Graph-based Semi-Supervised Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging

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

Abstract

This paper introduces a graph-based semi-supervised joint model of Chinese word segmentation and part-of-speech tagging. The proposed approach is based on a graph-based label propagation technique. One constructs a nearest-neighbor similarity graph over all trigrams of labeled and unlabeled data for propagating syntactic information, i.e., label distributions. The derived label distributions are regarded as virtual evidences to regularize the learning of linear conditional random fields (CRFs) on unlabeled data. An inductive character-based joint model is obtained eventually. Empirical results on Chinese tree bank (CTB-7) and Microsoft Research corpora (MSR) reveal that the proposed model can yield better results than the supervised baselines and other competitive semi-supervised CRFs in this task.

1 Introduction

Word segmentation and part-of-speech (POS) tagging are two critical and necessary initial procedures with respect to the majority of high-level Chinese language processing tasks such as syntax parsing, information extraction and machine translation. The traditional way of segmentation and tagging is performed in a pipeline approach, first segmenting a sentence into words, and then assigning each word a POS tag. The pipeline approach is very simple to implement, but frequently causes error propagation, given that wrong segmentations in the earlier stage harm the subsequent POS tagging (Ng and Low, 2004). The joint approaches of word segmentation and POS tagging (joint S&T) are proposed to resolve these two tasks simultaneously. They effectively alleviate the error propagation, because segmentation and tagging have strong interaction, given that most segmentation ambiguities cannot be resolved without considering the surrounding grammatical constructions encoded in a POS sequence (Qian and Liu, 2012).

In the past years, several proposed supervised joint models (Ng and Low, 2004; Zhang and Clark, 2008; Jiang et al., 2009; Zhang and Clark, 2010) achieved reasonably accurate results, but the outstanding problem among these models is that they rely heavily on a large amount of labeled data, i.e., segmented texts with POS tags. However, the production of such labeled data is extremely time-consuming and expensive (Jiao et al., 2006; Jiang et al., 2009). Therefore, semi-supervised joint S&T appears to be a natural solution for easily incorporating accessible unlabeled data to improve the joint S&T model. This study focuses on using a graph-based label propagation method to build a semi-supervised joint S&T model. Graph-based label propagation methods have recently shown they can outperform the state-of-the-art in several natural language processing (NLP) tasks, e.g., POS tagging (Subramanya et al., 2010), knowledge acquisition (Talukdar et al., 2008), shallow semantic parsing for unknown predicate (Das and Smith, 2011). As far as we know, however, these methods have not yet been applied to resolve the problem of joint Chinese word segmentation (CWS) and POS tagging.

Motivated by the works in (Subramanya et al., 2010; Das and Smith, 2011), for structured problems, graph-based label propagation can be employed to infer valuable syntactic information (n-gram-level label distributions) from labeled data to unlabeled data. This study extends this intuition to construct a similarity graph for propagating trigram-level label distributions. The derived label distributions are regarded as prior knowledge to regularize the learning of a sequential model, conditional random fields (CRFs) in this case, on both

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 770–779,
Sofia, Bulgaria, August 4-9 2013.

labeled and unlabeled data to achieve the semi-supervised learning. The approach performs the incorporation of the derived labeled distributions by manipulating a "virtual evidence" function as described in (Li, 2009). Experiments on the data from the Chinese tree bank (CTB-7) and Microsoft Research (MSR) show that the proposed model results in significant improvement over other comparative candidates in terms of F-score and out-of-vocabulary (OOV) recall.

This paper is structured as follows: Section 2 points out the main differences with the related work of this study. Section 3 reviews the background, including supervised character-based joint S&T model based on CRFs and graph-based label propagation. Section 4 presents the details of the proposed approach. Section 5 reports the experiment results. The conclusion is drawn in Section 6.

2 Related Work

Prior supervised joint S&T models present approximate 0.2% - 1.3% improvement in F-score over supervised pipeline ones. The state-of-the-art joint models include reranking approaches (Shi and Wang, 2007), hybrid approaches (Nakagawa and Uchimoto, 2007; Jiang et al., 2008; Sun, 2011), and single-model approaches (Ng and Low, 2004; Zhang and Clark, 2008; Kruengkrai et al., 2009; Zhang and Clark, 2010). The proposed approach in this paper belongs to the single-model type.

There are few explorations of semi-supervised approaches for CWS or POS tagging in previous works. Xu et al. (2008) described a Bayesian semi-supervised CWS model by considering the segmentation as the hidden variable in machine translation. Unlike this model, the proposed approach is targeted at a general model, instead of one oriented to machine translation task. Sun and Xu (2011) enhanced a CWS model by interpolating statistical features of unlabeled data into the CRFs model. Wang et al. (2011) proposed a semi-supervised pipeline S&T model by incorporating *n*-gram and lexicon features derived from unlabeled data. Different from their concern, our emphasis is to learn the semi-supervised model by injecting the label information from a similarity graph constructed from labeled and unlabeled data.

The induction method of the proposed approach also differs from other semi-supervised CRFs algorithms. Jiao et al. (2006), extended by Mann and McCallum (2007), reported a semi-supervised CRFs model which aims to guide the learning by minimizing the conditional entropy of unlabeled data. The proposed approach regularizes the CRFs by the graph information. Subramanya et al. (2010) proposed a graph-based self-train style semi-supervised CRFs algorithm. In the proposed approach, an analogous way of graph construction intuition is applied. But overall, our approach differs in three important aspects: first, novel feature templates are defined for measuring the similarity between vertices. Second, the critical property, i.e., sparsity, is considered among label propagation. And third, the derived label information from the graph is smoothed into the model by optimizing a modified objective function.

3 Background

3.1 Supervised Character-based Model

The character-based joint S&T approach is operated as a sequence labeling fashion that each Chinese character, i.e., *hanzi*, in the sequence is assigned with a tag. To perform segmentation and tagging simultaneously in a uniform framework, according to Ng and Low (2004), the tag is composed of a word boundary part, and a POS part, e.g., "B_NN" refers to the first character in a word with POS tag "NN". In this paper, 4 word boundary tags are employed: B (beginning of a word), M (middle part of a word), E (end of a word) and S (single character). As for the POS tag, we shall use the 33 tags in the Chinese tree bank. Thus, the potential composite tags of joint S&T consist of 132 (4×33) classes.

The first-order CRFs model (Lafferty et al., 2001) has been the most common one in this task. Given a set of labeled examples $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{l}$, where $x_i = x_i^1 x_i^2 ... x_i^N$ is the sequence of characters in the *i*th sentence, and $y_i = y_i^1 y_i^2 ... y_i^N$ is the corresponding label sequence. The goal is to learn a CRFs model in the form,

$$p(y_i|x_i;\Lambda) = \frac{1}{Z(x_i;\Lambda)} \exp\{\sum_{j=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_i^{j-1}, y_i^j, x_i, j)\} \quad (1)$$

where $Z(x_i;\Lambda)$ is the partition function that normalizes the exponential form to be a probability distribution, and $f_k(y_i^{j-1}, y_i^j, x_i, j)$. In this study,

the baseline feature templates of joint S&T are the ones used in (Ng and Low, 2004; Jiang et al., 2008), as shown in Table 1. $\Lambda = \{\lambda_1 \lambda_2 ... \lambda_K\} \in \mathbb{R}^K$ are the weight parameters to be learned. In supervised training, the aim is to estimate the Λ that maximizes the conditional likelihood of the training data while regularizing model parameters:

$$\mathcal{L}(\Lambda) = \sum_{i=1}^{l} \log p(y_i|x_i; \Lambda) - R(\Lambda) \qquad (2)$$

$R(\Lambda)$ can be any standard regularizer on parameters, e.g., $R(\Lambda) = \| \Lambda \| /2\delta^2$, to limit overfitting on rare features and avoid degeneracy in the case of correlated features. This objective function can be optimized by the stochastic gradient method or other numerical optimization methods.

Type	Font Size
Unigram	$C_n(n = -2, -1, 0, 1, 2)$
Bigram	$C_n C_{n+1}(n = -2, -1, 0, 1)$
Date, Digit and	$T(C_{-2})T(C_{-1})T(C_0)$
Alphabetic Letter	$T(C_1)T(C_2)$

Table 1: The feature templates of joint S&T.

3.2 Graph-based Label Propagation

Graph-based label propagation, a critical subclass of semi-supervised learning (SSL), has been widely used and shown to outperform other SSL methods (Chapelle et al., 2006). Most of these algorithms are transductive in nature, so they cannot be used to predict an unseen test example in the future (Belkin et al., 2006). Typically, graph-based label propagation algorithms are run in two main steps: graph construction and label propagation. The graph construction provides a natural way to represent data in a variety of target domains. One constructs a graph whose vertices consist of labeled and unlabeled examples. Pairs of vertices are connected by weighted edges which encode the degree to which they are expected to have the same label (Zhu et al., 2003). Popular graph construction methods include *k*-nearest neighbors (*k*-NN) (Bentley, 1980; Beygelzimer et al., 2006), *b*-matching (Jebara et al., 2009) and local reconstruction (Daitch et al., 2009). Label propagation operates on the constructed graph. The primary objective is to propagate labels from a few labeled vertices to the entire graph by optimizing a loss function based on the constraints or properties derived from the graph, e.g., smoothness (Zhu et al., 2003; Subramanya et al., 2010; Talukdar et al., 2008), or sparsity (Das and Smith, 2012). State-of-the-art label propagation algorithms include LP-ZGL (Zhu et al., 2003), Adsorption (Baluja et al., 2008), MAD (Talukdar and Crammer, 2009) and Sparse Inducing Penalties (Das and Smith, 2012).

4 Method

The emphasis of this work is on building a joint S&T model based on two different kinds of data sources, labeled and unlabeled data. In essence, this learning problem can be treated as incorporating certain gainful information, e.g., prior knowledge or label constraints, of unlabeled data into the supervised model. The proposed approach employs a transductive graph-based label propagation method to acquire such gainful information, i.e., label distributions from a similarity graph constructed over labeled and unlabeled data. Then, the derived label distributions are injected as virtual evidences for guiding the learning of CRFs.

Algorithm 1 semi-supervised joint S&T induction

Input:
$\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{l}$ labeled sentences
$\mathcal{D}_u = \{(x_i)\}_{i=l+1}^{l+u}$ unlabeled sentences
Output:
Λ: a set of feature weights
1: **Begin**
2: $\{\mathcal{G}\} =$ **construct_graph** $(\mathcal{D}_l, \mathcal{D}_u)$
3: $\{q_0\} =$ **init_labelDist** $(\{\mathcal{G}\})$
4: $\{q\} =$ **propagate_label** $(\{\mathcal{G}\}, \{q_0\})$
5: $\{\Lambda\} =$ **train_crf** $(\mathcal{D}_l \cup \mathcal{D}_u, \{q\})$
6: **End**

The model induction includes the following steps (see Algorithm 1): firstly, given labeled and unlabeled data, i.e., $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{l}$ with l labeled sentences and $D_u = \{(x_i)\}_{i=l+1}^{l+u}$ with u unlabeled sentences, a specific similarity graph $\mathcal{G}$ representing $\mathcal{D}_l$ and $\mathcal{D}_u$ is constructed (**construct_graph**). The vertices (Section 4.1) in the constructed graph consist of all *trigrams* that occur in labeled and unlabeled sentences, and edge weights between vertices are computed using the cosine distance between pointwise mutual information (PMI) statistics. Afterwards, the estimated label distributions q_0 of vertices in the graph $\mathcal{G}$ are randomly initialized (**init_labelDist**). Subsequently,

the label propagation procedure (**propagate_label**) is conducted for projecting label distributions q from labeled vertices to the entire graph, using the algorithm of Sparse-Inducing Penalties (Das and Smith, 2012) (Section 4.2). The final step (**train_crf**) of the induction is incorporating the inferred trigram-level label distributions q into CRFs model (Section 4.3).

4.1 Graph Construction

In most graph-based label propagation tasks, the final effect depends heavily on the quality of the graph. Graph construction thus plays a central role in graph-based label propagation (Zhu et al., 2003). For character-based joint S&T, unlike the unstructured learning problem whose vertices are formed directly by labeled and unlabeled instances, the graph construction is non-trivial. Das and Petrov (2011) mentioned that taking individual characters as the vertices would result in various ambiguities, whereas the similarity measurement is still challenging if vertices corresponding to entire sentences.

This study follows the intuitions of graph construction from Subramanya et al. (2010) in which vertices are represented by character trigrams occurring in labeled and unlabeled sentences. Formally, given a set of labeled sentences $\mathcal{D}_l$, and unlabeled ones $\mathcal{D}_u$, where $\mathcal{D} \triangleq \{\mathcal{D}_l, \mathcal{D}_u\}$, the goal is to form an undirected weighted graph $\mathcal{G} = (V, E)$, where V is defined as the set of vertices which covers all trigrams extracted from $\mathcal{D}_l$ and $\mathcal{D}_u$. Here, $V = V_l \cup V_u$, where V_l refers to trigrams that occurs at least once in labeled sentences and V_u refers to trigrams that occur only in unlabeled sentences. The edges $E \in V_l \times V_u$, connect all the vertices. This study makes use of a symmetric k-NN graph (k = 5) and the edge weights are measured by a symmetric similarity function (Equation (3)):

$$w_{i,j} = \begin{cases} \mathrm{sim}(x_i, x_j) & \text{if } j \in \mathcal{K}(i) \text{ or } i \in \mathcal{K}(j) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $\mathcal{K}(i)$ is the set of the k nearest neighbors of $x_i (|\mathcal{K}(i) = k, \forall i|)$ and $\mathrm{sim}(x_i, x_j)$ is a similarity measure between two vertices. The similarity is computed based on the co-occurrence statistics over the features in Table 2. Most features we adopted are selected from those of (Subramanya et al., 2010). Note that a novel feature in the last row encodes the classes of surrounding characters, where four types are defined: number, punctuation, alphabetic letter and other. It is especially helpful for the graph to make connections with trigrams that may not have been seen in labeled data but have similar label information. The pointwise mutual information values between the trigrams and each feature instantiation that they have in common are summed to sparse vectors, and their cosine distances are computed as the similarities.

Description	Feature
Trigram + Context	$x_1x_2x_3x_4x_5$
Trigram	$x_2x_3x_4$
Left Context	x_1x_2
Right Context	x_4x_5
Center Word	x_3
Trigram - Center Word	x_2x_4
Left Word + Right Context	$x_2x_4x_5$
Right Word + Left Context	$x_1x_2x_3$
Type of Trigram: number, punctuation, alphabetic letter and other	$t(x_2)t(x_3)t(x_4)$

Table 2: Features employed to measure the similarity between two vertices, in a given text "$x_1x_2x_3x_4x_5$", where the trigram is "$x_2x_3x_4$".

The nature of the similarity graph enforces that the connected trigrams with high weight appearing in different texts should have similar syntax configurations. Thus, the constructed graph is expected to provide additional information that cannot be expressed directly in a sequence model (Subramanya et al., 2010). One primary benefit of this property is on enriching vocabulary coverage. In other words, the new features of various trigrams only occurring in unlabeled data can be discovered. As the excerpt in Figure 1 shows, the trigram "天津港" (Tianjin port) has no any label information, as it only occurs in unlabeled data, but fortunately its neighborhoods with similar syntax information, e.g., "上海港" (Shanghai port), "广州港" (Guangzhou port), can assist to infer the correct tag "M_NN".

4.2 Label Propagation

In order to induce trigram-level label distributions from the graph constructed by the previous step, a label propagation algorithm, Sparsity-Inducing Penalties, proposed by Das and Smith (2012), is employed. This algorithm is used because it captures the property of sparsity that only a few labels

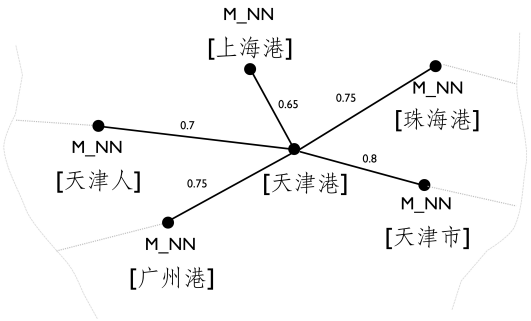

Figure 1: An excerpt from the similarity graph over trigrams on labeled and unlabeled data.

are typically associated with a given instance. In fact, the sparsity is also a common phenomenon among character-based CWS and POS tagging. The following convex objective is optimized on the similarity graph in this case:

$$\begin{aligned}&\operatorname*{argmin}_{q} \sum_{j=1}^{l} \parallel q_j - r_j \parallel^2 \\ &+\mu \sum_{i=1,k\in\mathcal{N}(i)}^{l+u} w_{ik} \parallel q_i - q_k \parallel^2 +\lambda \sum_{i=1}^{l+u} \parallel q_i \parallel^2 \\ &\text{s.t. } q_i \geq 0, \forall i \in V\end{aligned} \tag{4}$$

where r_j denotes empirical label distributions of labeled vertices, and q_i denotes unnormalized estimate measures in every vertex. The w_{ik} refers to the similarity between the *i*th trigram and the *k*th trigram, and $\mathcal{N}(i)$ is a set of neighbors of the *i*th trigram. μ and λ are two hyperparameters whose values are discussed in Section 5. The squared-loss criterion[1] is used to formulate the objective function. The first term in Equation (4) is the seed match loss which penalizes the estimated label distributions q_j, if they go too far away from the empirical labeled distributions r_j. The second term is the edge smoothness loss that requires q_i should be smooth with respect to the graph, such that two vertices connected by an edge with high weight should be assigned similar labels. The final term is a regularizer to incorporate the prior knowledge, e.g., uniform distributions used in (Talukdar et al., 2008; Das and Smith, 2011). This study applies the squared norm of q to encourage sparsity per vertex. Note that the estimated label distribution q_i in Equation (4) is relaxed to be unnormalized, which simplifies the optimization. Thus, the objective function can be optimized by L-BFGS-B (Zhu et al., 1997), a generic quasi-Newton gradient-based optimizer. The partial derivatives of Equation (4) are computed for each parameter of q and then passed on to the optimizer that updates them such that Equation (4) is maximized.

[1]It can be seen as a multi-class extension of quadratic cost criterion (Bengio et al., 2006) or as a variant of the objective in (Zhu et al., 2003). An entropic distance measure could also be used, e.g., KL-divergence (Subramanya et al., 2010; Das and Smith, 2012).

4.3 Semi-Supervised CRFs Training

The trigram-level label distributions inferred in the propagation step can be viewed as a kind of valuable "prior knowledge" to regularize the learning on unlabeled data. The final step of the induction is thus to incorporate such prior knowledge into CRFs. Li (2009) generalizes the use of virtual evidence to undirected graphical models and, in particular, to CRFs for incorporating external knowledge. By extending the similar intuition, as illustrated in Figure 2, we modify the structure of a regular linear-chain CRFs on unlabeled data for smoothing the derived label distributions, where virtual evidences, i.e., q in our case, are donated by $\{v_1, v_2, \ldots, v_T\}$, in parallel with the state variables $\{y_1, y_2, \ldots, y_T\}$. The modified CRFs model allows us to flexibly define the interaction between estimated state values and virtual evidences by potential functions. Therefore, given labeled and unlabeled data, the learning objective is defined as follows:

$$\mathcal{L}(\Lambda) + \sum_{i=l+1}^{l+u} E_{p(y_i|x_i,v_i;\Lambda^g)}[\log p(y_i, v_i|x_i;\Lambda)] \tag{5}$$

where the conditional probability in the second term is denoted as

$$\begin{aligned}&p(y_i, v_i|x_i;\Lambda) = \\ &\frac{1}{Z'(x_i;\Lambda)} \exp\{\sum_{j=1}^{N}\sum_{k=1}^{K} \lambda_k f_k(y_i^{j-1}, y_i^j, x_i, j) \\ &+\alpha \sum_{t=1}^{N} s(y_i^t, v_i^t)\}\end{aligned} \tag{6}$$

The first term in Equation (5) is the same as Equation (2), which is the traditional CRFs learning objective function on the labeled data. The second term is the expected conditional likelihood of unlabeled data. It is directed to maximize the conditional likelihood of hidden states with the derived label distributions on unlabeled data, i.e., $p(y, v|x)$, where y and v are jointly modeled but

the probability is still conditional on x. Here, $Z'(x;\Lambda)$ is the partition function of normalization that is achieved by summing the numerator over both y and v. A virtual evidence feature function of $s(y_i^t, v_i^t)$ with pre-defined weight α is defined to regularize the conditional distributions of states over the derived label distributions. The learning is impacted by the derived label distributions as Equation (7): firstly, if the trigram $x_i^{t-1}x_i^t x_i^{t+1}$ at current position does have no corresponding derived label distributions ($v_i^t = null$), the value of zero is assigned to all state hypotheses so that the posteriors would not affected by the derived information. Secondly, if it does have a derived label distribution, since the virtual evidence in this case is a distribution instead of a specific label, the label probability in the distribution under the current state hypothesis is assigned. This means that the values of state variables are constrained to agree with the derived distributions.

$$s(y_i^t, v_i^t) = \begin{cases} q_{x_i^{t-1}x_i^t x_i^{t+1}}(y_i^t) & \text{if } v_i^t \neq null \\ 0 & \text{else} \end{cases} \tag{7}$$

The second term in Equation (5) can be optimized by using the expectation maximization (EM) algorithm in the same fashion as in the generative approach, following (Li, 2009). One can iteratively optimize the Q function $Q(\Lambda) = \sum_y p(y_i|x_i;\Lambda^g)\log p(y_i, v_i|x_i;\Lambda)$, in which Λ^g is the model estimated from the previous iteration. Here the gradient of the Q function can be measured by:

$$\frac{\partial Q(\Lambda)}{\partial \Lambda_k} = \sum_t \sum_{y_i^{t-1}, y_i^t} f_k(y_i^{t-1}, y_i^t, x_i, t). \\ (p(y_i^{t-1}, y_i^t|x_i, v_i;\Lambda) - p(y_i^{t-1}, y_i^t|x_i;\Lambda)) \tag{8}$$

The forward-backward algorithm is used to measure $p(y_i^{t-1}, y_i^t|x_i, v_i;\Lambda)$ and $p(y_i^{t-1}, y_i^t|x_i;\Lambda)$. Thus, the objective function Equation (5) is optimized as follows: for the instances $i = 1, 2, ..., l$, the parameters Λ are learned as the supervised manner; for the instances $i = l+1, l+2, ..., u+l$, in the E-step, the expected value of Q function is computed, based on the current model Λ^g. In the M-step, the posteriors are fixed and updated Λ that maximizes Equation (5).

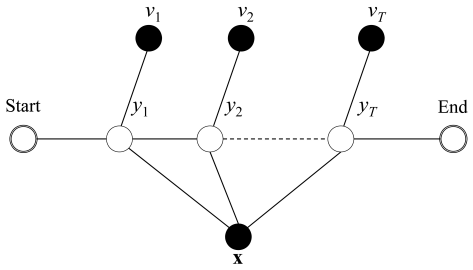

Figure 2: Modified linear-chain CRFs integrating virtual evidences on unlabeled data.

5 Experiment

5.1 Setting

The experimental data are mainly taken from the Chinese tree bank (CTB-7) and Microsoft Research (MSR)[2]. CTB-7 consists of over one million words of annotated and parsed text from Chinese newswire, magazine news, various broadcast news and broadcast conversation programs, web newsgroups and weblogs. It is a segmented, POS tagged[3] and fully bracketed corpus. The train, development and test sets[4] from CTB-7 and their corresponding statistics are reported in Table 3. To satisfy the characteristic of the semi-supervised learning problem, the train set, i.e., the labeled data, is formed by a relatively small amount of annotated texts sampled from CTB-7. For the unlabeled data in this experiment, a greater amount of texts is extracted from CTB-7 and MSR, which contains 53,108 sentences with 2,418,690 characters.

The performance measurement indicators for word segmentation and POS tagging (joint S&T) are balance F-score, F = 2PR/(P+R), the harmonic mean of precision (P) and recall (R), and out-of-vocabulary recall (OOV-R). For segmentation, a token is regarded to be correct if its boundaries match the ones of a word in the gold standard. For the POS tagging, it is correct only if both the boundaries and the POS tags are perfect matches.

The experimental platform is implemented based on two toolkits: Mallet (McCallum and Kachites, 2002) and Junto (Talukdar and Pereira, 2010). Mallet is a java-based package for statistical natural language processing, which includes the CRFs implementation. Junto is a graph-

[2]It can be download at: www.sighan.org/bakeoff2005.

[3]There is a total of 33 POS tags in CTB-7.

[4]The extracted sentences in train, development and test set were assigned with the composite tags as described in Section 3.1.

based label propagation toolkit that provides several state-of-the-art algorithms.

Data	#Sent	#Word	#Char	#OOV
Train	17,968	374,697	596,360	
Develop	1,659	46,637	79,283	0.074
Test	2,037	65,219	104,502	0.089

Table 3: Training, development and testing data.

5.2 Baseline and Proposed Models

In the experiment, the baseline supervised pipeline and joint S&T models are built only on the train data. The proposed model will also be compared with the semi-supervised pipeline S&T model described in (Wang et al., 2011). In addition, two state-of-the-art semi-supervised CRFs algorithms, Jiao's CRFs (Jiao et al., 2006) and Subramanya's CRFs (Subramanya et al., 2010), are also used to build joint S&T models. The corresponding settings of the above candidates are listed below:

- **Baseline I**: a supervised CRFs pipeline S&T model. The feature templates are from Zhao et al. (2006) and Wu et al. (2008).
- **Wang's model**: a semi-supervised CRFs pipeline S&T model. The same feature templates in (Wang et al., 2011) are used, i.e., "+*n*-gram+cluster+lexicon".
- **Baseline II**: a supervised CRFs joint S&T model. The feature templates introduced in Section 3.1 are used.
- **Jiao's model**: a semi-supervised CRFs joint S&T model trained using the entropy regularization (ER) criteria (Jiao et al., 2006). The optimization method proposed by Mann and McCallum (2007) is applied.
- **Subramanya's model**: a self-train style semi-supervised CRFs joint S&T model based on the same parameters used in (Subramanya et al., 2010).
- **Our model**: several parameters in our model are needed to tune based on the development set, e.g., μ, λ and α.

In all the CRFs models above, the Gaussian regularizer and stochastic gradient descent method are employed.

5.3 Main Results

This experiment yielded a similarity graph that consists of 462,962 trigrams from labeled and unlabeled data. The majority (317,677 trigrams) occurred only in unlabeled data. Based on the development data, the hyperparameters of our model were tuned among the following settings: for the graph propagation, $\mu \in \{0.2, 0.5, 0.8\}$ and $\lambda \in \{0.1, 0.3, 0.5, 0.8\}$; for the CRFs training, $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The best performed joint settings are $\mu = 0.5$, $\lambda = 0.3$ and $\alpha = 0.7$. With the chosen set of hyperparameters, the test data was used to measure the final performance.

Model	Segmentation		POS Tagging	
	F_1	OOV-R	F_1	OOV-R
Baseline I	94.27	60.12	91.08	51.72
Wang's	95.17	63.10	91.64	53.29
Baseline II	95.14	61.52	91.61	52.29
Jiao's	95.58	63.05	92.11	53.27
Subramanya's	96.30	67.12	92.46	57.15
Our model	**96.85**	**68.09**	**92.89**	**58.36**

Table 4: The performance of segmentation and POS tagging on testing data.

Table 4 summarizes the performance of segmentation and POS tagging on the test data, in comparison with the other five models. Firstly, as expected, for the two supervised baselines, the joint model outperforms the pipeline one, especially on segmentation. It obtains 0.92% and 2.32% increase in terms of F-score and OOV-R respectively. This outcome verifies the commonly accepted fact that the joint model can substantially improve the pipeline one, since POS tags provide additional information to word segmentation (Ng and Low, 2004). Secondly, it is also noticed that all four semi-supervised models are able to benefit from unlabeled data and greatly improve the results with respect to the baselines. On the whole, for segmentation, they achieve average improvements of 1.02% and 6.8% in F-score and OOV-R; whereas for POS tagging, the average increments of F-sore and OOV-R are 0.87% and 6.45%. An interesting phenomenon is found among the comparisons with baselines that the supervised joint model (Baseline II) is even competitive with semi-supervised pipeline one (Wang et al., 2011). This illustrates the effects of error propagation in the pipeline approach. Thirdly, in what concerns the semi-supervised approaches, the three joint S&T models, i.e., Jiao's, Subramanya's and our model, are superior to the pipeline model, i.e., Wang's

model. Moreover, the two graph-based approaches, i.e., Subramanya's and our model, outperform the others. Most importantly, the boldface numbers in the last row illustrate that our model does achieve the best performance. Overall, for word segmentation, it obtains average improvements of 1.43% and 8.09% in F-score and OOV-R over others; for POS tagging, it achieves average improvements of 1.09% and 7.73%.

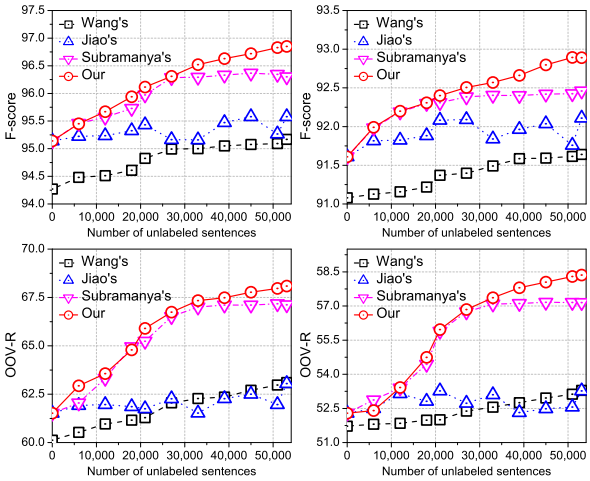

Figure 3: The learning curves of semi-supervised models on unlabeled data, where left graphs are segmentation and the right ones are tagging.

5.4 Learning Curve

An additional experiment was conducted to investigate the impact of unlabeled data for the four semi-supervised models. Figure 3 illustrates the curves of F-score and OOV-R for segmentation and tagging respectively, as the unlabeled data size is progressively increased in steps of 6,000 sentences. It can be clearly observed that all curves of our model are able to mount up steadily and achieve better gains over others consistently. The most competitive performance of the other three candidates is achieved by Subramanya's model. This strongly reveals that the knowledge derived from the similarity graph does effectively strengthen the model. But in Subramanya's model, when the unlabeled size ascends to approximately 30,000 sentences the curves become nearly asymptotic. The semi-supervised pipeline model, Wang's model, presents a much slower growth on all curves over the others and also begins to overfit with large unlabeled data sizes (>25,000 sentences). The figure also shows an erratic fluctuation of Jiao's model. Since this approach aims at minimizing conditional entropy over unlabeled data and encourages finding putative labelings for unlabeled data, it results in a data-sensitive model (Li et al., 2009).

5.5 Analysis & Discussion

A statistical analysis of the segmentation and tagging results of the supervised joint model (Baseline II) and our model is carried out to comprehend the influence of the graph-based semi-supervised behavior. For word segmentation, the most significant improvement of our model is mainly concentrated on two kinds of words which are known for their difficulties in terms of CWS: a) named entities (NE), e.g., "天津港" (Tianjin port) and "保税区" (free tax zone); and b) Chinese numbers (CN), e.g., "八点五亿" (eight hundred and fifty million) and "百分之七十二" (seventy two percent). Very often, these words do not exist in the labeled data, so the supervised model is hard to learn their features. Part of these words, however, may occur in the unlabeled data. The proposed semi-supervised approach is able to discover their label information with the help of a similarity graph. Specifically, it learns the label distributions from similar words (neighborhoods), e.g., "上海港" (Shanghai port), "保护区" (protection zone), "九点七亿" (nine hundred and seventy million). The statistics in Table 5 demonstrate significant error reductions of 50.44% and 48.74% on test data, corresponding to NE and CN respectively.

Type	#word	#baErr	#gbErr	ErrDec%
NE	471	226	112	50.44
CN	181	119	61	48.74

Table 5: The statistics of segmentation error for named entities (NE) and Chinese numbers (CN) in test data. **#baErr** and **#gbErr** denote the count of segmentations by Baseline II and our model; **ErrDec%** denotes the error reduction.

On the other hand, to better understand the tagging results, we summarize the increase and decrease of the top five common tagging error patterns of our model over Baseline II for the correctly segmented words, as shown in Table 6. The error pattern is defined by "A→B" that refers the true tag of "A" is annotated by a tag of "B". The obvious improvement brought by our model occurs with the tags "NN", "CD", "NR", "JJ" and "NR", where errors are reduced 60.74% on aver-

Pattern	#baErr	↓	Pattern	#baErr	↑
NN→VV	58	38	NN→NR	13	6
CD→NN	41	27	IJ→ON	9	5
NR→VV	29	17	VV→NN	4	3
JJ→NN	18	11	NR→NN	1	3
NR→VA	19	10	JJ→AD	1	2

Table 6: The statistics of POS tagging error patterns in test data. **#baErr** denote the count of tagging error by Baseline II, while ↓ and ↑ denotes the number of error reduced or increased by our model.

age. More impressively, there is a large portion of fixed error pattern instances stemming from OOV words. Meanwhile, it is also observed that the disambiguation of error patterns in the right portion of the table slightly suffers from our approach. In reality, it is impossible and unrealistic to request a model to be "no harms but only benefits" under whatever circumstances.

6 Conclusion

This study introduces a novel semi-supervised approach for joint Chinese word segmentation and POS tagging. The approach performs the semi-supervised learning in the way that the trigram-level distributions inferred from a similarity graph are used to regularize the learning of CRFs model on labeled and unlabeled data. The empirical results indicate that the similarity graph information and the incorporation manner of virtual evidences present a positive effect to the model induction.

Acknowledgments

The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau for the funding support for our research, under the reference No. 017/2009/A and RG060/09-10S/CS/FST. The authors also wish to thank the anonymous reviewers for many helpful comments.

References

Shumeet Baluja, Rohan Seth, D. Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravich, and Mohamed Aly. 2008. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of WWW*, pages 895-904, Beijing, China.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. 2006. Manifold regularization. *Journal of machine learning research*, 7:2399–2434.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 2006. Label propogation and quadratic criterion. *MIT Press*.

Jon Louis Bentley. 1980. Multidimensional divide-and -conquer. *Communications of the ACM*, 23(4):214 - 229.

Alina Beygelzimer, Sham Kakade, and John Langford. 2006. Cover trees for nearest neighbor. In *Proceedings of ICML*, pages 97-104, New York, USA

Olivier Chapelle, Bernhard Schö lkopf, and Alexander Zien. 2006. Semi-supervised learning. *MIT Press*.

Samuel I. Daitch, Jonathan A. Kelner, and Daniel A. Spielman. 2009. Fitting a graph to vector data. In *Proceedings of ICML*, 201-208, NY, USA.

Dipanjan Das and Noah A. Smith. 2011. Semi-supervised framesemantic parsing for unknown predicates. In *Proceedings of ACL*, pages 1435-1444, Portland, Oregon, USA.

Dipanjan Das and Slav Petrov. 2011. Unsupervised Part-of-Speech Tagging with Bilingual Graph-based Projections. In *Proceedings of ACL*, pages 1435-1444, Portland, Oregon, USA.

Dipanjan Das and Noah A. Smith. 2012. Graph-based lexicon expansion with sparsity-inducing penalties. In *Proceedings of NAACL*, pages 677-687, Montréal, Canada.

Tony Jebara, Jun Wang, and Shih-Fu Chang. 2009. Graph construction and b-matching for semi-supervised learning. In *Proceedings of ICML*, 441-448, New York, USA.

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

Wenbin Jiang, Liang Huang, and Qun Liu. 2009. Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging – A Case Study. In *Proceedings of he ACL and the 4th IJCNLP of the AFNLP*, pages 522–530, Suntec, Singapore.

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

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging. In *Proceedings of ACL and IJCNLP of the AFNLP*, pages 513- 521, Suntec, Singapore August.

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

Xiao Li. 2009. On the use of virtual evidence in conditional random fields. In *Proceedings of EMNLP*, pages 1289-1297, Singapore.

Xiao Li, Ye-Yi Wang, and Alex Acero. 2009. Extracting structured information from user queries with semi-supervised conditional random fields In *Proceedings of ACM SIGIR*, pages 572-579, Boston, USA.

Gideon S. Mann and Andrew McCallum. 2007. Efficient computation of entropy gradient for semi-supervised conditional random fields. In *Proceedings of NAACL*, pages 109-112, New York, USA.

McCallum and Andrew Kachites. 2002. MALLET: A Machine Learning for Language Toolkit. Software at *http://mallet.cs.umass.edu*.

Tetsuji Nakagawa and Kiyotaka Uchimoto. 2007. A hybrid approach to word segmentation and POS tagging. In *Proceedings of ACL Demo and Poster Session*, pages 217–220, Prague, Czech Republic.

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

Xian Qian and Yang Liu. 2012. Joint Chinese Word Segmentation, POS Tagging and Parsing. In *Proceedings of EMNLP-CoNLL*, pages 501-511, Jeju Island, Korea.

Yanxin Shi and Mengqiu Wang. 2007. A dual-layer CRF based joint decoding method for cascade segmentation and labelling tasks. In *Proceedings of IJCAI*, Hyderabad, India.

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, Massachusetts, USA.

Weiwei Sun. 2011. A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *Proceedings of ACL*, pages 1385–1394, Portland, Oregon.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese word segmentation using unlabeled data. In *Proceedings of EMNLP*, pages 970-979, Scotland, UK.

Partha Pratim Talukdar, Joseph Reisinger, Marius Pasca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. 2008. Weakly Supervised Acquisition of Labeled Class Instances using Graph Random Walks. In *Proceedings of EMNLP*, pages 582-590, Hawaii, USA.

Partha Pratim Talukdar and Koby Crammer. 2009. New Regularized Algorithms for Transductive Learning. In *Proceedings of ECML-PKDD*, pages 442 - 457, Bled, Slovenia.

Partha Pratim Talukdar and Fernando Pereira. 2010. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of ACL*, pages 1473-1481, Uppsala, Sweden.

Yiou Wang, Jun'ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Kentaro Torisawa. 2011. Improving Chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of IJCNLP*, pages 309–317, Chiang Mai, Thailand.

Yu-Chieh Wu Jie-Chi Yang, and Yue-Shi Lee. 2008. Description of the NCU Chinese Word Segmentation and Part-of-Speech Tagging for SIGHAN Bakeoff. In *Proceedings of the SIGHAN Workshop on Chinese Language Processing*, pages 161-166, Hyderabad, India.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised chinese word segmentation for statistical machine translation. In *Proceedings of COLING*, pages 1017-1024, Manchester, UK.

Yue Zhang and Stephen Clark. 2008. Joint word segmentation and POS tagging using a single perceptron. In *Proceedings of EMNLP*, pages 888-896, Columbus, Ohio.

Yue Zhang and Stephen Clark. 2010. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. In *Proceedings of EMNLP*, pages 843-852, Massachusetts, USA.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006. Effective tag set selection in Chinese word segmentation via conditional random field modeling. In *Proceedings of PACLIC*, pages 87-94, Wuhan, China.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of ICML*, pages 912–919, Washington DC, USA.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. 1997. L-BFGS-B: Fortran subroutines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23:550-560.