Semi-supervised Chinese Word Segmentation based on Bilingual Information

Wei Chen
Institute of Automation,
Chinese Academy of Sciences,
Beijing, 100190, China
wei.chen.media@ia.ac.cn

Bo Xu
Institute of Automation,
Chinese Academy of Sciences,
Beijing, 100190, China
xubo@ia.ac.cn

Abstract

This paper presents a bilingual semi-supervised Chinese word segmentation (CWS) method that leverages the natural segmenting information of English sentences. The proposed method involves learning three levels of features, namely, character-level, phrase-level and sentence-level, provided by multiple sub-models. We use a sub-model of conditional random fields (CRF) to learn monolingual grammars, a sub-model based on character-based alignment to obtain explicit segmenting knowledge, and another sub-model based on transliteration similarity to detect out-of-vocabulary (OOV) words. Moreover, we propose a sub-model leveraging neural network to ensure the proper treatment of the semantic gap and a phrase-based translation sub-model to score the translation probability of the Chinese segmentation and its corresponding English sentences. A cascaded log-linear model is employed to combine these features to segment bilingual unlabeled data, the results of which are used to justify the original supervised CWS model. The evaluation shows that our method results in superior results compared with those of the state-of-the-art monolingual and bilingual semi-supervised models that have been reported in the literature.

1 Introduction

Chinese word segmentation (CWS) is generally accepted to be a necessary first step in most Chinese NLP tasks because Chinese sentences are written in continuous sequences of characters with no explicit delimiters (e.g., the spaces in English). Many studies have been conducted in this area, resulting in extensive investigation of the problem of CWS using machine learning techniques in recent years. However, the reliability of CWS that can be achieved using machine learning techniques relies heavily on the availability of a large amount of high-quality, manually segmented data. Because hand-labeling individual words and word boundaries is very difficult (Jiao et al., 2006), producing segmented Chinese texts is very time-consuming and expensive. Although a number of manually segmented datasets have been constructed by various organizations, it is not feasible to combine them into a single complete dataset because of their incompatibility due to the use of various segmenting standards. Thus, it is difficult to build a large-scale manually segmented corpus, and the resulting lack of such a corpus is detrimental to further enhancement of the accuracy of CWS.

To address the scarcity of manually segmented corpora, a number of semi-supervised CWS approaches have been intensively investigated in recent years. These approaches attempt to either learn the predicted label distribution (Jiao et al., 2006) or extract mutual information ((Liang et al., 2005); (Sun and Xu, 2011); (Zeng et al., 2013a)) from large-scale monolingual unlabeled data to update the baseline model (from manually segmented corpora). In addition to these techniques, several co-training approaches (Zeng et al., 2013b) using character-based and word-based models have also been employed. However, because monolingual unlabeled data contain limited natural segmenting information, in most semi-supervised methods, the objective function tends to be optimized based on the personal experience and knowledge of the researchers. This practice means that these methods can typically yield high performance in certain specialized domains, but they lack generalizability. In contrast with these methods, we propose to leverage bilingual unlabeled data, i.e., a Chinese-English corpus with sentence alignment. Because English sentences

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1207–1216,
Lisbon, Portugal, 17-21 September 2015.

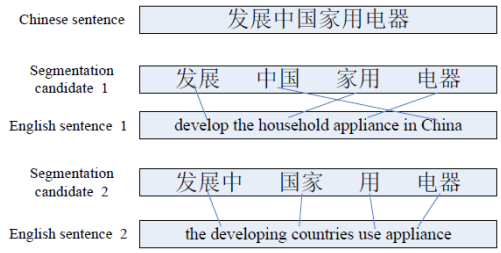

Figure 1: The examples of different segmentation on the same Chinese sentences guided by the English sentences

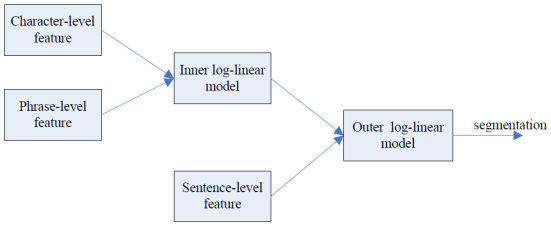

Figure 2: The structure of cascaded log-linear model with multilevel features

are naturally segmented, extracting information from a bilingual corpus is a much more objective task. As the example presented in Fig 1 shows, the English sentences that correspond to Chinese text can easily help guide better segmentation, and thus, the learning of segmenting information from bilingual data is a very promising approach.

In this paper, to obtain high-quality segmenting information from bilingual unlabeled data, we leverage multilevel features using the following steps: first, we integrate character-level features calculated using a conditional random field (CRF) model, which is used to capture the monolingual grammars. Then, we employ a statistical aligner to perform character-based alignment. Given the results of this character-based alignment, we apply several phrase-level features to extract explicit and implicit segmenting information: (1) we use two types of English-Chinese co-occurrence features (one-to-many and many-to-many) to learn the explicit segmenting information of the English sentences, (2) we use the transliteration similarity feature to detect out-of-vocabulary (OOV) words using a phrase-based translation model, and (3) we employ a neural network to calculate the semantic gap between the Chinese and English words to ensure that the Chinese segmentation follows the semantic meanings of the corresponding English sentences as closely as possible. Finally, we employ another phrase-based translation model to perform a sentence-level calculation of the translation probability of the Chinese segmentation and its corresponding English sentences. After obtaining these multilevel features, we normalize them and combine them into two log-linear models in a cascaded structure, which is illustrated in Fig 2. Finally, we segment the bilingual unlabeled data using the proposed model and use the segmentation of those data to justify the original supervised CWS model, which was trained on a standard manually segmented corpus.

In fact, several semi-supervised CWS methods have previously been proposed that leverage bilingual unlabeled data ((Xu et al., 2008); (Chang et al., 2008); (Ma and Way, 2009);(Chung et al., 2009); (Xi et al., 2012)). However, most were developed for statistical machine translation (SMT), causing them to focus on decreasing the perplexity of the bilingual data and the word alignment process rather than on achieving more accurate segmentation. These methods achieve significant improvement in SMT performance but are not very suitable for common NLP tasks because in many situations, they ignore the standard grammars to satisfy the needs of SMT. By contrast, we employ various types of features to capture both monolingual standard grammars and bilingual segmenting information, which allows our semi-supervised CWS model to be very efficient at other NLP tasks and endows it with higher generalizability.

Our evaluation also shows that our method significantly outperforms the state-of-the-art monolingual and bilingual semi-supervised approaches.

2 Related Work

First, we review related work on monolingual supervised and semi-supervised CWS methods. Then, we review bilingual semi-supervised CWS.

2.1 Monolingual Supervised and Semi-supervised CWS Methods

Considerable efforts have been made in the NLP community in the study of Chinese word segmentation. The most popular supervised approach treats word segmentation as a sequence labeling problem, as first proposed by (Xue et al., 2003). Most previous systems have addressed this task using linear statistical models with carefully designed features ((Peng et al., 2004); (Asahara et

al., 2005); (Zhang and Clark, 2007); (Zhao et al., 2010)). However, the primary shortcoming of these approaches is that they rely heavily on a large amount of labeled data, which is very time-consuming and expensive to produce. Thus, the scale of available manually labeled data has placed considerable limitations on the further enhancement of supervised CWS methods.

To address this problem, a number of semi-supervised CWS approaches have been intensively investigated in recent years. For example, (Sun and Xu, 2011) enhanced their segmentation results by interpolating statistics-based features derived from unlabeled data into a CRF model. (Zeng et al., 2013a) introduced a graph-based semi-supervised joint model of Chinese word segmentation and part-of-speech tagging and regularized the learning of a linear CRF model based on the label distributions derived from unlabeled data. However, because monolingual unlabeled data lack natural segmenting information, most previous semi-supervised CWS methods have required certain assumptions to be made regarding their objective functions based on the researchers' personal experiences. By contrast, we leverage bilingual unlabeled data that contain the natural segmentation that is present in English sentences and can therefore extract linguistic knowledge without any manual assumptions or bias.

2.2 Bilingual Semi-supervised CWS Methods

Some previous work ((Xu et al., 2008); (Chang et al., 2008); (Ma and Way, 2009);(Chung et al., 2009); (Xi et al., 2012)) has been performed on leveraging bilingual unlabeled data to achieve better segmentation, although most such studies have focused on statistical machine translation (SMT). These approaches leverage the mappings of individual English words to one or more consecutive Chinese characters either to construct a Chinese word dictionary for maximum-matching segmentation (Xu et al., 2004) or to form a labeled dataset for training a sequence labeling model (Peng et al., 2004). (Zeng et al., 2014) also used such mappings to bias a supervised segmentation model toward a better solution for SMT. However, because most of these approaches focus on SMT performance, they emphasize decreasing the perplexity of the bilingual data and word alignment rather than improving the CWS accuracy. Thus, they sometimes ignore the standard grammars during segmentation in favor of satisfying the needs of SMT, thereby causing these methods to be rather unsuitable for other NLP tasks. By contrast, we propose to use various types of features to capture syntactic and semantic information and a cascaded log-linear model to maintain balance between the monolingual grammars and the bilingual knowledge.

3 Multilevel Features

In this section, we describe the three levels of features used in our approach. We propose to use character-level features to capture monolingual grammars and phrase-level and sentence-level features to obtain bilingual segmenting information. Moreover, we describe a cascaded log-linear model by proposing both inner and outer log-linear models.

3.1 Character-level Feature

The conditional random field (CRF) (Lafferty et al., 2001) model was first used for CWS tasks by (Xue et al., 2003) who treated the CWS task as a sequence tagging problem and demonstrated this model's effectiveness in detecting OOV words.

In this paper, we score the character-level feature in the same manner defined by (Xue et al., 2003). For the jth character c_j in the sentence $c_1^J = c_1...c_J$, the score can be calculated as follows:

$$f_{CRF}(j) = \sum_k \lambda_k f_k(y_{j-1}, y_j, c_1^J, j) \quad (1)$$

where $f_k(y_{j-1}, y_j, c_1^J, j)$ is a feature function and λ_k is a learned weight that corresponds to the feature f_k. j represents the index of the character in the sentence. y_{j-1} and y_j represent the tags of the previous and current characters, respectively.

We do not introduce the CRF-based CWS model in detail here, but more information can be obtained from (Lafferty et al., 2001) and (Xue et al., 2003).

3.2 Phrase-level Features

In this section, we first describe English-Chinese character-based alignment. Then, we propose several phrase-level features to obtain explicit and implicit segmenting information from the character-based alignment. Finally, we describe the inner log-linear model that is used to combine the character-level and phrase-level features.

3.2.1 English-Chinese Character-based Alignment

To avoid introducing omissions and mistakes into the linguistic information in the initial segmentations of the bilingual data, we perform a statistical character-based alignment: First, every Chinese character in the bitexts is separated by white spaces so that individual characters are recognized as unique "words" or alignment targets. Then, they are associated with English words using a statistical word aligner.

By representing the English and Chinese sentences as $e_1^I = e_1 e_2 ... e_I$ and $c_1^J = c_1 c_2 ... c_J$, respectively, where e_i and c_j represent single elements of the sentences, we define their alignment as a_1^K, of which each element is a span $a_k =< s, t >$ and represents the alignment of the English word e_s with the Chinese character c_t. Then, the corpus of unlabeled bilingual data can be represented as the set of sentence tuples $<e_1^I, c_1^J, a_1^K>$

To obtain the character-based alignment, we employ an open-source toolkit Pialign[1] ((Neubig et al., 2011); (Neubig et al., 2012)) which uses Bayesian learning and inversion transduction grammars.

3.2.2 Features Obtained from the Character-based Alignment

Given the English-Chinese character-based alignment a_1^K, we extract several phrase-level features to optimize the segmentation. For the jth character in c_1^J, we assume that one of the segmentations of the substring c_1^j can be represented as $w_1^{N+1} = w_1 w_2 w_3 ... w_{N+1} = c_1^{j_1} c_{j_1+1}^{j_2} ... c_{j_N+1}^{j}$. Then, we calculate the scores of each Chinese word $w_n = c_{j_m}^{j_n}$ $(j_m = j_{n-1} + 1)$ in w_1^{N+1} using the following features.

English-Chinese One-to-Many Alignment

To evaluate the probability that a sequence of Chinese characters $c_{j_m}^{j_n} = c_{j_m} c_{j_m+1} ... c_{j_n}$ should be combined into a word w_n based on the corresponding English sentence, we integrate the feature of English-Chinese one-to-many alignment (one English word is aligned with multiple Chinese characters). First, for any English word e_i in e_1^I, the phrase tuple $< e_i, c_{j_m}^{j_n} >$ can be defined as an aligned One-to-Many phrase tuple if it satisfies the following conditions:

(1) $< i, j_m > \in a_1^K, < i, j_n > \in a_1^K$

(2) $\forall j' \notin [j_m, j_n], < i, j' > \notin a_1^K$

(3) $\forall i' \neq i,\ \forall j' \in [j_m, j_n], < i', j' > \notin a_1^K$

Then, for any phrase tuple $< \overline{e_i}, c_{j_m}^{j_n} >$ that satisfies these conditions, the span $< \overline{i}, j_m, j_n >$ is defined as a One-to-Many span and as a member of the set A_{One}.

Thus, for each span $< i, j_m, j_n >$, the One-to-Many score can be calculated as follows:

$$s(< i, j_m, j_n >) = \begin{cases} t(c_{j_m}^{j_n} | e_i) & if < i, j_m, j_n > \in A_{One} \\ 0 & else \end{cases} \quad (2)$$

where $t(c_{j_m}^{j_n} | e_i)$ represents the translation probability of the phrase tuple $c_{j_m}^{j_n} | e_i$.

Finally, the score for the feature of English-Chinese One-to-Many alignment for $w_n = c_{j_m}^{j_n}$ is derived as follows:

$$f_{One-to-Many}(n) = \underset{i \in [1,I]}{\operatorname{argmax}}\, s(< i, j_m, j_n >) \quad (3)$$

English-Chinese Many-to-Many Alignment

The second phrase-level feature, called English-Chinese Many-to-Many Alignment (multiple English words are aligned with multiple Chinese characters), is used to evaluate the probability that a space should be inserted between c_n and c_{n+1}. Similar to One-to-Many alignment, for any sequence of English words $e_{i_1}^{i_2}$ and the Chinese word $w_n = c_{j_m}^{j_n}$, the phrase tuple $< e_{i_1}^{i_2}, c_{j_1}^{j_n} >$ is defined as an aligned Many-to-Many phrase tuple if it satisfies the following conditions:

(1) $j_1 \leq j_m$, and j_1 is the beginning character of a word in w_1^n

(1) $< i_1, j_1 > \in a_1^K, < i_2, j_m > \in a_1^K$

(2) $\forall j' \notin [j_1, j_m], \forall i' \in [i_1, i_2], < i', j' > \notin a_1^K$

(3) $\forall j' \in [j_1, j_m], \forall i' \notin [i_1, i_2], < i', j' > \notin a_1^K$

Then, for any phrase tuple $< \overline{e_{i_1}^{i_2}}, c_{j_m}^{j_n} >$ that satisfies these conditions, the span $< \overline{i_1, i_2}, j_1, j_n >$ is defined as a Many-to-Many span and as a member of the set A_{Many}.

Thus, for each span $< i_1, i_2, j_1, j_n >$, the Many-to-Many score can be calculated as follows:

$$s(< i_1, i_2, j_1, j_n >) = \begin{cases} t(c_{j_1}^{j_n} | e_{i_1}^{i_2}) \\ \quad if < i_1, i_2, j_1, j_n > \in A_{Many} \\ 0 \quad else \end{cases} \quad (4)$$

where $t(c_{j_1}^{j_n} | e_{i_1}^{i_2})$ represents the translation probability of the phrase tuple $< e_{i_1}^{i_2}, c_{j_1}^{j_n} >$.

[1] http://www.phontron.com/pialign/

Finally, the score for the feature of English-Chinese Many-to-Many alignment for $w_n = c_{j_m}^{j_n}$ is derived as follows:

$$f_{Many-to-Many}(n) = \underset{i_1 \in [1,I] i_2 \in [i_1,I] j_1 \leq j_m}{\operatorname{argmax}} s(i_1, i_2, j_1, j_n) \tag{5}$$

Transliteration Feature

To account for named entities (NEs), which suffer from sparsity and thus make it difficult to calculate the probabilities discussed above, we introduce a transliteration feature to evaluate the similarities between the pronunciations of Chinese and English words because many NEs are translated via transliteration. To perform this task, we first introduce an initial NE dictionary and convert each dictionary item—for example, we convert "爱丽斯/Alice" into "ai l i s i/a l i c e" —by transforming the Chinese word into its pronunciation (represented by the function $F_{py}(\cdot)$) and splitting the English word into its constituent letters (represented by the function $F_{let}(\cdot)$). Then, we train two phrase-based translation models (Chinese-English and English-Chinese) on the data obtained from the converted NE dictionary.

Specifically, we apply two standard log-linear phrase-based SMT models. The GIZA++ aligner is adopted to obtain word alignments (Och and Ney, 2000) from the converted NE dictionary. The heuristic strategy of grow-diag-final-and (Koehn et al., 2003) is used to combine the bidirectional alignments to extract phrase translations and to reorder tables. A 5-gram language model with Kneser-Ney smoothing is trained using SRILM (Stolcke et al., 2002) on the target language. Moses (Koehn et al., 2007) is used as a decoder. Minimum error rate training (MERT) (Och et al., 2003) is applied to tune the feature parameters on the development dataset.

Given these two phrase-based translation models, we calculate each span $< i, j_m, j_n >$ in A_{One} for the Chinese word w_n using the following formula:

$$\begin{aligned} S_{tr}(< i, j_m, j_n >) = S_{ch-en}(< i, j_m, j_n >) \\ + S_{en-ch}(< i, j_m, j_n >) \end{aligned} \tag{6}$$

where $S_{ch-en}(<i, j_m, j_n>) = D_{Lev}(F_{let}e_i, PT_{ch-en}(F_{py}(c_{j_m}^{j_n})))$ means that the pronunciation conversion in the Chinese-English direction is performed as follows: First, the English word e_i is split into its constituent letters; Second, the sequence of Chinese characters $c_{j_m}^{j_n}$ is converted into its pronunciation; Third, this pronunciation is input into the Chinese-English phrase-based translation model, and the corresponding translation result is obtained; And finally, the Levenshtein distance between the English letters and the translation result is returned.

$S_{en-ch}(<i, j_m, j_n>)$ can be calculated in exactly the same way.

We set any span that does not belong to A_{One} to zero, and the transliteration feature score of a word $w_n = c_{j_m}^{j_n}$ is derived as follows:

$$f_{transliteration}(n) = \underset{i \in [1,I]}{\operatorname{argmax}} S_{tr}(< i, j_m, j_n >) \tag{7}$$

English-Chinese semantic gap feature

To guarantee that the semantic meanings of the Chinese segmentation match those of the corresponding English sentences as closely as possible, we propose to use a feature based on the English-Chinese semantic gap to ensure the retention of semantic meaning during the segmentation process.

First, we pre-train word embeddings using the open-source toolkit Word2Vec (Mikolov et al., 2013) on the Chinese (segmented using character-level features only) and English sentences separately, thereby obtaining the vocabularies V_{ch} and V_{en} and their corresponding embedding matrixes $L_{ch} \in R^{n \times |V_{ch}|}$ and $L_{en} \in R^{n \times |V_{en}|}$. Given a Chinese word w_n with an index i in the vocabulary, it is then straightforward to retrieve the word's vector representation via simple multiplication with a binary vector d that is equal to zero at all positions except that with index i:

$$X_i = L_{ch} d_i \in R^n \tag{8}$$

Because the word embeddings for the two languages (L_{ch} and L_{en}) are learned separately and located in different vector spaces, we suppose that a transformation exists between these two semantic embedding spaces. Thus, we collect all the One-to-Many phrase tuples $< e_1, c_{j_1}^{j_2} >$ that satisfy $e_1 \in V_{en}$ and $c_{j_1}^{j_2} \in V_{ch}$ from the entire corpus of bilingual data. Then, we insert the word embedding tuple of each One-to-Many phrase tuple into the set A_{embed}. Let us consider a word embedding tuple $< p_s, p_t >$ in A_{embed} as an example. We define a bidirectional semantic distance using the parameter θ as follows:

$$E_{sem}(p_s, p_t; \theta) = E_{sem}(p_s|p_t, \theta) + E_{sem}(p_t|p_s, \theta) \quad (9)$$

Here, $E_{sem}(p_s|p_t, \theta) = E_{sem}(p_t, f(W_{en}^{ch} p_s + b_{en}^{ch}))$ represents the transformation of p_s and is performed as follows: We first multiply a parameter matrix W_{en}^{ch} by p_s, and after adding a bias term b_{en}^{ch}, we apply an element-wise activation function $f = tanh(\cdot)$. Finally, we calculate their Euclidean distance:

$$E_{sem}(p_s|p_t, \theta) = \frac{1}{2}||p_t - f(W_{en}^{ch} p_s + b_{en}^{ch})||^2 \quad (10)$$

$E_{sem}(p_t|p_s, \theta)$ can be calculated in exactly the same way.

Given the definition of the semantic distance of each word-embedding tuple in A_{embed}, we wish to minimize the following objective function:

$$J = \sum_{<p_s, p_t> \in A_{embed}} E_{sem}(p_s, p_t; \theta) \quad (11)$$

We apply the Stochastic Gradient Descent (SGD) algorithm to optimize each parameter and ultimately obtain the optimized parameters θ^*.

Using θ^*, we can calculate the semantic gap for any possible span for w_n, such as $< i, j_m, j_n >$, as follows:

$$S_{gap}(< i, j_m, j_n >) = \begin{cases} \frac{1}{E_{sem}(p'_s|p'_t, \theta^*)} \\ \quad if \quad e_i \in V_{en} \quad c_{j_m}^{j_n} \in V_{ch} \\ \quad\quad\quad < i, j_m, j_n > \in A_{One} \\ 0 \quad else \end{cases} \quad (12)$$

where p'_s and p'_t are the word vector representation of e_i and $c_{j_m}^{j_n}$, respectively. Thus, the semantic gap feature score of the word $w_n = c_{j_m}^{j_n}$ is derived as follows:

$$f_{sem}(w_n) = \underset{i \in [1,I]}{\operatorname{argmax}} S_{gap}(< i, j_m, j_n >) \quad (13)$$

3.2.3 Normalization and the Inner Log-Linear Model

Because the output scores of each sub-model described above are not probabilistic and they vary by orders of magnitude, we must first normalize the output scores of each sub-model. After normalization, the scores have means and standard deviations of zero. We represent the normalization function by $Norm(\cdot)$.

Thus, for the substring c_1^j ($j \in [1, J)$) in c_1^J of the sentence tuple $<e_1^I, c_1^J, a_1^K>$, assuming that one of its candidate segmentations is $w_1^{N+1} = w_1 w_2 w_3 ... w_{N+1} = c_1^{j_1} c_{j_1+1}^{j_2} ... c_{j_N+1}^{j}$, the feature score of the inner log-linear model is derived as follows:

$$f_{inner} = \sum_{j' \in [1,j]} Norm(f_{CRF}(j')) + \lambda_1 \sum_{n \in [1, N+1]} (\sum_k Norm(f_k(n))) \quad (14)$$

where $f_k(n)$ represents the phrase-level features.

Then, we tune the weight λ_1 from 0 to 1 in equal increments of 0.1 to optimize its value.

3.3 Sentence-level Features

In this section, we describe the sentence-level features calculated using the phrase-based translation model and the outer log-linear model that is used to combine the sentence-level features with the features in the inner log-linear model.

3.3.1 Features Obtained from the Phrase-based Translation Model

Let us consider the last character c_J in c_1^J and assume that its candidate segmentation (according to the inner log-linear model only) is $w_1^{N+1} = w_1 w_2 w_3 ... w_{N+1}$. We now add a sentence-level feature to incorporate into the inner log-linear model. This sentence-level feature is obtained using a phrase-based translation model. We segment the Chinese sentences from the bilingual unlabeled data using character-level features only and train a phrase-based translation model on the bilingual data that is similar to the phrase-based translation model used for the transliteration features.

Unlike the usage of the phrase-based translation model in the case of the transliteration features, here, we input both the source and target sentences and achieve the output of translation probability. Thus, we perform a force decoding for the sentence tuple $<w_1^{N+1}, e_1^I>$ and obtain the set of decoding paths P(w_1^{N+1}), where each element acts as a decoding path that can translate w_1^{N+1} into e_1^I. Finally, we define the sentence-level feature score of $<w_1^{N+1}, e_1^I>$ as follows:

$$f_{sent}(w_1^{N+1}) = \underset{p(w_1^{N+1}) \in P(w_1^{N+1})}{\operatorname{argmax}} F_{trans}(p(w_1^{N+1})) \quad (15)$$

where $F_{trans}(\cdot)$ returns the translation score of the given decoding path based on the phrase-based translation model.

3.3.2 The Outer Log-Linear Model

Finally, we normalize the sentence-level features in a manner similar to that described previously and construct the outer log-linear model by combining the inner log-linear model and the sentence-level features as follows:

$$f_{outer} = f_{inner} + \lambda_2 Norm(f_{sent}(w_1^{N+1})) \quad (16)$$

Then, we also tune the weight λ_2 from 0 to 1 in equal increments of 0.1 to optimize its value.

3.3.3 Decoder

A traditional viterbi beam search procedure is applied in the decoder to seek the segmented sequence with the highest score. Given a sentence tuple $< e_1^I, c_1^J, a_1^K >$, the decoding procedure will proceed in a left-right fashion using a dynamic programming approach. At each position j in the sequence c_1^J, we maintain a vector of size N to store the top N candidate segmentations of subsequence c_1^j which are scored using the inner log-linear model ($j \in [1, J)$) or the outer log-linear model ($j = J$). Finally, we return the best segmentation.

4 Justifying the Original CWS Model

We justify the original CWS model (the CRF-based model trained on manually segmented data) using the new CRF model trained on the segmentation of unlabeled bilingual data. To avoid over-weakening the influence of the small-scale manually segmented data, we again utilized a log-linear model to balance their weights. The formula can be described as follows:

$$\begin{aligned} f_{new\_mono} = & \sum_{k_1} \lambda_{k_1} f_{k_1}(y_{j-1}, y_j, c_1^J, j) \\ & + \theta_3 \sum_{k_2} \lambda_{k_2} f_{k_2}(y_{j-1}, y_j, c_1^J, j) \end{aligned} \quad (17)$$

where θ_3 represents the weights of the second CRF model, which are set via minimum error rate training using the developing dataset, and λ_{k_i} (i =1, 2) represents the learned weights of the features of the CRF models.

5 The Datasets

In this paper, we conduct our experiments on the corpus of People's daily of 1998 (from January to June) as the standard (manually segmented) training corpus, the corpus of Bakeoff-2 CWS evaluation as the developing and testing dataset. As the corpus of Bakeoff-2 is made up of several sets provided by different organizations, we only select two sets whose segmenting standards are similar to the training corpus. For each set, we take 3000 sentences as the developing dataset and the others as the testing dataset. The statistics of every set and the standard training corpus are shown in Table 1.

Data Set	of sent.	of words
Training	120K	7.28M
AS	708K	5.45M
PKU	19K	1.1M

Table 1: Statistics of training and testing datasets

Moreover, the bilingual unlabeled data is formed by a large in-house Chinese-English parallel corpus (Tian et al., 2014). There are in total 2,215,000 Chinese-English sentence pairs crawled from online resources, concentrated in 5 different domains including laws, novels, spoken, news and miscellaneous.

6 Experiments

In our evaluation, the F-score was used as the accuracy measure. The precision p is defined as the percentage of words in the decoder output that are segmented correctly, and the recall r is the percentage of gold-standard output words that are correctly segmented by the decoder. The balanced F- score is calculated as 2pr/(p + r). We also report the recall of OOV words in our experiments. In the following, we refer to our methods as "SLBD" (segmenter leveraging bilingual data).

Initially, we evaluated state-of-the-art supervised CWS methods, i.e., those of (Peng et al., 2004) (Peng); (Asahara et al., 2005) (Asahara); (Zhang and Clark, 2007) ($Z\&C$); (Zhao et al., 2010) (Zhao), whose models are trained only on manually segmented data. Moreover, we also evaluated the performance of our sub-models by

methods	AS		PKU	
	F	OOV	F	OOV
Peng	91.6	52.5	91.1	59
Asahara	92.2	63.1	91.4	61.6
$Z\&C$	92.9	69.9	91.6	67.9
Zhao	93.1	72	92.3	60.6
character-level	92.3	58.6	92.9	60.8
Inner log-linear	95.9	78.8	96.1	81
Outer log-linear	96.7	80.8	97.1	85

Table 2: Word segmentation performance of SLBD and supervised CWS methods[%]

segmenting the bilingual unlabeled dataset using character-level features only, the inner log-linear model (which includes character-level and phrase-level features) and the outer log-linear model (the full SLBD approach). After applying these three segmentations using the different sub-models, we trained the new CRF models on the results of the three segmentations to justify the original CWS model. The evaluation results for the supervised CWS methods and the sub-models are presented in Table 2.

It can be seen that we achieved significant improvement in performance when we combined the character-level and phrase-level features in the inner log-linear model, demonstrating that the proposed phrase-level features can be used to efficiently obtain bilingual segmenting information. Moreover, the outer log-linear model achieves a further enhancement, thereby demonstrating that the sentence-level features can be used to effectively re-rank the candidate segmentations produced by the inner log-linear model.

Next, we compared the SLBD method with several state-of-the-art monolingual semi-supervised methods, including those of (Sun et al., 2012) (Sun); (Sun and Xu, 2011) ($S\&X$); (Zeng et al., 2013b) (Zeng). To ensure a fair comparison, we performed the evaluation in two steps. First, we input the entire bilingual unlabeled dataset into the SLBD method and input only the Chinese sentences from the bilingual unlabeled dataset into the other semi-supervised methods. Then, because the available monolingual unlabeled dataset was much larger than the bilingual unlabeled dataset in natural, we used the XIN_CMN portion of Chinese Gigaword 2.0 as an additional unlabeled dataset for the monolingual semi-supervised methods. which contains 204 million words, more than ten times

methods	Bilingual data		Monolingual data	
	F	OOV	F	OOV
Sun	93.9	63.1	94.6	67.9
$S\&X$	94.1	66	94.4	71
Zeng	94.0	64.5	94.8	63.2
SLBD	96.7	80.8	-	-

Table 3: Word segmentation performance of SLBD and other monolingual semi-supervised CWS methods[%]

methods	AS		PKU	
	F	OOV	F	OOV
Xu	92.8	70.5	92.1	66
Ma	93.1	73	92.6	71.1
Xi	90.2	63	90.9	67.2
Zeng2014	93.5	76	93.2	73.3
SLBD	96.7	80.8	97.1	85

Table 4: Word segmentation performance of SLBD and other bilingual semi-supervised CWS methods[%]

the number of words in the bilingual unlabeled dataset. The testing data was the set of AS only. The evaluation is summarized in Table 3.

The results demonstrate that either leveraging the same unlabeled data or providing a much larger unlabeled dataset for the monolingual semi-supervised methods, the SLBD method can significantly outperform the evaluated monolingual semi-supervised methods, which indicates that the segmenting information obtained using SLBD is much more efficient at optimizing segmentation.

Finally, we evaluated SLBD in comparison with other bilingual semi-supervised methods, including (Xu et al., 2008) (Xu); (Ma and Way, 2009) (Ma); (Xi et al., 2012) (Xi);(Zeng et al., 2014) (Zeng2014). The results presented in Table 4 indicate that SLBD demonstrates much stronger performance, primarily because these other methods were developed with a focus on SMT, which causes them to preferentially decrease the perplexity of the subsequent SMT steps rather than producing a highly accurate segmentation. In contrast to these methods, the SLBD method exhibits greater generalizability.

7 Conclusion

In this paper, we propose a cascaded log-linear model to involve learning three levels of bilingual linguistic features to semi-supervisedly learn a new CWS model. Different from other monolingual and bilingual semi-supervised approaches, we employ various types of features to capture both monolingual grammars and bilingual segmenting information, which allows our model to be very efficient at other NLP tasks and endows it with higher generalizability. The evaluation shows that our method significantly outperforms the state-of-the-art monolingual and bilingual semi-supervised approaches.

References

Masayuki Asahara, Kenta Fukuoka, Ai Azuma, ChooiLing Goh, Yotaro Watanabe, Yuji Matsumoto, and Takahashi Tsuzuki. 2005. *Combination of machine learning methods for optimum chinese word segmentation*. In Proceedings of The Fourth SIGHAN Workshop, pages 134-137.

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. *Optimizing Chinese word segmentation for machine translation performance*. In Proceedings of WMT, pages 224-232. Association for Computational Linguistics.

Tagyoung Chung and Daniel Gildea. 2009. *Unsupervised Tokenization for Machine Translation*. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 718-726.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In Proc. 18th International Conf. on Machine Learning.

Percy Liang. 2005. *Semi-supervised learning for natural language.*. Master' s thesis.

Yanjun Ma and Andy Way. 2009. *Bilingually motivated domain-adapted word segmentation for statistical machine translation*. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 549-557.

Mikolov, Tomas, et al. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems. 2013.*

Graham Neubig, Taro Watanabe et al. 2011. *An Unsupervised Model for Joint Phrase Alignment and Extraction*. Proceedings of ACL 2011.

Graham Neubig, Taro Watanabe et al. 2012. *Machine Translation without Words through Substring Alignment*. Proceedings of ACL 2012.

Feng Jiao, Shaojun Wang, and Chi-Hoon Lee. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In Proceedings of ACL, pages 209-216, Sydney, Australia.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.*

Koehn, Philipp, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics, 2007: 177-180.*

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL*, pages 440-447.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003: 160-167.*

Fuchun Peng, Fangfang Feng and Andrew McCallum. 2004. *Chinese segmentation and new word detection using conditional random fields*. Proceedings of the 20th international conference on Computational Linguistics.

Stolcke, Andreas. 2002. SRILM-an extensible language modeling toolkit. In *INTERSPEECH. 2002*

Weiwei Sun and Jia Xu. 2011. *Enhancing chinese word segmentation using unlabeled data.* . In Proceedings of EMNLP 2011.

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

Tian, Liang, et al. 2014. UM-Corpus: a large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation. ELRA Reykjavik, Iceland, 2014.*

Ning Xi, Guangchao Tang, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2012. *Enhancing statistical machine translation with character alignment*. In Proceedings of ACL, pages 285-290. Association for Computational Linguistics.

Jia Xu, Richard Zens, and Hermann Ney. 2004. *Do we need Chinese word segmentation for statistical machine translation?*. Proceedings of the Third SIGHAN Workshop on Chinese Language Learning, pages 122 – 128.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Herman Ney. 2008. *Bayesian Semi-Supervised Chinese Word Segmentation for Statistical Machine Translation*. Proceedings of Coling 2008.

Nianwen Xue. 2003. *Chinese word segmentation as character tagging*. Computational Linguistics and Chinese Language Processing, pages 29 – 48.

Zhang and Clark. 2007. *Chinese segmentation with a word-based perceptron algorithm.*. Proceedings of ACL 2007.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2010. *A unified character-based tagging framework for chinese word segmentation*. ACM Transactions on Asian Language Information Processing, 9(2):5:1-5:32, June.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao and Isabel Trancoso. 2013a. *Graph-based Semi-Supervised Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao and Isabel Trancoso 2013b. *Co-regularizing character-based and word-based models for semi-supervised Chinese word segmentation*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.

Xiaodong Zeng, Derek F. Wong et al. 2014. *Toward Better Chinese Word Segmentation for SMT via Bilingual Constraints*. Proceedings of the Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2014: 1360-1369.