Co-regularizing character-based and word-based models for semi-supervised Chinese word segmentation

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 presents a semi-supervised Chinese word segmentation (CWS) approach that co-regularizes character-based and word-based models. Similarly to multi-view learning, the "segmentation agreements" between the two different types of view are used to overcome the scarcity of the label information on unlabeled data. The proposed approach trains a character-based and word-based model on labeled data, respectively, as the initial models. Then, the two models are constantly updated using unlabeled examples, where the learning objective is maximizing their segmentation agreements. The agreements are regarded as a set of valuable constraints for regularizing the learning of both models on unlabeled data. The segmentation for an input sentence is decoded by using a joint scoring function combining the two induced models. The evaluation on the Chinese tree bank reveals that our model results in better gains over the state-of-the-art semi-supervised models reported in the literature.

1 Introduction

Chinese word segmentation (CWS) is a critical and a necessary initial procedure with respect to the majority of high-level Chinese language processing tasks such as syntax parsing, information extraction and machine translation, since Chinese scripts are written in continuous characters without explicit word boundaries. Although supervised CWS models (Xue, 2003; Zhao et al., 2006; Zhang and Clark, 2007; Sun, 2011) proposed in the past years showed some reasonably accurate results, the outstanding problem is that they rely heavily on a large amount of labeled data. However, the production of segmented Chinese texts is time-consuming and expensive, since hand-labeling individual words and word boundaries is very hard (Jiao et al., 2006). So, one cannot rely only on the manually segmented data to build an everlasting model. This naturally provides motivation for using easily accessible raw texts to enhance supervised CWS models, in semi-supervised approaches. In the past years, however, few semi-supervised CWS models have been proposed. Xu et al. (2008) described a Bayesian semi-supervised model by considering the segmentation as the hidden variable in machine translation. Sun and Xu (2011) enhanced the segmentation results by interpolating the statistics-based features derived from unlabeled data to a CRFs model. Another similar trial via "feature engineering" was conducted by Wang et al. (2011).

The crux of solving semi-supervised learning problem is the learning on unlabeled data. Inspired by multi-view learning that exploits redundant views of the same input data (Ganchev et al., 2008), this paper proposes a semi-supervised CWS model of co-regularizing from two different views (intrinsically two different models), character-based and word-based, on unlabeled data. The motivation comes from that the two types of model exhibit different strengths and they are mutually complementary (Sun, 2010; Wang et al., 2010). The proposed approach begins by training a character-based and word-based model on labeled data respectively, and then both models are regularized from each view by their segmentation agreements, i.e., the identical outputs, of unlabeled data. This paper introduces segmentation agreements as gainful knowledge for guiding the learning on the texts without label information. Moreover, in order to better combine the strengths of the two models, the proposed approach uses a joint scoring function in a log-linear combination form for the decoding in the segmentation phase.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 171–176,
Sofia, Bulgaria, August 4-9 2013.

2 Segmentation Models

There are two classes of CWS models: character-based and word-based. This section briefly reviews two supervised models in these categories, a character-based CRFs model, and a word-based Perceptrons model, which are used in our approach.

2.1 Character-based CRFs Model

Character-based models treat word segmentation as a sequence labeling problem, assigning labels to the characters in a sentence indicating their positions in a word. A 4 tag-set is used in this paper: ***B*** (beginning), ***M*** (middle), ***E*** (end) and ***S*** (single character). Xue (2003) first proposed the use of CRFs model (Lafferty et al., 2001) in character-based CWS. Let $x = (x^1 x^2 ... x^{|x|}) \in \mathcal{X}$ denote a sentence, where each character and $y = (y^1 y^2 ... y^{|y|}) \in \mathcal{Y}$ denote a tag sequence, $y^i \in \mathcal{T}$ being the tag assigned to x^i. The goal is to achieve a label sequence with the best score in the form,

$$p_{\theta_c}(y|x) = \frac{1}{Z(x;\theta_c)} \exp\{f(x,y) \cdot \theta_c\} \quad (1)$$

where $Z(x;\theta_c)$ is a partition function that normalizes the exponential form to be a probability distribution, and $f(x,y)$ are arbitrary feature functions. The aim of CRFs is to estimate the weight parameters θ_c that *maximizes* the conditional likelihood of the training data:

$$\hat{\theta_c} = \underset{\theta_c}{\operatorname{argmax}} \sum_{i=1}^{l} \log p_{\theta_c}(y^i|x^i) - \gamma \|\theta_c\|_2^2 \quad (2)$$

where $\gamma\|\theta_c\|_2^2$ is a regularizer on parameters to limit overfitting on rare features and avoid degeneracy in the case of correlated features. In this paper, this objective function is optimized by stochastic gradient method. For the decoding, the Viterbi algorithm is employed.

2.2 Word-based Perceptrons Model

Word-based models read a input sentence from left to right and predict whether the current piece of continuous characters is a word. After one word is identified, the method moves on and searches for a next possible word. Zhang and Clark (2007) first proposed a word-based segmentation model using a discriminative Perceptrons algorithm. Given a sentence x, let us denote a possible segmented sentence as $w \in \mathbf{w}$, and the function that enumerates a set of segmentation candidates as GEN:$\mathbf{w} = \mathrm{GEN}(x)$ for x. The objective is to *maximize* the following problem for all sentences:

$$\hat{\theta}_w = \underset{\mathbf{w}=\mathrm{GEN}(x)}{\operatorname{argmax}} \sum_{i=1}^{|\mathbf{w}|} \phi(x, w_i) \cdot \theta_w \quad (3)$$

where it maps the segmented sentence w to a global feature vector ϕ and denotes θ_w as its corresponding weight parameters. The parameters θ_w can be estimated by using the Perceptrons method (Collins, 2002) or other online learning algorithms, e.g., Passive Aggressive (Crammer et al., 2006). For the decoding, a beam search decoding method (Zhang and Clark, 2007) is used.

2.3 Comparison Between Both Models

Character-based and word-based models present different behaviors and each one has its own strengths and weakness. Sun (2010) carried out a thorough survey that includes theoretical and empirical comparisons from four aspects. Here, two critical properties of the two models supporting the co-regularization in this study are highlighted. Character-based models present better prediction ability for new words, since they lay more emphasis on the internal structure of a word and thereby express more nonlinearity. On the other side, it is easier to define the word-level features in word-based models. Hence, these models have a greater representational power and consequently better recognition performance for in-of-vocabulary (IV) words.

3 Semi-supervised Learning via Co-regularizing Both Models

As mentioned earlier, the primary challenge of semi-supervised CWS concentrates on the unlabeled data. Obviously, the learning on unlabeled data does not come for "free". Very often, it is necessary to discover certain gainful information, e.g., label constraints of unlabeled data, that is incorporated to guide the learner toward a desired solution. In our approach, we believe that the segmentation agreements (§ 3.1) from two different views, character-based and word-based models, can be such gainful information. Since each of the models has its own merits, their consensuses signify high confidence segmentations. This naturally leads to a new learning objective that maximizes segmentation agreements between two models on unlabeled data.

This study proposes a co-regularized CWS model based on character-based and word-based models, built on a small amount of segmented sentences (labeled data) and a large amount of raw sentences (unlabeled data). The model induction process is described in Algorithm 1: given labeled dataset D_l and unlabeled dataset D_u, the first two steps are training a CRFs (character-based) and Perceptrons (word-based) model on the labeled data D_l , respectively. Then, the parameters of both models are continually updated using unlabeled examples in a learning cycle. At each iteration, the raw sentences in D_u are segmented by current character-based model θ_c and word-based model θ_w. Meanwhile, all the segmentation agreements $\mathcal{A}$ are collected (§ 3.1). Afterwards, the agreements $\mathcal{A}$ are used as a set of constraints to bias the learning of CRFs (§ 3.2) and Perceptron (§ 3.3) on the unlabeled data. The convergence criterion is the occurrence of a reduction of segmentation agreements or reaching the maximum number of learning iterations. In the final segmentation phase, given a raw sentence, the decoding requires both induced models (§ 3.4) in measuring a segmentation score.

Algorithm 1 Co-regularized CWS model induction

Require: n labeled sentences $\mathcal{D}_l$; m unlabeled sentences $\mathcal{D}_u$
Ensure: θ_c and θ_w
1: $\theta_c^0 \leftarrow$ **crf_train**$(\mathcal{D}_l)$
2: $\theta_w^0 \leftarrow$ **perceptron_train**$(\mathcal{D}_l)$
3: for $t = 1...T_{max}$ do
4: $\quad \mathcal{A}^t \leftarrow$ **agree**$(\mathcal{D}_u, \theta_c^{t-1}, \theta_w^{t-1})$
5: $\quad \theta_c^t \leftarrow$ **crf_train_constraints**$(\mathcal{D}_u, \mathcal{A}^t, \theta_c^{t-1})$
6: $\quad \theta_w^t \leftarrow$ **perceptron_train_constraints**$(\mathcal{D}_u, \mathcal{A}^t, \theta_w^{t-1})$
7: end for

3.1 Agreements Between Two Models

Given a raw sentence, e.g., "我正在北京看奥运会开幕式。(I am watching the opening ceremony of the Olympics in Beijing.)", the two segmentations shown in Figure 1 are the predictions from a character-based and word-based model. The segmentation agreements between the two models correspond to the identical words. In this example, the five words, i.e. "我 (I)", "北京 (Beijing)", "看 (watch)", "开幕式 (opening ceremony)" and "。(.)", are the agreements.

3.2 CRFs with Constraints

For the character-based model, this paper follows (Täckström et al., 2013) to incorporate the segmentation agreements into CRFs. The main idea is to constrain the size of the tag sequence lattice according to the agreements for achieving simplified learning. Figure 2 demonstrates an example of the constrained lattice, where the bold node represents that a definitive tag derived from the agreements is assigned to the current character, e.g., "我 (I)" has only one possible tag "S" because both models segmented it to a word with a single character. Here, if the lattice of all admissible tag sequences for the sentence x is denoted as $\mathcal{Y}(x)$, the constrained lattice can be defined by $\hat{\mathcal{Y}}(x, \tilde{y})$, where $\tilde{y}$ refers to tags inferred from the agreements. Thus, the objective function on unlabeled data is modeled as:

$$\hat{\theta}_c' = \underset{\theta_c}{\operatorname{argmax}} \sum_{i=1}^{m} \log p_{\theta_c}(\hat{\mathcal{Y}}(x^i, \tilde{y}^i)|x^i) - \gamma\|\theta_c\|_2^2 \quad (4)$$

It is a marginal conditional probability given by the total probability of all tag sequences consistent with the constrained lattice $\hat{\mathcal{Y}}(x, \tilde{y})$. This objective can be optimized by using LBFGS-B (Zhu et al., 1997), a generic quasi-Newton gradient-based optimizer.

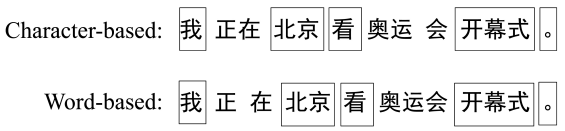

Figure 1: The segmentations given by character-based and word-based model, where the words in "□" refer to the segmentation agreements.

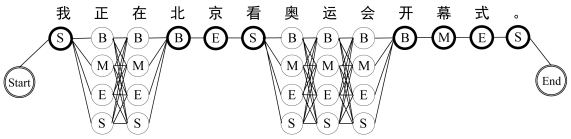

Figure 2: The constrained lattice representation for a given sentence, "我正在北京看奥运会开幕式。".

3.3 Perceptrons with Constraints

For the word-based model, this study incorporates segmentation agreements by a modified parameter update criterion in Perceptrons online training, as shown in Algorithm 2. Because there are no "gold segmentations" for unlabeled sentences, the output sentence predicted by the current model is compared with the agreements instead of the "answers" in the supervised case. At each parameter

update iteration k, each raw sentence x_u is decoded with the current model into a segmentation z_u. If the words in output z_u do not match the agreements $\mathcal{A}(x_u)$ of the current sentence x_u, the parameters are updated by adding the global feature vector of the current training example with the agreements and subtracting the global feature vector of the decoder output, as described in lines 3 and 4 of Algorithm 2.

Algorithm 2 Parameter update in word-based model

1: for $k = 1...K, u = 1...m$ do
2: calculate $z_u = \underset{\mathbf{w}=\mathrm{GEN}(x)}{\operatorname{argmax}} \sum_{i=1}^{|\mathbf{w}|} \phi(x_u, w_i) \cdot \theta_w^{k-1}$
3: if $z_u \neq \mathcal{A}(x_u)$
4: $\theta_w^k = \theta_w^{k-1} + \phi(\mathcal{A}(x_u)) - \phi(z_u)$
5: end for

3.4 The Joint Score Function for Decoding

There are two co-regularized models as results of the previous induction steps. An intuitive idea is that both induced models are combined to conduct the segmentation, for the sake of integrating their strengths. This paper employs a log-linear interpolation combination (Bishop, 2006) to formulate a joint scoring function based on character-based and word-based models in the decoding:

$$Score(w) = \alpha \cdot \log(p_{\theta_c}(y|x)) + (1-\alpha) \cdot \log(\phi(x, w) \cdot \theta_w) \quad (5)$$

where the two terms of the logarithm are the scores of character-based and word-based models, respectively, for a given segmentation w. This composite function uses a parameter α to weight the contributions of the two models. The α value is tuned using the development data.

4 Experiment

4.1 Setting

The experimental data is taken from the Chinese tree bank (CTB). In order to make a fair comparison with the state-of-the-art results, the versions of CTB-5, CTB-6, and CTB-7 are used for the evaluation. The training, development and testing sets are defined according to the previous works. For CTB-5, the data split from (Jiang et al., 2008) is employed. For CTB-6, the same data split as recommended in the CTB-6 official document is used. For CTB-7, the datasets are formed according to the way in (Wang et al., 2011). The corresponding statistic information on these data splits is reported in Table 1. The unlabeled data in our experiments is from the XIN_CMN portion of Chinese Gigaword 2.0. The articles published in 1991-1993 and 1999-2004 are used as unlabeled data, with 204 million words.

The feature templates in (Zhao et al., 2006) and (Zhang and Clark, 2007) are used in train-ing the CRFs model and Perceptrons model, respectively. The experimental platform is implemented based on two popular toolkits: CRF++ (Kudo, 2005) and Zpar (Zhang and Clark, 2011).

Data	#Sent-train	#Sent-dev	#Sent-test	OOV-dev	OOV-test
CTB-5	18,089	350	348	0.0811	0.0347
CTB-6	23,420	2,079	2,796	0.0545	0.0557
CTB-7	31,131	10,136	10,180	0.0549	0.0521

Table 1: Statistics of CTB-5, CTB-6 and CTB-7 data.

4.2 Main Results

The development sets are mainly used to tune the values of the weight factor α in Equation 5. We evaluated the performance (F-score) of our model on the three development sets by using different α values, where α is progressively increased in steps of 0.1 ($0 < \alpha < 1.0$). The best performed settings of α for CTB-5, CTB-6 and CTB-7 on development data are 0.7, 0.6 and 0.6, respectively. With the chosen parameters, the test data is used to measure the final performance.

Table 2 shows the F-score results of word segmentation on CTB-5, CTB-6 and CTB-7 testing sets. The line of "ours" reports the performance of our semi-supervised model with the tuned parameters. We first compare it with the supervised "baseline" method which joints character-based and word-based model trained only on the training set[1]. It can be observed that our semi-supervised model is able to benefit from unlabeled data and greatly improves the results over the supervised baseline. We also compare our model with two state-of-the-art semi-supervised methods of Wang '11 (Wang et al., 2011) and Sun '11 (Sun and Xu, 2011). The performance scores of Wang '11 are directly taken from their paper, while the results of Sun '11 are obtained, using the program provided by the author, on the same experimental data. The

[1]The "baseline" uses a different training configuration so that the α values in the decoding are also need to be tuned on the development sets. The tuned α values are {0.6, 0.6, 0.5} for CTB-5, CTB-6 and CTB-7.

bold scores indicate that our model does achieve significant gains over these two semi-supervised models. This outcome can further reveal that using the agreements from these two views to regularize the learning can effectively guide the model toward a better solution. The third comparison candidate is Hatori '12 (Hatori et al., 2012) which reported the best performance in the literature on these three testing sets. It is a supervised joint model of word segmentation, POS tagging and dependency parsing. Impressively, our model still outperforms Hatori '12 on all three datasets. Although there is only a 0.01 increase on CTB-5, it can be seen as a significant improvement when considering Hatori '12 employs much richer training resources, i.e., sentences tagged with syntactic information.

Method	CTB-5	CTB-6	CTB-7
Ours	**98.27**	**96.33**	**96.72**
Baseline	97.58	94.71	94.87
Wang '11	98.11	95.79	95.65
Sun '11	98.04	95.44	95.34
Hatori '12	98.26	96.18	96.07

Table 2: F-score (%) results of five CWS models on CTB-5, CTB-6 and CTB-7.

5 Conclusion

This paper proposed an alternative semi-supervised CWS model that co-regularizes a character- and word-based model by using their segmentation agreements on unlabeled data. We perform the agreements as valuable knowledge for the regularization. The experiment results reveal that this learning mechanism results in a positive effect to the segmentation performance.

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 MYRG076(Y1-L2)-FST13-WF. The authors also wish to thank the anonymous reviewers for many helpful comments.

References

Christopher M. Bishop. 2006. *Pattern recognition and machine learning*.

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

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of ma-chine learning research*, 7:551-585.

Kuzman Ganchev, Joao Graca, John Blitzer, and Ben Taskar. 2008. Multi-View Learning over Structured and Non-Identical Outputs. In *Proceedings of CUAI*, pages 204-211, Helsinki, Finland.

Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2012. Incremental Joint Approach to Word Segmentation, POS Tagging, and Dependency Parsing in Chinese. In *Proceedings of ACL*, pages 1045-1053, Jeju, Republic of Korea.

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 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 *Proceedings of ACL and the 4th IJCNLP of the AFNLP*, pages 209-216, Strouds-burg, PA, USA.

Taku Kudo. 2005. CRF++: Yet another CRF toolkit. Software available at *http://crfpp.sourceforge. net*.

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.

Weiwei Sun. 2001. Word-based and character-based word segmentation models: comparison and combination. In *Proceedings of COLING*, pages 1211-1219, Bejing, China.

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.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. In *Transactions of the Association for Computational Linguistics*, 1:1-12.

Kun Wang, Chengqing Zong, and Keh-Yih Su. 2010. A Character-Based Joint Model for Chinese Word Segmentation. In *Proceedings of COLING*, pages 1173-1181, Bejing, China.

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, 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.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29-48.

Yue Zhang and Stephen Clark. 2007. Chinese segmentation using a word-based perceptron algorithm. In *Proceedings of ACL*, pages 840-847, Prague, Czech Republic.

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.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105-151.

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.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. 2006. L-BFGS-B: Fortran subroutines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23:550-560.