cross-domain tests. There are four different domains: Literature (denoted as **Lit.**), Computer (**Cmp.**), Medicine (**Med.**) and Finance (**Fin.**). To obtain the weights of different factors in Equations (4) and (5), we randomly selected 1% from the original training corpus as the development set, and

[6] CIPS-SIGHAN-2010 is the first cross-domain Chinese Word Segmentation (CWS) bake-off competition which involves 18 systems. To our knowledge, this data-set is the most well-known and widely adopted (also publically available) one for cross-domain CWS test.

[7] The PKU-News training data of CIPS-SIGHAN-2010 is the same with the PKU training data of SIGHAN-2005.

regard the remaining part as the new training set. The updated corpora statistics are shown at TABLE 1.

Corpora	Domain	Characters	Tokens	Word Types	OOV Rate
Training	News	1,820,456	1,109,947	55,303	N/A
Develop.	News	99,381	60,585	11,216	0.028
Testing	News	172,733	104,372	13,148	0.058
	Lit.	50,637	35,736	6,364	0.069
	Cmp.	53,382	35,319	4,150	0.152
	Med.	50,969	31,490	5,076	0.110
	Fin.	53,253	33,028	4,918	0.087

TABLE 1 – Corpus statistics for CIPS-SIGHAN-2010

Besides, the SRI Language Modelling Toolkit (SRILM)[8] (Stolcke, 2002) is used to train $P([u,t]_i \mid [u,t]_{i-2}^{i-1})$ with the modified Kneser-Ney smoothing method (Chen and Goodman, 1996). Also, the Factored Language Model in SRILM is adopted to train $P_{BF}(TM_i \mid MWL_i, DC_i, u_{i-2}^{i})$, and $P_{BF}(TM_i \mid MWL_i, DC_i, u_{i-2}^{i})$ sequentially back-off to $P_{GT}(TM_i \mid MWL_i, DC_i)$, where the subscripts "BF" and "GT" denote back-off and Good-Turing estimations, respectively. For the discriminative approach, the ME Package[9] provided by Zhang Le is adopted for training. Last, all adjustable weights are first selected only based on the development set. Those obtained optimal values are then fixed and applied to the testing-set.

4.1 Enhanced generative model

4.1.1 Effect of having character context

It is observed that tag matching status is less reliable when there are several dictionary matching words for the given character. Therefore, the character context is introduced to help disambiguate them. To show the influence of character context information, we test two different enhanced generative models. The first one adopts the factor $P(TM_i \mid MWL_i, DC_i)$ (denoted as G1), and the second one adopts the factor $P(TM_i \mid MWL_i, DC_i, u_{i-2}^{i})$ (denoted as G2; with character context). In addition, the performance of the original generative model (denoted as B for baseline) is also shown for comparison. The dictionary adopted here contains all the training words and the testing words excluding the named entities (which are usually not covered by domain dictionaries).

TABLE 2 shows that character context effectively improves the performance for four out of five domains in F-score (except the Computer domain), which mainly resultes from the inconsistent segmentation criteria between the News training set and the computer testing set. For example, 18.3% of "就是" (just like) occurrences are segmented as a single word, and 81.7% of them are segmented into two single-character words "就" (just) and "是" (be) under similar context in the training set. However, this string is always treated as a single word in the Computer testing set. When the context is not considered, G1 prefers the longer word and gives correct answer for all the occurrences in the Computer corpus. While the context is considered, G2 will prefer to follow those occurrences in the training set and thus give wrong result most of the time.

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

[9] http://homepages.inf.ed.ac.uk/lzhang10/maxent.html

Model	News	Lit.	Cmp.	Med.	Fin.	OA
B	0.952	0.928	0.929	0.904	0.950	0.939
G1	0.959	0.951	0.969	0.963	0.971	0.962
G2	0.968	0.953	0.966	0.964	0.973	0.965

TABLE 2 – Effect of adopting character context in F-score (testing-set)

Corpus	Character-Tag-Bigram	B	G1	G2
News	seen	0.018	0.030	0.016
	unseen	0.183	0.045	0.045
Lit.	seen	0.028	0.030	0.022
	unseen	0.203	0.046	0.047
Cmp.	Seen	0.026	0.017	0.015
	unseen	0.183	0.014	0.016
Med.	Seen	0.024	0.016	0.012
	unseen	0.251	0.022	0.023
Fin.	Seen	0.019	0.018	0.011
	unseen	0.155	0.015	0.017

TABLE 3 –Tagging error rates for seen/unseen cases in the testing-set

Overall, G2 outperforms B and G1 by 2.6% and 0.3%, respectively. This phenomenon can be explained by classifying the tagging errors into two groups according to whether the associated character-tag bigram is seen or not in the training set. TABLE 3 shows that the original character-tag trigram factor (i.e., B) works well when the bigram is seen, but it performs poorly when this bigram is unseen. On the other hand, G1 (without context) mainly boosts the performance for unseen cases. However, G2 (with context) also boosts the performance for seen cases. Therefore, it will not let the newly added tag-matching factor contaminate the original trigram model when associated character-tag bigrams are seen. The above observations hold for both development-set and testing-set. G2 is thus adopted for the enhanced generative model and integrated model.

4.1.2 Effect of dictionary coverage rate

Since no dictionary can cover all OOV words for real applications, we would like to know how this enhanced generative model performs under different dictionary coverage rates. We extract two dictionaries: the first one (D1) includes all the training words; and the second one (D2) contains all the OOV words in the testing set (excluding named entities). TABLE 4 gives the results for various combinations of D1 and D2 with $\alpha = 0.5$, where the first row “None” denotes that no dictionary (even D1) is adopted; also, the last column “OA” gives the overall performance of various domains (except News).

It can be seen that the improvements with the dictionary information from the training set are not obvious (None vs. D1). The reason is that the original character-tag trigram model already handles IV words well enough and the information of IV words seems redundant to this model. However, when the dictionary starts to cover OOV words, the performance rises sharply according to the OOV coverage rate. Anyway, the enhanced model always outperforms the original model even when the dictionary only covers a few OOV words.

Dict.	News	Lit.	Cmp.	Med.	Fin.	OA
None	0.952	0.928	0.929	0.904	0.950	0.928
D1	0.953	0.930	0.929	0.907	0.951	0.929
+20%D2	0.955	0.933	0.934	0.919	0.955	0.935
+40%D2	0.958	0.939	0.940	0.931	0.959	0.942
+60%D2	0.961	0.943	0.949	0.941	0.964	0.949
+80%D2	0.964	0.948	0.958	0.952	0.968	0.956
+D2	**0.968**	**0.953**	**0.966**	**0.964**	**0.973**	**0.964**

TABLE 4 – F-score versus different OOV coverage rates for the enhanced character-based generative model (testing-set). OA: overall performance of those four cross-domains. Boldface indicates the best result under each column.

Dict.	News	Lit.	Cmp.	Med.	Fin.	OA
D1	0.935	0.901	0.895	0.861	0.933	0.914
+20%D2	0.940	0.909	0.908	0.880	0.940	0.923
+40%D2	0.945	0.918	0.924	0.901	0.947	0.932
+60%D2	0.950	0.928	0.937	0.923	0.955	0.942
+80%D2	0.955	0.938	0.952	0.945	0.962	0.952
+D2	**0.961**	**0.949**	**0.967**	**0.969**	**0.969**	**0.962**

TABLE 5 – F-score versus different OOV coverage rates for the word-based trigram model

Nonetheless, not every model possesses the robustness for varying dictionary coverage rate. For example, the corresponding result of the word-based generative trigram model[10], given at TABLE 5, shows that it is quite fragile in comparison with our model. In this model, all words kept in the dictionary are used to construct the word lattice in the decoding process. Those OOV words will be treated as unseen events and given a very low score. However, it can be seen that although the results with full dictionary are satisfactory, the performance drops dramatically while the OOV coverage rate decreases. This indicates that this model is quite sensitive to those OOV words, due to its incapability of identifying OOV words beyond the dictionary. This model is thus not useful for real applications, as it is impossible to know the corresponding dictionary coverage rate in the testing set in advance. Therefore, checking the robustness of dictionary-based models for different dictionary coverage rates is important in selecting an appropriate model.

4.1.3 Effect of varying weights

The F-scores of the enhanced generative model versus various α values (the weight of $P([u,t]_i \mid [u,t]_{i-2}^{i-1})$ in Equation (4) are evaluated on the development set, and are shown in FIGURE 3. It can be seen that all the curves are flat near their peaks, which indicates that this enhanced model is not sensitive to which α value is picked. Besides, although the performance decreases when the OOV coverage rate drops, the α locations of peaks for various curves are almost the same (all around $\alpha = 0.4$). This indicates that the best α value is not sensitive to the OOV coverage rate.

[10] This well-known model adopts the form : $WSeq = \arg\max \prod_{i=1}^{m} P(w_i \mid w_{i-2}^{i-1})$ (Wang et al., 2012).

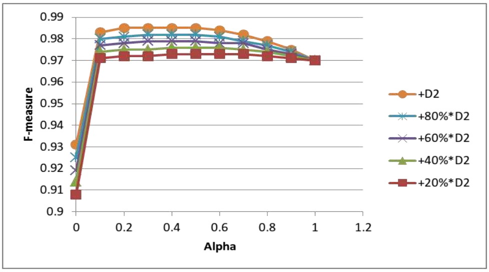


FIGURE 3 – F-score of the enhanced generative model versus various weights α on the development set

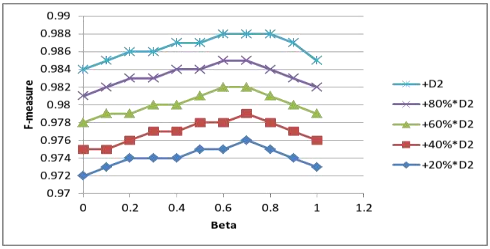


FIGURE 4 –F-score of the enhanced integrated model versus various weights β on the development set

4.2 Enhanced integrated model

Since the original generative model and the discriminative model are found to complement each other (Wang et al., 2011), it is expected that the enhanced generative model and the enhanced discriminative model (Low et al., 2005), which is re-implemented, should also complement each other. This inference is supported by the statistics that these two approaches share only 31.9% of their errors in the literature domain (similarly in other domains). Therefore, integrating these two models is expected to achieve a better result too. For the enhanced integrated model, we fix the weight α of $P([u,t]_i \mid [u,t]_{i-2}^{i-1}$ to be 0.4 (according to section 4.1.3). Afterwards, we adjust the weight of the enhanced generative model and the enhanced discriminative approach on the same development set. FIGURE 4 gives F-scores with different dictionaries versus various β values. The β locations of peaks for various curves are also almost the same (all around β =0.7) for various dictionary coverage rates. This indicates that the β weight is not sensitive to the OOV coverage rate. This figure also shows that the peak of the integrated model is robust for different dictionary coverage rates.

Last, to fairly compare different models in a more realistic condition, Table 6 shows the results of the enhanced discriminative model (denoted by ED) and the integrated model (denoted by EI) with an external dictionary (which roughly corresponds to 65% D2 coverage rate in TABLE 4, and is specified in the next section). Note that our ED result (0.968) is a little bit different from that reported in (Low et al., 2005), which gives 0.965 F-score with a smaller dictionary that is a part of ours (see the next section). It can be seen that our enhanced integrated model achieves the best results on all five corpora.

	News	**Lit.**	**Cmp.**	**Med.**	**Fin.**
SBest	0.969	**0.955**	0.950	0.938	0.960
ED	0.968	0.951	0.946	0.938	0.961
EG	0.967	0.946	0.950	**0.944**	0.962
EI	**0.973**	**0.955**	**0.951**	**0.944**	**0.963**

TABLE 6 - F-scores on the testing sets. SBest: best results from SIGHAN 2005 (News) and CIPS-SIGHAN 2010 (other domains). Boldface indicates the best result.

4.3 Comparison with other state-of-the-art systems

To provide publically accessible dictionaries for open comparison, we combine a general dictionary downloaded from the Internet[11] and another technical dictionary extracted from Wiki[12] as our external dictionary. The first general dictionary is also adopted by (Low et al., 2005). For simplicity, we adopted the same dictionary (the union of above two dictionaries) for all five different domains. This external dictionary includes 458,165 words in total which roughly corresponds to 65% D2 coverage rate in TABLE 4. Since the external dictionary is expected to be collected by the user in real applications, dictionary words should be consistent with his/her own segmentation criterion. Therefore, to give true evaluation for reflecting the real situation, words in the dictionary are first transformed into their corresponding ones according to the same criteria adopted in various given corpora. For example, "免疫系统" (immune system) is converted into "免疫" (immune) and "系统" (system) according to the gold criterion adopted in CIPS-SIGHAN-2010.

The results of the enhanced generative model (denoted by EG) with the external dictionary and the SIGHAN best results in each domain are also given in TABLE 6. They are summarized as follows: Low et al. (2005) added the dictionary information to the discriminative approach and adopted additional corpora. They achieved the best result (0.969 F-score) on PKU News corpus in the open test of SIGHAN-2005. On the other hand, Huang et al. (2010) adopted HMM and some rules to post-process the output of the CRF discriminative approach. They achieved the best in the Literature domain in CIPS-SIGAHN-2010. Last, (Gao and Vogel, 2010) combined several classifiers with a large margin classifier and won the best on other three remaining domains. It shows our enhanced generative model achieves the best results on three out of four cross-domains, and our enhanced integrated model outperforms all the systems reported in the literature.

To further check if the difference between various models listed in Table 6 is really statistically significant, we adopt the bootstrapping technique (Koehn, 2004; Zhang et al., 2004) to conduct the significant tests. We follow the work of (Wang et al., 2010) and take the re-sampling size to be 2,000. 95% confidence interval is adopted in our tests. TABLE 7 shows that our enhanced generative model is superior to the enhanced discriminative model in overall comparison on those four cross-domains. Furthermore, the enhanced integrated model is either superior or comparable to all other models.

[11] http://ccl.pku.edu.cn/alcourse/nlp/2010/word_freq_list.rar
[12] http://dumps.wikimedia.org/zhwiki/20111017/zhwiki-20111017-all-titles-in-ns0.gz

Systems		News	Lit.	Cmp.	Med.	Fin.
A	B					
EG	SBest	<	<	~	>	~
ED	SBest	~	<	~	~	~
EG	ED	<	<	>	>	~
EI	SBest	>	~	~	>	~
	ED	>	>	>	>	~
	EG	>	>	~	~	>

TABLE 7 - Statistical significance test of F-Score among various systems. SBest: best results of the SIGHAN 2005 (News) and the CIPS-SIGHAN 2010 (others). “>” means that A is significantly better than B; “<” means A is worse than B; “~” means that they are not different.

5 Error analysis and discussion

Following the work of (Sun, 2010), we get the upper bound of our enhanced generative model (EG) and enhanced integrated model (EI) by regarding each factor as an independent model, which is shown in TABLE 8. Compared with the results in TABLE 6, we can find that there is still a large room to further improve our proposed models.

Model	News	Lit.	Cmp.	Med.	Fin.
EG	0.984	0.973	0.971	0.968	0.978
EI	0.986	0.976	0.973	0.970	0.980

TABLE 8 - F-score upper bound for EG and EI models

Furthermore, we collect and analyse the remaining errors generated by the enhanced integrated model on the Medicine corpus, which contains a large number of technical terms and is most far away from the upper bound. It is found that 66.6% (out of 1,385) of error words are related to OOV (not seen in the training-set). Among those 922 OOV errors, 758 (82.2%) of them are not covered by the dictionary and 401 (52.9%) out of 758 are technical terms. Therefore, it again confirms how important a dictionary is. Also, 88 (21.9%) of those uncovered terms are with prefix/suffix. For example, “造影术” (radiography) is an OOV word with suffix “术” (technique), while the word “造影” (radiograph) is contained in the dictionary. However, it is wrongly split into “造影” and “术”, since the longest word in the dictionary is preferred. This problem will be our future work.

6 Related work

The word-based generative model (Gao et al., 2003; Zhang et al., 2003) is a classical approach for CWS. However, this approach needs an additional module to recognize OOV words. Therefore, the character-based discriminative model (Xue, 2003; Low et al., 2005; Zhang et al., 2006; Jiang et al., 2008; Zhao et al., 2010) has become the main stream due to its capability in handling OOV words.

However, the character-based discriminative model cannot give satisfactory performance for IV words. Wang et al. (2010) thus proposed a generative model to fix this problem. Afterwards, they further proposed an integrated model to integrate generative and discriminative approaches, as these two approaches complement each other.

On the other hand, dictionary information has been utilized in the discriminative approach in the previous works of (Low et al., 2005; Zhao et al., 2010). However, they focus on improving the in-domain word segmentation accuracy, while we investigate how the domain invariant feature (based on dictionary information) helps for cross-domain tasks. Besides, the effect of varying OOV words coverage rates is studied in this paper for the first time.

In addition to dictionary feature, Zhao and Kit (2007; 2008), Sun and Xu (2011) too, also adopted the accessor variety feature to gain better generalization ability. Since this feature can be extracted from unlabelled corpora, it is suitable to be adopted for domain adaptation. Again, all their works focus on in-domain performance. Other works that focus on in-domain performance also include (Zhang and Clark, 2007), (Fu et al., 2008), (Jiang et al., 2008), (Lin, 2009), (Xiong et al., 2009), and (Zhang and Clark, 2011).

Last, (Ben-David et al., 2007) pointed out that a good feature representation for domain adaptation should minimize the difference between its distributions in source and target domains. The proposed abstract feature is also inspired by their conclusion.

Our approach differs from those previous works in several ways. First, we do not simply add the dictionary matching information as an additional feature under the Maximum Entropy framework. In contrast, we derive a new generative model with dictionary information starting from the problem formulation, and solve the problem in a principled way. Second, the robustness of the proposed model for varying dictionary coverage rate is first studied and checked in this paper. As explained in Section 4.1.2, this issue is important for selecting a model for real applications.

7 Conclusion

Current character-based approaches are not robust for cross domain Chinese word segmentation, because those surface features adopted in the model frequently possess different tag distributions for the same character in various domains. This paper thus proposes a new abstract aggregate candidate-feature, which indicates if the assigned tag follows the corresponding position-tag of the longest dictionary matching word. With this novel domain invariant feature, we then derive an enhanced generative model for cross-domain CWS to solve the problem in a principled way. Experiments show that the proposed approach is robust for various OOV coverage rates and outperforms the best system in three out of five corpora.

The proposed model is further integrated with an enhanced discriminative approach because they complement each other. With the help of a publically accessible external dictionary, experiments on the SIGHAN-2005 and CIPS-SIGHAN-2010 show that our integrated approach outperforms all the systems in open test and achieves the best F-score in each corpus across five different specified domains.

Acknowledgments

The research work has been funded by the Natural Science Foundation of China under Grant No. 6097 5053 and the Hi-Tech Research and Development Program ("863" Program) of China under Grant No. 2012AA011102 and 2011AA01A207.

References

Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F. (2007). Analysis of representations for domain adaptation. Advances in neural information processing systems, volume 19, pages 137.

Chen, S. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In Proceedings of the 34th annual meeting on Association for Computational Linguistics, pages 310-318.

Emerson, T. (2005). The second international Chinese word segmentation bakeoff. In Proceedings of the fourth SIGHAN workshop, pages 123-133.

Fu, G. Kit, C. and Webster, J. (2008). Chinese word segmentation as morpheme-based lexical chunking. Information Sciences, volume 178, pages 2282-2296.

Gao, J., Li, M. and Huang, C. (2003). Improved Source-Channel Models for Chinese Word Segmentation. In Proceedings of the 41th Annual Meeting of Association of Computational Linguistics (ACL), pages 272-279.

Gao, Q. and Vogel, S. (2010). A Multi-layer Chinese Word Segmentation System Optimized for Out-of-domain Tasks. In Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP2010), pages 210-215.

Huang, D., Tong, D. and Luo, Y. (2010). HMM Revises Low Marginal Probability by CRF for Chinese Word Segmentation. In Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP2010), pages 216-220.

Jiang, H. and Dong, Z. (2010). An Double Hidden HMM and an CRF for Segmentation Tasks with Pinyin’s Finals. In Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP2010), pages 277-281.

Jiang, W., Huang, L., Liu, Q. and Lu, Y. (2008). A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In Proceedings of ACL, pages 897-904, Columbus, Ohio, USA.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Proceedings of EMNLP, pages 388-395, Barcelona, Spain.

Lin, D. (2009). Combining Language Modeling and Discriminative Classification for Word Segmentation. In Proceedings of CICLing 2009, pages 170-182.

Low, J., Ng, H. and Guo W. (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.

Ng, H. and Low, J. (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, Barcelona, Spain.

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

Sun, W. (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.

Sun, W. and Xu, J. (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.

Sun, X., Wang, H. and Li, W. (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.

Wang, K., Zong, C. and Su, K. (2009). Which is More Suitable for Chinese Word Segmentation, the Generative Model or the Discriminative One? In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC23), pages 827-834, Hong Kong, China.

Wang, K., Zong, C. and Su, K. (2010). A Character-Based Joint Model for Chinese Word Segmentation. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1173-1181, Beijing, China.

Wang, K., Zong, C. and Su, K. (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

Xiong, Y., Zhu, J., Huang, H. and Xu, H. (2009). Minimum tag error for discriminative training of conditional random fields. Information Sciences, volume 179, pages 169-179.

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

Zhang, H., Yu, H., Xiong, D. and Liu, Q. (2003). HHMM-based Chinese lexical analyzer ICTCLAS. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pages 184-187.

Zhang, R., Kikui, G. and Sumita, E. (2006). Subword-based Tagging for Confidence-dependent Chinese Word Segmentation. In Proceedings of the COLING/ACL, pages 961-968, Sydney, Australia.

Zhang, Y. and Clark, S. (2007). Chinese Segmentation with a Word-Based Perceptron Algorithm. In Proceedings of ACL, pages 840-847, Prague, Czech Republic.

Zhang, Y. and Clark, S. (2011). Syntactic processing using the generalized perceptron and beam search. Computational Linguistics, volume 37, pages 105-151.

Zhang, Y., Vogel, S. and Waibel, A. (2004). Interpreting BLEU/NIST scores: How much improvement do we need to have a better system. In Proceedings of the Fourth International Conference on Language Resource and Evaluation (LREC), pages 2051–2054.

Zhao, H. and Kit, C. (2007). Incorporating Global Information into Supervised Learning for Chinese Word Segmentation. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 66-74.

Zhao, H. and Kit, C. (2008). Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In Sixth SIGHAN Workshop on Chinese Language Processing.

Zhao, H. and Liu, Q. (2010). The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff. In Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP2010), pages 199-209, Beijing, China.

Zhao, H., Huang, C., Li, M. and Lu, B. (2010). A Unified Character-Based Tagging Framework for Chinese Word Segmentation. ACM Transactions on Asian Language Information Processing (TALIP), 9 (2). pages 1-32.