Toward Better Chinese Word Segmentation for SMT via Bilingual Constraints

Xiaodong Zeng[†] **Lidia S. Chao**[†] **Derek F. Wong**[†] **Isabel Trancoso**[‡] **Liang Tian**[†]
[†]NLP[2]CT Lab / Department of Computer and Information Science, University of Macau
[‡]INESC-ID / Instituto Superior Ténico, Lisboa, Portugal
nlp2ct.samuel@gmail.com, {lidiasc, derekfw}@umac.mo,
isabel.trancoso@inesc-id.pt,tianliang0123@gmail.com

Abstract

This study investigates on building a better Chinese word segmentation model for statistical machine translation. It aims at leveraging word boundary information, automatically learned by bilingual character-based alignments, to induce a preferable segmentation model. We propose dealing with the induced word boundaries as soft constraints to bias the continuous learning of a supervised CRFs model, trained by the treebank data (labeled), on the bilingual data (unlabeled). The induced word boundary information is encoded as a graph propagation constraint. The constrained model induction is accomplished by using posterior regularization algorithm. The experiments on a Chinese-to-English machine translation task reveal that the proposed model can bring positive segmentation effects to translation quality.

1 Introduction

Word segmentation is regarded as a critical procedure for high-level Chinese language processing tasks, since Chinese scripts are written in continuous characters without explicit word boundaries (e.g., space in English). The empirical works show that word segmentation can be beneficial to Chinese-to-English statistical machine translation (SMT) (Xu et al., 2005; Chang et al., 2008; Zhao et al., 2013). In fact most current SMT models assume that parallel bilingual sentences should be segmented into sequences of tokens that are meant to be "words" (Ma and Way, 2009). The practice in state-of-the-art MT systems is that Chinese sentences are tokenized by a monolingual supervised word segmentation model trained on the hand-annotated treebank data, e.g., Chinese treebank (CTB) (Xue et al., 2005). These models are conducive to MT to some extent, since they commonly have relatively good aggregate performance and segmentation consistency (Chang et al., 2008). But one outstanding problem is that these models may leave out some crucial segmentation features for SMT, since the output words conform to the treebank segmentation standard designed for monolingually linguistic intuition, rather than specific to the SMT task.

In recent years, a number of works (Xu et al., 2005; Chang et al., 2008; Ma and Way, 2009; Xi et al., 2012) attempted to build segmentation models for SMT based on bilingual unsegmented data, instead of monolingual segmented data. They proposed to learn gainful bilingual knowledge as golden-standard segmentation supervisions for training a bilingual unsupervised model. Frequently, the bilingual knowledge refers to the mappings of an individual English word to one or more consecutive Chinese characters, generated via statistical character-based alignment. They leverage such mappings to either constitute a Chinese word dictionary for maximum-matching segmentation (Xu et al., 2004), or form labeled data for training a sequence labeling model (Paul et al., 2011). The prior works showed that these models help to find some segmentations tailored for SMT, since the bilingual word occurrence feature can be captured by the character-based alignment (Och and Ney, 2003). However, these models tend to miss out other linguistic segmentation patterns as monolingual supervised models, and suffer from the negative effects of erroneously alignments to word segmentation.

This paper proposes an alternative Chinese Word Segmentation (CWS) model adapted to the SMT task, which seeks not only to maintain the advantages of a monolingual supervised model, having hand-annotated linguistic knowledge, but also to assimilate the relevant bilingual segmenta-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1360–1369,
Baltimore, Maryland, USA, June 23-25 2014.

tion nature. We propose leveraging the bilingual knowledge to form learning constraints that guide a supervised segmentation model toward a better solution for SMT. Besides the bilingual motivated models, character-based alignment is also employed to achieve the mappings of the successive Chinese characters and the target language words. Instead of directly merging the characters into concrete segmentations, this work attempts to extract word boundary distributions for character-level trigrams (types) from the "chars-to-word" mappings. Furthermore, these word boundaries are encoded into a graph propagation (GP) expression, in order to widen the influence of the induced bilingual knowledge among Chinese texts. The GP expression constrains similar types having approximated word boundary distributions. Crucially, the GP expression with the bilingual knowledge is then used as side information to regularize a CRFs (conditional random fields) model's learning over treebank and bitext data, based on the posterior regularization (PR) framework (Ganchev et al., 2010). This constrained learning amounts to a jointly coupling of GP and CRFs, i.e., integrating GP into the estimation of a parametric structural model.

This paper is structured as follows: Section 2 points out the main differences with the related works of this study. Section 3 presents the details of the proposed segmentation model. Section 4 reports the experimental results of the proposed model for a Chinese-to-English MT task. The conclusion is drawn in Section 5.

2 Related Work

In the literature, many approaches have been proposed to learn CWS models for SMT. They can be put into two categories, monolingual-motivated and bilingual-motivated. The former primarily optimizes monolingual supervised models according to some predefined segmentation properties that are manually summarized from empirical MT evaluations. Chang et al. (2008) enhanced a CRFs segmentation model in MT tasks by tuning the word granularity and improving the segmentation consistence. Zhang et al. (2008) produced a better segmentation model for SMT by concatenating various corpora regardless of their different specifications. Distinct from their behaviors, this work uses automatically learned constraints instead of manually defined ones. Most importantly, the constraints have a better learning guidance since they originate from the bilingual texts. On the other hand, the bilingual-motivated CWS models typically rely on character-based alignments to generate segmentation supervisions. Xu et al. (2004) proposed to employ "chars-to-word" alignments to generate a word dictionary for maximum matching segmentation in SMT task. The works in (Ma and Way, 2009; Zhao et al., 2013) extended the dictionary extraction strategy. Ma and Way (2009) adopted co-occurrence frequency metric to iteratively optimize "candidate words" extract from the alignments. Zhao et al. (2013) attempted to find an optimal subset of the dictionary learned by the character-based alignment to maximize the MT performance. Paul et al. (2011) used the words learned from "chars-to-word" alignments to train a maximum entropy segmentation model. Rather than playing the "hard" uses of the bilingual segmentation knowledge, i.e., directly merging "char-to-word" alignments to words as supervisions, this study extracts word boundary information of characters from the alignments as soft constraints to regularize a CRFs model's learning.

The graph propagation (GP) technique provides a natural way to represent data in a variety of target domains (Belkin et al., 2006). In this technique, the constructed graph has vertices consisting of labeled and unlabeled examples. Pairs of vertices are connected by weighted edges encoding the degree to which they are expected to have the same label (Zhu et al., 2003). Many recent works, such as by Subramanya et al. (2010), Das and Petrov (2011), Zeng et al. (2013; 2014) and Zhu et al. (2014), proposed GP for inferring the label information of unlabeled data, and then leverage these GP outcomes to learn a semi-supervised scalable model (e.g., CRFs). These approaches are referred to as pipelined learning with GP. This study also works with a similarity graph, encoding the learned bilingual knowledge. But, unlike the prior pipelined approaches, this study performs a joint learning behavior in which GP is used as a learning constraint to interact with the CRFs model estimation.

One of our main objectives is to bias CRFs model's learning on unlabeled data, under a non-linear GP constraint encoding the bilingual knowledge. This is accomplished by the posterior regularization (PR) framework (Ganchev et

al., 2010). PR performs regularization on posteriors, so that the learned model itself remains simple and tractable, while during learning it is driven to obey the constraints through setting appropriate parameters. The closest prior study is constrained learning, or learning with prior knowledge. Chang et al. (2008) described constraint driven learning (CODL) that augments model learning on unlabeled data by adding a cost for violating expectations of constraint features designed by domain knowledge. Mann and McCallum (2008) and McCallum et al. (2007) proposed to employ generalized expectation criteria (GE) to specify preferences about model expectations in the form of linear constraints on some feature expectations.

3 Methodology

This work aims at building a CWS model adapted to the SMT task. The model induction is shown in Algorithm 1. The input data requires two types of training resources, segmented Chinese sentences from treebank $\mathcal{D}_l^c$ and parallel unsegmented sentences of Chinese and foreign language $\mathcal{D}_u^c$ and $\mathcal{D}_u^f$. The first step is to conduct character-based alignment over bitexts $\mathcal{D}_u^c$ and $\mathcal{D}_u^f$, where every Chinese character is an alignment target. Here, we are interested on n-to-1 alignment patterns, i.e., one target word is aligned to one or more source Chinese characters. The second step aims to collect word boundary distributions for all types, i.e., character-level trigrams, according to the n-to-1 mappings (Section 3.1). The third step is to encode the induced word boundary information into a k-nearest-neighbors (k-NN) similarity graph constructed over the entire set of types from $\mathcal{D}_l^c$ and $\mathcal{D}_u^c$ (Section 3.2). The final step trains a discriminative sequential labeling model, conditional random fields, on $\mathcal{D}_l^c$ and $\mathcal{D}_u^c$ under bilingual constraints in a graph propagation expression (Section 3.3). This constrained learning is carried out based on posterior regularization (PR) framework (Ganchev et al., 2010).

Algorithm 1 CWS model induction with bilingual constraints

Require:
Segmented Chinese sentences from treebank $\mathcal{D}_l^c$; Parallel sentences of Chinese and foreign language $\mathcal{D}_u^c$ and $\mathcal{D}_u^f$

Ensure:
θ: the CRFs model parameters

1: $\mathcal{D}^{c\leftrightarrow f} \leftarrow$ **char_align_bitext** $(\mathcal{D}_u^c, \mathcal{D}_u^f)$
2: $r \leftarrow$ **learn_word_bound** $(\mathcal{D}^{c\leftrightarrow f})$
3: $\mathcal{G} \leftarrow$ **encode_graph_constraint** $(\mathcal{D}_l^c, \mathcal{D}_u^c, r)$
4: $\theta \leftarrow$ **pr_crf_graph** $(\mathcal{D}_l^c, \mathcal{D}_u^c, \mathcal{G})$

3.1 Word Boundaries Learned from Character-based Alignments

The gainful supervisions toward a better segmentation solution for SMT are naturally extracted from MT training resources, i.e., bilingual parallel data. This study employs an approximated method introduced in (Xu et al., 2004; Ma and Way, 2009; Chung and Gildea, 2009) to learn bilingual segmentation knowledge. This relies on statistical character-based alignment: first, every Chinese character in the bitexts is divided by a white space so that individual characters are regarded as special "words" or alignment targets, and second, they are connected with English words by using a statistical word aligner, e.g., GIZA++ (Och and Ney, 2003). Note that the aligner is restricted to use an n-to-1 alignment pattern. The primary idea is that consecutive Chinese characters are grouped to a candidate word, if they are aligned to the same foreign word. It is worth mentioning that prior works presented a straightforward usage for candidate words, treating them as golden segmentations, either dictionary units or labeled resources. But this study treats the induced candidate words in a different way. We propose to extract the word boundary distributions[1] for character-level trigrams (*type*)[2], as shown in Figure 1, instead of the very specific words. There are two main reasons to do so. First, it is a more general expression which can reduce the impact amplification of erroneous character alignments. Second, boundary distributions can play more flexible roles as constraints over labelings to bias the model learning.

The type-level word boundary extraction is formally described as follows. Given the ith sentence pair $\langle x_i^c, x_i^f, \mathcal{A}_i^{c\rightarrow f}\rangle$ of the aligned bilingual corpus $\mathcal{D}^{c\leftrightarrow f}$, the Chinese sentence x_i^c consisting of m characters $\{x_{i,1}^c, x_{i,2}^c, ..., x_{i,m}^c\}$, and the foreign language sentence x_i^f, consisting of

[1] The distribution is on four word boundary labels indicating the character positions in a word, i.e., **B** (begin), **M** (middle), **E** (end) and **S** (single character).

[2] A word boundary distribution corresponds to the center character of a type. In fact, it aims at reducing label ambiguities to collect boundary information of character trigrams, rather than individual characters (Altun et al., 2006).

n words $\{x^f_{i,1}, x^f_{i,2}, ..., x^f_{i,n}\}$, $\mathcal{A}^{c\to f}_i$ represents a set of alignment pairs $a_j = \langle C_j, x^f_{i,j}\rangle$ that defines connections between a few Chinese characters $C_j = \{x^c_{i,j_1}, x^c_{i,j_2}, ..., x^c_{i,j_k}\}$ and a single foreign word $x^f_{i,j}$. For an alignment $a_j = \langle C_j, x^f_{i,j}\rangle$, only the sequence of characters $C_j = \{x^c_{i,j_1}, x^c_{i,j_2}, ..., x^c_{i,j_k}\}$ $\forall d \in [1, k-1], j_{d+1} - j_d = 1$ constitutes a valid candidate word. For the whole bilingual corpus, we assign each character in the candidate words with a word boundary tag $T \in \{B, M, E, S\}$, and then count across the entire corpus to collect the tag distributions $r_i = \{r_{i,t}; t \in T\}$ for each type $x^c_{i,j-1}x^c_{i,j}x^c_{i,j+1}$.

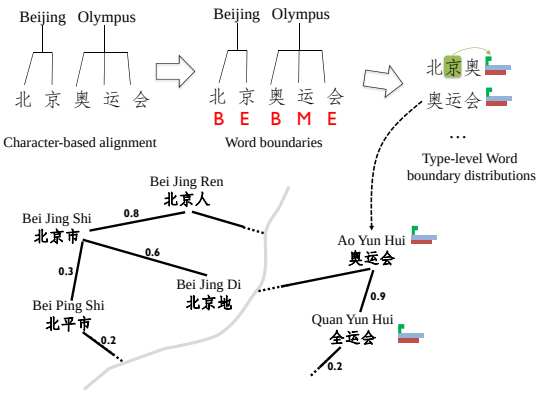

Figure 1: An example of similarity graph over character-level trigrams (types).

3.2 Constraints Encoded by Graph Propagation Expression

The previous step contributes to generate bilingual segmentation supervisions, i.e., type-level word boundary distributions. An intuitive manner is to directly leverage the induced boundary distributions as label constraints to regularize segmentation model learning, based on a constrained learning algorithm. This study, however, makes further efforts to elevate the positive effects of the bilingual knowledge via the graph propagation technique. We adopt a similarity graph to encode the learned type-level word boundary distributions. The GP expression will be defined as a PR constraint in Section 3.3 that reflects the interactions between the graph and the CRFs model. In other words, GP is integrated with estimation of parametric structural model. This is greatly different from the prior pipelined approaches (Subramanya et al., 2010; Das and Petrov, 2011; Zeng et al., 2013), where GP is run first and its propagated outcomes are then used to bias the structural model. This work seeks to capture the GP benefits during the modeling of sequential correlations.

In what follows, the graph setting and propagation expression are introduced. As in conventional GP examples (Das and Smith, 2012), a similarity graph $\mathcal{G} = (V, E)$ is constructed over N types extracted from Chinese training data, including treebank $\mathcal{D}^c_l$ and bitexts $\mathcal{D}^c_u$. Each vertex V_i has a $|T|$-dimensional estimated measure $v_i = \{v_{i,t}; t \in T\}$ representing a probability distribution on word boundary tags. The induced type-level word boundary distributions $r_i = \{r_{i,t}; t \in T\}$ are empirical measures for the corresponding M graph vertices. The edges $E \in V_i \times V_j$ connect all the vertices. Scores between pairs of graph vertices (types), w_{ij}, refer to the similarities of their syntactic environment, which are computed following the method in (Subramanya et al., 2010; Das and Petrov, 2011; Zeng et al., 2013). The similarities are measured based on co-occurrence statistics over a set of predefined features (introduced in Section 4.1). Specifically, the point-wise mutual information (PMI) values, between vertices and each feature instantiation that they have in common, are summed to sparse vectors, and their cosine distances are computed as the similarities. The nature of this similarity graph enforces that the connected types with high weights appearing in different texts should have similar word boundary distributions.

The quality (smoothness) of the similarity graph can be estimated by using a standard propagation function, as shown in Equation 1. The square-loss criterion (Zhu et al., 2003; Bengio et al., 2006) is used to formulate this function:

$$\mathcal{P}(v) = \sum_{t=1}^{T}\Bigg(\sum_{i=1}^{M}(v_{i,t} - r_{i,t})^2 + \mu\sum_{j=1}^{N}\sum_{i=1}^{N} w_{ij}(v_{i,t} - v_{j,t})^2 + \rho\sum_{i=1}^{N}(v_{i,t})^2\Bigg) \quad (1)$$

The first term in this equation refers to seed matches that compute the distances between the estimated measure v_i and the empirical probabilities r_i. The second term refers to edge smoothness that measures how vertices v_i are smoothed with respect to the graph. Two types connected by an edge with high weight should be assigned similar word boundary distributions. The third term, a ℓ_2 norm, evaluates the distribution sparsity (Das and

Smith, 2012) per vertex. Typically, the GP process amounts to an optimization process with respect to parameter v such that Equation 1 is minimized. This propagation function can be used to reflect the graph smoothness, where the higher the score, the lower the smoothness.

3.3 PR Learning with GP Constraint

Our learning problem belongs to semi-supervised learning (SSL), as the training is done on treebank labeled data $(\mathrm{X}_L, \mathrm{Y}_L) = \{(\mathrm{x}_1, \mathrm{y}_1), ..., (\mathrm{x}_l, \mathrm{y}_l)\}$, and bilingual unlabeled data $(\mathrm{X}_U) = \{\mathrm{x}_1, ..., \mathrm{x}_u\}$ where $\mathrm{x}_i = \{x^1, ..., x^m\}$ is an input word sequence and $\mathrm{y}_i = \{y^1, ..., y^m\}, y \in T$ is its corresponding label sequence. Supervised linear-chain CRFs can be modeled in a standard conditional log-likelihood objective with a Gaussian prior:

$$\mathcal{L}(\theta) = p_\theta(\mathrm{y}_i|\mathrm{x}_i) - \frac{\|\theta\|^2}{2\sigma} \tag{2}$$

The conditional probabilities p_θ are expressed as a log-linear form:

$$p_\theta(\mathrm{y}_i|\mathrm{x}_i) = \frac{\exp(\sum_{k=1}^{m} \theta^{\mathrm{T}} f(y_i^{k-1}, y_i^k, \mathrm{x}_i))}{Z_\theta(\mathrm{x}_i)} \tag{3}$$

Where $Z_\theta(\mathrm{x}_i)$ is a partition function that normalizes the exponential form to be a probability distribution, and $f(y_i^{k-1}, y_i^k, \mathrm{x}_i)$ are arbitrary feature functions.

In our setting, the CRFs model is required to learn from unlabeled data. This work employs the posterior regularization (PR) framework[3] (Ganchev et al., 2010) to bias the CRFs model's learning on unlabeled data, under a constraint encoded by the graph propagation expression. It is expected that similar types in the graph should have approximated expected taggings under the CRFs model. We follow the approach introduced by (He et al., 2013) to set up a penalty-based PR objective with GP: the CRFs likelihood is modified by adding a regularization term, as shown in Equation 4, representing the constraints:

$$\mathcal{R}_U(\theta, q) = \mathrm{KL}(q||p_\theta) + \lambda \mathcal{P}(v) \tag{4}$$

Rather than regularize CRFs model's posteriors $p_\theta(\mathcal{Y}|\mathrm{x}_i)$ directly, our model uses an auxiliary distribution $q(\mathcal{Y}|\mathrm{x}_i)$ over the possible labelings $\mathcal{Y}$ for x_i, and penalizes the CRFs marginal log-likelihood by a **KL-divergence** term[4], representing the distance between the estimated posteriors p and the desired posteriors q, as well as a **penalty** term, formed by the GP function. The hyperparameter λ is used to control the impacts of the penalty term. Note that the penalty is fired if the graph score computed based on the expected taggings given by the current CRFs model is increased vis-a-vis the previous training iteration. This nature requires that the penalty term $\mathcal{P}(v)$ should be formed as a function of posteriors q over CRFs model predictions[5], i.e., $\mathcal{P}(q)$. To state this, a mapping $\mathcal{M} : (\{1, ..., u\}, \{1, ..., m\}) \rightarrow V$ from words in the corpus to vertices in the graph is defined. We can thus decompose $v_{i,t}$ into a function of q as follows:

$$v_{i,t} = \frac{\sum_{a=1}^{u} \sum_{\substack{b=1; \\ \mathcal{M}(a,b)=V_i}}^{m} \sum_{c=1}^{T} \sum_{y \in \mathcal{Y}} \mathbf{1}(y^b = t, y^{b-1} = c) q(\mathrm{y}|\mathrm{x}_a)}{\sum_{a=1}^{u} \sum_{b=1}^{m} \mathbf{1}(\mathcal{M}(a,b) = V_i)} \tag{5}$$

The final learning objective combines the CRFs likelihood with the PR regularization term: $\mathcal{J}(\theta, q) = \mathcal{L}(\theta) + \mathcal{R}_U(\theta, q)$. This joint objective, over θ and q, can be optimized by an expectation maximization (EM) style algorithm as reported in (Ganchev et al., 2010). We start from initial parameters θ^0, estimated by supervised CRFs model training on treebank data. The E-step is to minimize $\mathcal{R}_U(\theta, q)$ over the posteriors q that are constrained to the probability simplex. Since the penalty term $\mathcal{P}(v)$ is a non-linear form, the optimization method in (Ganchev et al., 2010) via projected gradient descent on the dual is inefficient[6]. This study follows the optimization method (He et al., 2013) that uses exponentiated gradient descent (EGD) algorithm. It allows that the variable update expression, as shown in Equation 6, takes a multiplicative rather than an additive form.

$$q^{(w+1)}(\mathrm{y}|\mathrm{x}_i) = q^{(w)}(\mathrm{y}|\mathrm{x}_i) \exp(-\eta \frac{\partial \mathcal{R}}{\partial q^{(w)}(\mathrm{y}|\mathrm{x}_i)}) \tag{6}$$

where the parameter η controls the optimization rate in the E-step. With the contributions from

[3]The readers are refered to the original paper of Ganchev et al. (2010).

[4]The form of KL term: $\mathrm{KL}(q||p) = \sum_{q \in \mathcal{Y}} q(\mathrm{y}) \log \frac{q(\mathrm{y})}{p(\mathrm{y})}$.

[5]The original PR setting also requires that the penalty term should be a linear (Ganchev et al., 2010) or non-linear (He et al., 2013) function on q.

[6]According to (He et al., 2013), the dual of quadratic program implies an expensive matrix inverse.

the E-step that further encourage q and p to agree, the M-step aims to optimize the objective $\mathcal{J}(\theta, q)$ with respect to θ. The M-step is similar to the standard CRFs parameter estimation, where the gradient ascent approach still works. This EM-style approach monotonically increases $\mathcal{J}(\theta, q)$ and thus is guaranteed to converge to a local optimum.

$$\begin{aligned}\textbf{E-step:}\quad & q^{(t+1)} = \arg\min_{q} \mathcal{R}_U(\theta^{(t)}, q^{(t)}) \\ \textbf{M-step:}\quad & \theta^{(t+1)} = \arg\max_{\theta} \mathcal{L}(\theta) \\ & + \delta \sum_{i=1}^{u} \sum_{\mathrm{y} \in \mathcal{Y}} q^{(t+1)}(\mathrm{y}|\mathrm{x}_i) \log p_\theta(\mathrm{y}|\mathrm{x}_i)\end{aligned} \tag{7}$$

4 Experiments

4.1 Data and Setup

The experiments in this study evaluated the performances of various CWS models in a Chinese-to-English translation task. The influence of the word segmentation on the final translation is our main investigation. We adopted three state-of-the-art metrics, BLEU (Papineni et al., 2002), NIST (Doddington et al., 2000) and METEOR (Banerjee and Lavie, 2005), to evaluate the translation quality.

The monolingual segmented data, $\mathrm{train}_{\mathrm{TB}}$, is extracted from the Penn Chinese Treebank (CTB-7) (Xue et al., 2005), containing 51,447 sentences. The bilingual training data, $\mathrm{train}_{\mathrm{MT}}$, is formed by a large in-house Chinese-English parallel corpus (Tian et al., 2014). There are in total 2,244,319 Chinese-English sentence pairs crawled from online resources, concentrated in 5 different domains including *laws*, *novels*, *spoken*, *news* and *miscellaneous*[7]. This in-house bilingual corpus is the MT training data as well. The target-side language model is built on over 35 million monolingual English sentences, $\mathrm{train}_{\mathrm{LM}}$, crawled from online resources. The NIST evaluation campaign data, MT-03 and MT-05, are selected to comprise the MT development data, $\mathrm{dev}_{\mathrm{MT}}$, and testing data, $\mathrm{test}_{\mathrm{MT}}$, respectively.

For the settings of our model, we adopted the standard feature templates introduced by Zhao et al. (2006) for CRFs. The character-based alignment for achieving the "chars-to-word" mappings is accomplished by GIZA++ aligner (Och and Ney, 2003). For the GP, a 10-NNs similarity graph was constructed[8]. Following (Subramanya et al., 2010; Zeng et al., 2013), the features used to compute similarities between vertices were (Suppose given a type "$w_2w_3w_4$" surrounding contexts "$w_1w_2w_3w_4w_5$"): unigram (w_3), bigram (w_1w_2, w_4w_5, w_2w_4), trigram ($w_2w_3w_4$, $w_2w_4w_5$, $w_1w_2w_4$), trigram+context ($w_1w_2w_3w_4w_5$) and character classes in number, punctuation, alphabetic letter and other ($t(w_2)t(w_3)t(w_4)$). There are four hyperparameters in our model to be tuned by using the development data ($\mathrm{dev}_{\mathrm{MT}}$) among the following settings: for the graph propagation, $\mu \in \{0.2, 0.5, 0.8\}$ and $\rho \in \{0.1, 0.3, 0.5, 0.8\}$; for the PR learning, $\lambda \in \{0 \leq \lambda_i \leq 1\}$ and $\sigma \in \{0 \leq \sigma_i \leq 1\}$ where the step is 0.1. The best performed joint settings, $\mu = 0.5, \rho = 0.5, \lambda = 0.9$ and $\sigma = 0.8$, were used to measure the final performance.

The MT experiment was conducted based on a standard log-linear phrase-based SMT model. The GIZA++ aligner was also adopted to obtain word alignments (Och and Ney, 2003) over the segmented bitexts. The heuristic strategy of *grow-diag-final-and* (Koehn et al., 2007) was used to combine the bidirectional alignments for extracting phrase translations and reordering tables. A 5-gram language model with Kneser-Ney smoothing was trained with SRILM (Stolcke, 2002) on monolingual English data. Moses (Koehn et al., 2007) was used as decoder. The Minimum Error Rate Training (MERT) (Och, 2003) was used to tune the feature parameters on development data.

4.2 Various Segmentation Models

To provide a thorough analysis, the MT experiments in this study evaluated three baseline segmentation models and two off-the-shelf models, in addition to four variant models that also employ the bilingual constraints. We start from three baseline models:

- **Character Segmenter (CS)**: this model simply divides Chinese sentences into sequences of characters.
- **Supervised Monolingual Segmenter (SMS)**: this model is trained by CRFs on treebank training data ($\mathrm{train}_{\mathrm{TB}}$). The same feature templates (Zhao et al., 2006) are used. The standard four-tags (**B**, **M**, **E** and **S**) were used

[7]The in-house corpus has been manually validated, in a long process that exceeded 500 hours.

[8]We evaluated graphs with top k (from 3 to 20) nearest neighbors on development data, and found that the performance converged beyond 10-NNs.

as the labels. The stochastic gradient descent is adopted to optimize the parameters.

- **Unsupervised Bilingual Segmenter (UBS)**: this model is trained on the bitexts (trainMT) following the approach introduced in (Ma and Way, 2009). The optimal set of the model parameter values was found on $\mathrm{dev_{MT}}$ to be $k = 3, t_{AC} = 0.0$ and $t_{COOC} = 15$.

The comparison candidates also involve two popular off-the-shelf segmentation models:

- **Stanford Segmenter**: this model, trained by Chang et al. (2008), treats CWS as a binary word boundary decision task. It covers several features specific to the MT task, e.g., external lexicons and proper noun features.

- **ICTCLAS Segmenter**: this model, trained by Zhang et al. (2003), is a hierarchical HMM segmenter that incorporates parts-of-speech (POS) information into the probability models and generates multiple HMM models for solving segmentation ambiguities.

This work also evaluated four variant models[9] that perform alternative ways to incorporate the bilingual constraints based on two state-of-the-art graph-based SSL approaches.

- **Self-training Segmenters (STS)**: two variant models were defined by the approach reported in (Subramanya et al., 2010) that uses the supervised CRFs model's decodings, incorporating empirical and constraint information, for unlabeled examples as additional labeled data to retrain a CRFs model. One variant (STS-NO-GP) skips the GP step, directly decoding with type-level word boundary probabilities induced from bitexts, while the other (STS-GP-PL) runs the GP at first and then decodes with GP outcomes. The optimal hyperparameter values were found to be: STS-NO-GP ($\alpha = 0.8$) and $\eta = 0.6$) and STS-GP-PL ($\mu = 0.5, \rho = 0.3, \alpha = 0.8$ and $\eta = 0.6$).

- **Virtual Evidences Segmenters (VES)**: Two variant models based on the approach in (Zeng et al., 2013) were defined. The type-level word boundary distributions, induced by the character-based alignment (VES-NO-GP), and the graph propagation (VES-GP-PL), are regarded as virtual evidences to bias CRFs model's learning on the unlabeled data. The optimal hyperparameter values were found to be: VES-NO-GP ($\alpha = 0.7$) and VES-GP-PL ($\mu = 0.5, \rho = 0.3$ and $\alpha = 0.7$).

[9]Note that there are two variant models working with GP. To be fair, the same similarity graph settings introduced in this paper were used.

4.3 Main Results

Table 1 summarizes the final MT performance on the MT-05 test data, evaluated with ten different CWS models. In what follows, we summarized four major observations from the results. Firstly, as expected, having word segmentation does help Chinese-to-English MT. All other nine CWS models outperforms the CS baseline which does not try to identify Chinese words at all. Secondly, the other two baselines, SMS and UBS, are on a par with each other, showing less than 0.36 average performance differences on the three evaluation metrics. This outcome validated that the models, trained by either the treebank or the bilingual data, performed reasonably well. But they only capture partial segmentation features so that less gains for SMT are achieved when comparing to other sophisticated models. Thirdly, we notice that the two off-the-shelf models, Stanford and ICTCLAS, just brought minor improvements over the SMS baseline, although they are trained using richer supervisions. This behaviour illustrates that the conventional optimizations to the monolingual supervised model, e.g., accumulating more supervised data or predefined segmentation properties, are insufficient to help model for achieving better segmentations for SMT. Finally, highlighting the five models working with the bilingual constraints, most of them can achieve significant gains over the other ones without using the bilingual constraints. This strongly demonstrates that bilingually-learned segmentation knowledge does helps CWS for SMT. The models working with G-P, STS-GP-PL, VES-GP-PL and ours outperform all others. We attribute this to the role of GP in assisting the spread of bilingual knowledge on the Chinese side. Importantly, it can be observed that our model outperforms STS-GP, VES-GP, which greatly supports that joint learning of CRFs and GP can alleviate the error transfer by the pipelined models. This is one of the most crucial findings in this study. Overall, the boldface numbers in the last row illustrate that our model obtains average improvements of 1.89, 1.76 and 1.61 on BLEU,

NIST and METEOR over others.

Models	BLEU	NIST	METEOR
CS	29.38	59.85	54.07
SMS	30.05	61.33	55.95
UBS	30.15	61.56	55.39
Stanford	30.40	61.94	56.01
ICTCLAS	30.29	61.26	55.72
STS-NO-GP	31.47	62.35	56.12
STS-GP-PL	31.94	63.20	57.09
VES-NO-GP	31.98	62.63	56.59
VES-GP-PL	32.04	63.49	57.34
Our Model	**32.75**	**63.72**	**57.64**

Table 1: Translation performances (%) on MT-05 testing data by using ten different CWS models.

4.4 Analysis & Discussion

This section aims to further analyze the three primary observations concluded in Section 4.3: i) word segmentation is useful to SMT; ii) the treebank and the bilingual segmentation knowledge are helpful, performing segmentation of different nature; and iii) the bilingual constraints lead to learn segmentations better tailored for SMT.

The first observation derives from the comparisons between the CS baseline and other models. Our results, showing the significant CWS benefits to SMT, are consistent with the works reported in the literature (Xu et al., 2004; Chang et al., 2008). In our experiment, two additional evidences found in the translation model are provided to further support that NO tokenization of Chinese (i.e., the CS model's output) could harm the MT system. First, the SMT phrase extraction, i.e., building "phrases" on top of the character sequences, cannot fully capture all meaningful segmentations produced by the CS model. The character based model leads to missing some useful longer phrases, and to generate many meaningless or redundant translations in the phrase table. Moreover, it is affected by translation ambiguities, caused by the cases where a Chinese character has very different meanings in different contextual environments.

The second observation shifts the emphasis to SMS and UBS, based on the treebank and the bilingual segmentation, respectively. Our results show that both segmentation patterns can bring positive effects to MT. Through analyzing both models' segmentations for $\mathrm{train}_{\mathrm{MT}}$ and $\mathrm{test}_{\mathrm{MT}}$, we attempted to get a closer inspection on the segmentation preferences and their influence on MT. Our first finding is that the segmentation consensuses between SMS and UBS are positive to MT. There have about 35% identical segmentations produced by the two models. If these identical segmentations are removed, and the experiments are rerun, the translation scores decrease (on average) by 0.50, 0.85 and 0.70 on BLEU, NIST and METEOR, respectively. Our second finding is that SMS exhibits better segmentation consistency than UBS. One representative example is the segmentations for "孤零零 (lonely)". All the outputs of SMS were "孤零零", while UBS generated three ambiguous segmentations, "孤(alone)_零零(double zero)", "孤 零(lonely)_零(zero)" and "孤(alone)_零(zero)_零(zero)". The segmentation consistency of SMS rests on the high-quality treebank data and the robust CRFs tagging model. On the other hand, the advantage of UBS is to capture the segmentations matching the aligned target words. For example, UBS grouped "国(country)_际(border)_间(between)" to a word "国 际 间(international)", rather than two words "国际(international)_间(between)" (as given by SMS), since these three characters are aligned to a single English word "international". The above analysis shows that SMS and UBS have their own merits and combining the knowledge derived from both segmentations is highly encouraged.

The third observation concerns the great impact of the bilingual constraints to the segmentation models in the MT task. The use of the bilingual constraints is the prime objective of this study. Our first contribution for this purpose is on using the word boundary distributions to capture the bilingual segmentation supervisions. This representation contributes to reduce the negative impacts of erroneous "chars-to-word" alignments. The ambiguous types (having relatively uniform boundary distribution), caused by alignment errors, cannot directly bias the model tagging preferences. Furthermore, the word boundary distributions are convenient to make up the learning constraints over the labelings among various constrained learning approaches. They have successfully played in three types of constraints for our experiments: PR penalty (Our model), decoding constraints in self-training (STS) and virtual evidences (VES). The second contribution is the use of GP, illustrated by STS-GP-PL, VES-GP-PL and

Our model. The major effect is to multiply the impacts of the bilingual knowledge through the similarity graph. The graph vertices (types)[10], without any supervisions, can learn the word boundary information from their similar types (neighborhoods) having the empirical boundary probabilities. The segmentations given by the three GP models show about 70% positive segmentation changes, affected by the unlabeled graph vertices, with respect to the ones given by the NO-GP models, STS-NO-GP and VES-NO-GP. In our opinion, the learning mechanism of our approach, joint coupling of GP and CRFs, rather than the pipelined one as the other two models, contributes to maximizing the graph smoothness effects to the CRFs estimation so that the error propagation of the pipelined approaches is alleviated.

5 Conclusion

This paper proposed a novel CWS model for the SMT task. This model aims to maintain the linguistic segmentation supervisions from treebank data and simultaneously integrate useful bilingual segmentations induced from the bitexts. This objective is accomplished by three main steps: 1) learn word boundaries from character-based alignments; 2) encode the learned word boundaries into a GP constraint; and 3) training a CRFs model, under the GP constraint, by using the PR framework. The empirical results indicate that the proposed model can yield better segmentations for SMT.

Acknowledgments

The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau (Grant No. MYRG076 (Y1-L2)-FST13-WF and MYRG070 (Y1-L2)-FST12-CS) for the funding support for our research. The work of Isabel Trancoso was supported by national funds through FCT-Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013. The authors also wish to thank the anonymous reviewers for many helpful comments.

[10]This experiment yielded a similarity graph that consists of 11,909,620 types from $\mathrm{train}_{\mathrm{TB}}$ and $\mathrm{train}_{\mathrm{MT}}$, where there have 8,593,220 (72.15%) types without any empirical boundary distributions.

References

Yasemin Altun, David McAllester, and Mikhail Belkin. 2006. Maximum margin semi-supervised learning for structured variables. *Advances in Neural Information Processing Systems*, 18:33.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 2006. Label propagation and quadratic criterion. *Semi-Supervised Learning*, pages 193–216.

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. In *Proceedings of EMNLP*, pages 718–726. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL*, pages 600–609. Association for Computational Linguistics.

Dipanjan Das and Noah A Smith. 2012. Graph-based lexicon expansion with sparsity-inducing penalties. In *Proceedings of NAACL*, pages 677–687. Association for Computational Linguistics.

George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. 2000. The nist speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech Communication*, 31(2):225–254.

Kuzman Ganchev, Jõao Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049.

Luheng He, Jennifer Gillenwater, and Ben Taskar. 2013. Graph-based posterior regularization for semi-supervised structured prediction. In *Proceedings of CoNLL*, page 38. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Yanjun Ma and Andy Way. 2009. Bilingually motivated domain-adapted word segmentation for statistical machine translation. In *Proceedings of EACL*, pages 549–557. Association for Computational Linguistics.

Gideon S. Mann and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proceedings of ACL*, pages 870–878. Association for Computational Linguistics.

Andrew McCallum, Gideon Mann, and Gregory Druck. 2007. Generalized expectation criteria. *Computer Science Technical Note, University of Massachusetts, Amherst, MA*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318. Association for Computational Linguistics.

Michael Paul, Finch Andrew, and Sumita Eiichiro. 2011. Integration of multiple bilingually-trained segmentation schemes into statistical machine translation. *IEICE Transactions on Information and Systems*, 94(3):690–697.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of Interspeech*.

Amarnag Subramanya, Slav Petrov, and Fernando Pereira. 2010. Efficient graph-based semi-supervised learning of structured tagging models. In *Proceedings of EMNLP*, pages 167–176. Association for Computational Linguistics.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Shuo Li, Yiming Wang, and Yi Lu. 2014. UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of LREC*. European Language Resources Association.

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? In *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning*, pages 122–128. Association for Computational Linguistics.

Jia Xu, Evgeny Matusov, Richard Zens, and Hermann Ney. 2005. Integrated Chinese word segmentation in statistical machine translation. In *Proceedings of IWSLT*, pages 216–223. Association for Computational Linguistics.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Isabel Trancoso. 2013. Graph-based semi-supervised model for joint Chinese word segmentation and part-of-speech tagging. In *Proceedings of ACL*, pages 770–779. Association for Computational Linguistics.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, Isabel Trancoso, Liangye He, and Qiuping Huang. 2014. Lexicon expansion for latent variable grammars. *Pattern Recognition Letters*, 42:47–55.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi 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. Association for Computational Linguistics.

Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. 2008. Improved statistical machine translation by multiple Chinese word segmentation. In *Proceedings of WMT*, pages 216–223. Association for Computational Linguistics.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics.

Hai Zhao, Masao Utiyama, Eiichiro Sumita, and Bao-Liang Lu. 2013. An empirical study on word segmentation for Chinese machine translation. In *Computational Linguistics and Intelligent Text Processing*, pages 248–263. Springer.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of ICML*, volume 3, pages 912–919.

Ling Zhu, Derek F. Wong, and Lidia S. Chao. 2014. Unsupervised chunking based on graph propagation from bilingual corpus. *The Scientific World Journal*, 2014(401943):10.