Exploiting Heterogeneous Annotations for Weibo Word Segmentation and POS Tagging

Jiayuan Chao, Zhenghua Li(✉), Wenliang Chen, and Min Zhang

School of Computer Science and Technology, Soochow University, Suzhou 215006, Jiangsu, China
chaojiayuan.china@gmail.com, {zhli13,wlchen,minzhang}@suda.edu.cn

Abstract. This paper describes our system designed for the NLPCC 2015 shared task on Chinese word segmentation (WS) and POS tagging for Weibo Text. We treat WS and POS tagging as two separate tasks and use a cascaded approach. Our major focus is how to effectively exploit multiple heterogeneous data to boost performance of statistical models. This work considers three sets of heterogeneous data, i.e., Weibo (*WB*, 10K sentences), Penn Chinese Treebank 7.0 (*CTB7*, 50K), and People's Daily (*PD*, 280K). For WS, we adopt the recently proposed coupled sequence labeling to combine *WB*, *CTB7*, and *PD*, boosting F1 score from 93.76% (baseline model trained on only *WB*) to 95.58% (+1.82%). For POS tagging, we adopt an ensemble approach combining coupled sequence labeling and the guide-feature based method, since the three datasets have three different annotation standards. First, we convert *PD* into the annotation style of *CTB7* based on coupled sequence labeling, denoted by PD^{CTB}. Then, we merge *CTB7* and PD^{CTB} to train a POS tagger, denoted by $Tag_{CTB7+PD^{CTB}}$, which is further used to produce guide features on *WB*. Finally, the tagging F1 score is improved from 87.93% to 88.99% (+1.06%).

1 Introduction

Chinese word segmentation (WS) and part-of-speech (POS) tagging are two fundamental tasks in Chinese language processing. In past decades, supervised approaches have gained extensive progress on canonical texts, especially on texts from domains or genres similar to existing manually labeled data[1]. However, the upsurge of web data imposes great challenges on existing techniques. The performance of the state-of-the-art systems degrades dramatically on informal web texts, such as micro-blogs, product comments, and so on. Driven by this challenge, NLPCC 2015 organizes a shared task with an aim of promoting WS and POS tagging on Weibo (*WB*, Chinese pinyin of micro-blogs) text [10].

This work was supported by National Natural Science Foundation of China (Grant No.61432013, 61273319) and Jiangsu Planned Projects for Post-doctoral Research Funds (No.1401075B).

[1] Please refer to http://zhangkaixu.github.io/bibpage/cws.html for a long list of related papers.

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 495–506, 2015.
DOI: 10.1007/978-3-319-25207-0_46

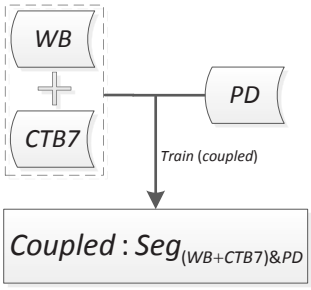

Fig. 1. Our final system for the WS subtask

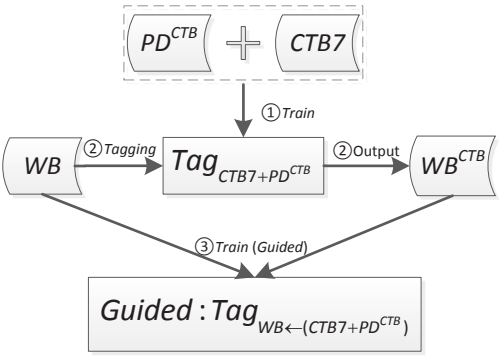

Fig. 2. Our final system for the POS tagging subtask

This paper describes our system designed for the shared task in detail. We treat WS and POS tagging as two separate tasks and use a cascaded approach. We treat both WS and POS tagging as sequence labeling problems, and build our model based on the standard conditional random field (CRF) [5] with bigram features. Our major focus is how to effectively exploit multiple heterogeneous data to boost performance of WS and POS tagging on *WB* (10*K* sentences). The third-party heterogeneous resources used in the work are Penn Chinese Treebank 7.0 (*CTB7*, 50*K*) and People's Daily (*PD*, 280*K*).

Figure 1 illustrates our final system for the WS subtask. We adopt the recently proposed coupled sequence labeling to combine *WB*, *CTB7*, and *PD* [6]. Strictly speaking, the three datasets have three different annotation standards. However, we empirically find that the annotation standard of *WB* and *CTB7* are very close while *PD* is different from both of them. Therefore, in our final system, we assume that *WB* and *CTB7* have the same annotation standard while *PD* follows a different annotation standard. Then, we train a coupled sequence labeling model on the three datasets with two sets of WS labels. The baseline model trained on the single *WB* achieve an F1 score of 93.76%, whereas our best model has an F1 score of 95.58%, which is an absolute improvement of 1.82%. Our WS system ranks the third place among all 17 systems (5 participating systems in the open track, 4 semi-open, 8 closed).

Figure 2 illustrates our final system for the POS tagging subtask. The three datasets have three different annotation standards. Meanwhile, the coupled sequence labeling model has an inefficient problem due to the large bundled POS tag set. Therefore, due to time limitation for the shared task, we adopt an ensemble approach combining coupled sequence labeling and the guide-feature based method [1]. First, we apply the coupled model described in Li et al. (2015) [6] to convert *PD* into the annotation style of *CTB7*, denoted by PD^{CTB}. Then, we merge $CTB7$ and PD^{CTB} to train a POS tagger, denoted by $Tag_{CTB7+PD^{CTB}}$, which is further used to produce guide features on *WB*. Finally, we train a POS

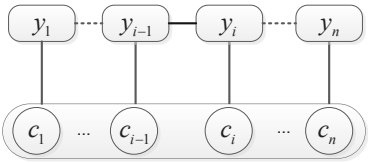

Fig. 3. Graphical structure of the traditional CRF

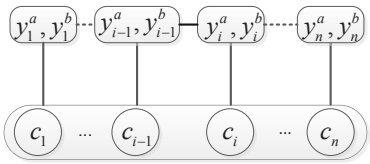

Fig. 4. Graphical structure of the coupled CRF

tagger on the WB with guide features. The tagging F1 score is improved from 87.93% (baseline model trained on only *WB*) to 88.99% (+1.06%). Our POS tagging system ranks the second place among all 12 participating systems (3 systems in the open track, 3 semi-open, 6 closed).

This paper is organized as follows. Since Li et al. (2015) [6] describe coupled sequence labeling for POS tagging, we introduce the traditional CRF model and the coupled CRF model for WS in Section 2 and Section 3 respectively. Section 4 presents experiments. We discuss closely related works in Section 5 and conclude this paper in Section 6.

2 Traditional CRF for WS

We adopt CRF for the WS subtask, which treats WS as a sequence labeling problem. We adopt the $\{B, M, E, S\}$ tag set, indicating the beginning of a word, the middle of a word, the end of a word and a single-character word [16].

Figure 3 shows the graphical structure of the coupled model. Given an input sentence, which is a sequence of n characters, denoted by $\mathbf{x} = c_1...c_n$, WS aims to determine the best tag sequence $\mathbf{y} = y_1...y_n$, where $y_i \in \{B, M, E, S\}$ is a segmentation label for c_i. As a log-linear model, CRF defines the probability of a tag sequence as:

$$\begin{aligned} P(\mathbf{y}|\mathbf{x};\theta) &= \frac{exp(Score(\mathbf{x},\mathbf{y};\theta))}{\sum_{\mathbf{y}'} exp(Score(\mathbf{x},\mathbf{y}';\theta))} \\ Score(\mathbf{x},\mathbf{y};\theta)) &= \sum_{1\le i\le n+1} \theta \cdot \mathbf{f}(\mathbf{x}, i, y_{i-1}, y_i) \end{aligned} \tag{1}$$

where $Score(\mathbf{x},\mathbf{y};\theta)$ is a scoring function; $\mathbf{f}(\mathbf{x}, i, y_{i-1}, y_i)$ is the feature vector at the i^{th} character and θ is the feature weight vector. Please note that c_0 and c_{n+1} are two pseudo characters marking the beginning and end of the sentence. We use the features described in zhang et al. (2014) [21], as shown in Table 1.

Suppose the training data is $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{y}_i$ is the provided gold-standard tag sequence for $\mathbf{x}_i$. Then the log likelihood of $\mathcal{D}$ is:

$$\mathcal{L}(\mathcal{D};\theta) = \sum_{i=1}^{N} \log P(\mathbf{y}_i|\mathbf{x}_i;\theta) \tag{2}$$

Table 1. Feature templates for $\mathbf{f}(\mathbf{x}, i, y_{i-1}, y_i)$. $T(c_i)$ returns the type of the character c_i (time, number, punctuation, special symbols, else). $I(c_i, c_j)$ judges whether the two characters c_i and c_j are the same.

Unigram		Bigram
01: $y_i \circ c_k$	$i-2 \leq k \leq i+2$	09: $y_{i-1} \circ y_i$
02: $y_i \circ c_{k-1} \circ c_k$	$i-1 \leq k \leq i+2$	10: $y_{i-1} \circ y_i \circ c_i$
03: $y_i \circ c_{k-1} \circ c_k \circ c_{k+1}$	$i-1 \leq k \leq i+1$	11: $y_{i-1} \circ y_i \circ c_{i-1} \circ c_i$
04: $y_i \circ T(c_k)$	$i-1 \leq k \leq i+1$	
05: $y_i \circ T(c_{k-1}) \circ T(c_k)$	$i \leq k \leq i+1$	
06: $y_i \circ T(c_{i-1}) \circ T(c_i) \circ T(c_{i+1})$		
07: $y_i \circ I(c_i, c_k)$	$i-2 \leq k \leq i+2, k \neq i$	
08: $y_i \circ I(c_{i-1}, c_{i+1})$		

The training objective is to find an optimal θ to maximize the log likelihood, which can be resolved using the standard gradient descend algorithms. We omit the details for space limitation.

3 Coupled CRF for WS

In this section, we introduce the coupled model proposed in Li et al. (2015) [6], which can learn and predict two heterogeneous annotations simultaneously. Figure 4 shows the graphical structure of the coupled CRF. We can see that the only difference from the traditional CRF is that in the coupled model, two tags $[y_i^a, y_i^b]$ are simultaneously assigned for one character, which we name as a pair of *bundled tags*. Then, the score of a bundled tag sequence $[\mathbf{y}^a, \mathbf{y}^b]$ is defined as:

$$Score(\mathbf{x}, [\mathbf{y}^a, \mathbf{y}^b]; \theta) = \sum_{1 \leq i \leq n+1} \theta \cdot \begin{bmatrix} \mathbf{f}(\mathbf{x}, i, [y_{i-1}^a, y_{i-1}^b], [y_i^a, y_i^b]) \\ \mathbf{f}(\mathbf{x}, i, y_{i-1}^a, y_i^a) \\ \mathbf{f}(\mathbf{x}, i, y_{i-1}^b, y_i^b) \end{bmatrix} \tag{3}$$

where the first item of the extended feature vector is called *joint features*, which can be obtained by replacing y_i with bundled tags $[y_i^a, y_i^b]$ in Table 1; the second and third items are called *separate features*, which are based on single-side tags. The coupled model provides us the flexibility of using both kinds of features.

3.1 Creating Bundled Tags

The key challenge for the idea of coupled sequence labeling is that the heterogeneous datasets are non-overlapping and each contains only one-side tags. Therefore, the problem is how to construct training data for the coupled model. Li et al. (2015) [6] study how to construct bundled POS tags and show that best performance can be achieved using a complete mapping between two tag sets. In this work, we use two tag sets $\{B, M, E, S\}$ and $\{b, m, e, s\}$ to distinguish the labels in two heterogeneous data. Then, we map a one-side tag into a

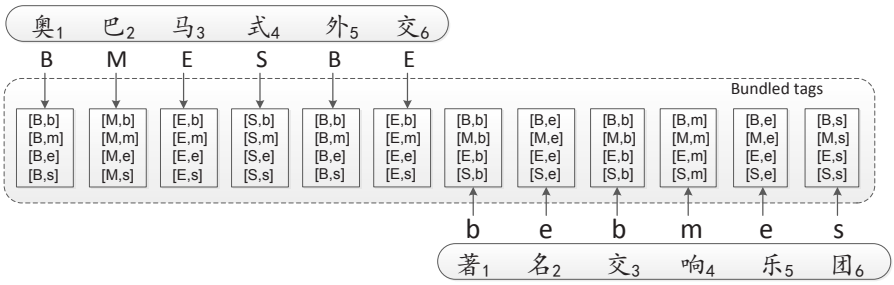

Fig. 5. An example to illustrate how to obtain bundled tags from one-side tags of *CTB7* (above) and *PD* (below).

set of bundled tags by considering all tags at the missing side, as illustrated in Figure 5. In the resulting training data, each sentence has multiple possible tag sequences as its gold-standard reference, which is known as *ambiguous labelings*. In the coupled model, the tag space contains $4 \times 4 = 16$ bundled tags.

3.2 Training Objective with Ambiguous Labelings

After our transformation, each character in a sentence has four bundled tags as gold-standard references, known as ambiguous labels, as shown in Figure 5. Given a sentence $\mathbf{x}$, we denote the set of ambiguous tag sequences as $\mathcal{V}$. The probability of $\mathcal{V}$ is the sum of probabilities of all tag sequences contained in $\mathcal{V}$.

$$P(\mathcal{V}|\mathbf{x};\theta) = \sum_{\mathbf{y}\in\mathcal{V}} P(\mathbf{y}|\mathbf{x};\theta) \tag{4}$$

Suppose the training data is $\mathcal{D} = \{(\mathbf{x_i}, \mathcal{V}_i)\}_{i=1}^{N}$. Then the log likelihood of is:

$$\mathcal{L}(\mathcal{D};\theta) = \sum_{i=1}^{N} \log P(\mathcal{V}_i|\mathbf{x}_i;\theta) \tag{5}$$

Following Li et al. (2015) [6], we can derive the gradient of the likelihood function and apply standard gradient descent algorithms to learn θ. We omit the details for space limitation.

3.3 Stochastic Gradient Descent (SGD) Training with Two Datasets

In this work, we adopt SGD to train both the traditional and coupled CRF models. The basic idea is to derive the gradient w.r.t. θ using a small batch of sampled training data, and use it update θ at each step. However, we need to train a model on two separate data, and one dataset may be overwhelmed by another due to imbalance in scale if directly merging two datasets into one (see table 2). Therefore, we adopt the corpus-weighting strategy proposed in

Li et al. (2015) [6], as shown in Algorithm 1. D_k^b is a subset of training data used in k^{th} step update; b is the batch size; η_k is a update step. The idea is to randomly select instances from each training data in a certain proportion, and use them for one-iteration training. In this work, we use $b = 30$, and follow the implementation in CRFsuite[2] to decide η_k. Actually, in our experiments, we in some cases simultaneously train our model on three datasets, which requires slightly extending the procedure in Algorithm 1.

Algorithm 1. SGD training with two labeled datasets

Input: Two labeled datasets: $\mathcal{D}^1 = \{(x_i^1, v_i^1)\}_{i=1}^N$, $\mathcal{D}^2 = \{(x_i^2, v_i^2)\}_{i=1}^M$;
Parameters: I(Iterations) , N', M', b
Output: θ
Initialization: $\theta_0 = 0, k = 0$;
for $i = 1;\ i <= I;\ i++$ **do**
  $\mathcal{D}_i = N^{'}$ instances from $\mathcal{D}^1 + M^{'}$ instrances from $\mathcal{D}^2$ (randomly select) ;
  Shuffle $\mathcal{D}_i$;
  while *Traverse* $\mathcal{D}_i$ *and* $\mathcal{D}_k^b \subseteq \mathcal{D}_i$ **do**
    $\theta_{k+1} = \theta_k + \eta_k \frac{1}{b} \bigtriangledown \mathcal{L}(\mathcal{D}_k^b; \theta_k)$;
    $k = k + 1$;
  end
end

4 Experiments

4.1 Datasets

Tabel 2 shows the datasets explored in this work. For labeled data, the NLPCC 2015 shared task organizer provides us only a training dataset consisting 10,000 *WB* sentences. To develop our model and tune the parameter settings, we randomly split the training data into two sets: a train set to learn model parameters and a dev set for model evaluation [10]. However, for the final submission, we use the whole training data to train our model. Actually, the organizer also provides a large set of unlabeled *WB* text, which is not considered in this work.

For the third party resources, we follow the suggestion in the data description guideline for data split for *CTB7*, and use the data of Li et al. (2015) [6] for *PD*.

4.2 Experiments on WS

In this section, we conduct a lot of experiments to find the best strategy to utilize multiple heterogeneous data that performs best on *WB*.

[2] http://www.chokkan.org/software/crfsuite/

Table 2. Data statistics

Dataset	Partition	Sentences	Words	Characters
WB	train	8,000	172,800	279,676
	dev	2,000	42,372	68,535
	test	5,000	106,741	172,339
CTB7	train	46,572	1,039,774	168,2485
	dev	2,079	59,955	100,316
	test	2,796	81,578	134,149
PD	train	283,862	6,797,623	11,308,278
	dev	4,965	116,769	194,641
	test	9,913	236,234	393,185

Coupled Models on *CTB7* and *PD*: For the first time, we apply the recently proposed coupled sequence labeling to the task of Chinese WS. We would like to know how the approach performs on canonical data. Table 3 shows the performance of the coupled CRF on *CTB7* using both *CTB7* and *PD* training data, denoted as $Seg_{CTB7\&PD}$. We reimplement the guide-feature based method of Jiang et al. [1], using *PD* to produce guide feature on *CTB7*, denoted as $Seg_{CTB7\leftarrow PD}$. The baseline model Seg_{CTB7} is a traditional CRF trained on *CTB7*. The results show that *the coupled model outperforms both the baseline and the guide-feature based method on the task of WS*.

Model Selection on *WB*-dev: We then try to find the best strategy of utilizing the three datasets that performs best on *WB*. We experiment with different methods of combining the three training datasets (*WB*, *CTB7*, and *PD*) and report F1 scores on the corresponding dev dataset. The results are shown in Table 4.

The first major row refers to baseline CRF models, which are trained on a single training dataset, and evaluated against all three dev datasets. We can see that all models performs best on the dev dataset that corresponds to its training dataset. Moreover, we notice that Seg_{CTB7} has an F1 score on *WB*-dev which is very close to that of Seg_{WB}, indicating that the WS guidelines in *WB* and *CTB7* may be close.

In the second major row, we assume that two training datasets follow the same guideline, and directly combine the two datasets to train traditional CRF models. Algorithm 1 is adopted for training on two datasets with $N' = 5K$

Table 3. WS F1 scores on *CTB7*

Approaches	dev	test
Baseline: Seg_{CTB7}	95.39	94.95
Guide-Feature: $Seg_{CTB7\leftarrow PD}$	95.80 (+0.41)	95.29 (+0.34)
Coupled: $Seg_{CTB7\&PD}$	95.81 (+0.42)	95.45 (+0.50)

Table 4. WS F1 scores on *WB/CTB7/PD* dev datasets

Approaches		*WB*	*CTB7*	*PD*
Baseline	Seg_{WB}	**93.13**	87.90	85.80
	Seg_{CTB7}	92.31	**95.39**	89.33
	Seg_{PD}	89.72	90.23	**97.06**
Merge two datasets	$Seg_{WB+CTB7}$	**94.81**	**95.46**	89.79
	Seg_{WB+PD}	93.70	90.84	**96.38**
	$Seg_{WB+PD^{CTB}}$	94.78	93.43	92.75
Coupled	$Seg_{WB\&CTB7}$	95.06	95.14	–
	$Seg_{WB\&PD}$	95.13	–	**96.96**
	$Seg_{(WB+CTB7)\&PD}$	**95.54**	**95.52**	96.54
	$Seg_{WB\&(CTB7+PD^{CTB})}$	95.14	95.04	–

and $M' = 5K$. For the results, we can see that the model trained on combined *WB* and *CTB7*, denoted as $Seg_{WB+CTB7}$, achieves much higher performance on *WB*-dev than the baseline Seg_{WB} (+1.68%). This clearly shows that *CTB7 can be directly used as extra training data for WS on WB*. The model also slightly outperforms the baseline Seg_{CTB7} on *CTB7*-dev (+0.07%). Directly combining *WB* and *PD* also lead to improved F1 score on *WB*-dev over the baseline model, but is much inferior the case of *CTB7*. PD^{CTB} refer to a modified *PD* data which is converted into the *CTB7* WS annotation using $Seg_{CTB7\&PD}$ (described in Table 3). As described in Li et al. (2015) [6], the coupled model can be naturally used to conduct annotation conversion, based on constrained decoding, given one-side tags. $Seg_{WB+PD^{CTB}}$ achieve very close F1 score to $Seg_{WB+CTB7}$, indicating the *PD can also effectively help WS on WB*. Moreover, all these results also suggest that the WS guidelines in *WB* and *CTB7* may be close.

In the third major row, we conduct experiments with the coupled models trained on multiple datasets. The coupled model trained on *WB* and *CTB7*, denoted as $Seg_{WB\&CTB7}$, achieves slightly higher F1 score on *WB*-dev than $Seg_{WB+CTB7}$ (0.25%), indicating *there still exists some differences in the guidelines of two dataset*. $Seg_{WB\&(CTB7+PD^{CTB})}$ is trained on three dataset, i.e., *WB*, *CTB7*, and PD^{CTB}, assuming the latter two having the homogeneous annotations. However, the model only achieves slightly higher accuracy when using only *CTB7* or PD^{CTB} as the second extra training dataset. The best-performing model is $Seg_{(WB+CTB7)\&PD}$, which assumes that *WB* and *CTB7* are under the same annotation guideline and uses *PD* as the second training dataset. Therefore, we choose $Seg_{(WB+CTB7)\&PD}$ as our final system for result submission. Our next plan is to directly train a coupled model on the three datasets that uses three-side bundled tags.

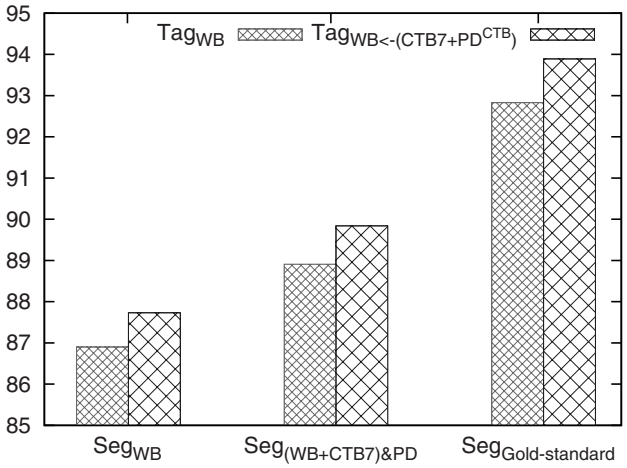

Fig. 6. POS F1 scores on *WB*-dev, with three different settings of WS results

4.3 Experiments on POS Tagging

Our initial plan is to combine *WB*, *CTB7*, and *PD* with the coupled sequence labeling approach, which has an inefficient problem due to the large bundled POS tag set. Due to time limitation, we adopt an ensemble approach, as illustrated in Figure 2. PD^{CTB} refers to a modified *PD* converted into the annotation guideline of *CTB* using the coupled model described in Li et al (2015) [6]. Please note that the coupled model of Li et al (2015) [6] is built on *CTB5* and *PD*.

To better comprehend the effect of different WS approaches on POS F1 scores, we experiment with three settings of WS results on *WB*-dev. In the first setting, *WB*-dev is automatically segmented by the baseline Seg_{WB}; the second setting uses the best-performing model $Seg_{(WB+CTB7)\&PD}$ to produce WS results on *WB*-dev; the third setting uses gold-standard WS results of *WB*-dev. Figure 6 shows the results. We can see that our guide-feature based model $Tag_{WB\leftarrow(CTB7+PD^{CTB})}$ consistently outperforms the baseline model Tag_{WB}.

4.4 Final Results on Test Data

For the final submission, we re-train our models on the whole *WB* training data, using the iteration number that leads to best performance on *WB*-dev during model development. Table 5 shows the results on the test data used for final evaluation. Our effort leads to an improvement on WS F1 score from 93.76% to 95.58% (+1.82%), and on POS F1 score from 87.93% to 88.99% (+1.06%).

Table 5. F1 scores on *WB*-test data

	WS F1	POS F1
Baseline: Seg_{WB}+Tag_{WB}	93.76	87.93
Our best: $Seg_{(WB+CTB7)\&PD}$+$Tag_{WB\leftarrow(CTB7+PD^{CTB})}$	95.58 (**+1.82**)	88.99 (**+1.06**)

5 Related Work

In recent years, Chinese WS and POS tagging has drawn extensive attention and made a lot of progress. Due to the space limitation, we can only make limited references to research closely related with this work. As a promising direction, researchers have proposed different semi-supervised approaches to explore unlabeled data [9,14,18,19]. Particularly, the use of natural annotation has become a hot topic in Chinese word segmentation [3,8,17]. The idea is to derive segmentation boundaries from implicit information encoded in web texts, such as anchor texts and punctuation marks, and use them as partially labeled training data in sequence labeling models.

In the line of exploiting multiple labeled data, Jiang et al. (2009) [1] propose the simple yet effective guild-feature based method, which is further extended in [7,12,13]. Qiu et al. (2013) [11] propose a model that performs heterogeneous Chinese word segmentation and POS tagging and produces two sets of results following *CTB* and *PD* styles respectively. Their model is based on linear perceptron, and uses approximate inference. Instead, this work is based on the coupled sequence labeling model of Li et al. (2015) [6], which is a natural extension of CRFs.

Although we adopt a cascaded framework for WS and POS tagging, it is more popular to jointly solve the two tasks [2,4,23]. For recent works on Chinese lexical analysis on web data, please refer to [15,20,22].

6 Conclusion

We have participated in the NLPCC 2015 shared task on Chinese WS and POS tagging for Weibo Text. For the first time, we apply the recently proposed coupled sequence labeling to the task of WS, and find that coupled sequence labeling also slightly outperforms the guide-feature based method on *CTB7* with extra *PD*. Furthermore, we experiment with different strategies to combine the three datasets (plus *WB*), and achieve promising results. Due to the high time complexity of coupled POS tagging, we instead adopt the guide-feature based method for the POS tagging subtask.

For future work, we plan to advance our progress on exploiting multiple heterogeneous resources in the following directions. First, we will try to experiment with coupled sequence labeling for POS tagging on *WB*, and make comparison with the guide-feature based methods. Second, we would like to seek effective ways to further incorporate unlabeled WB texts. Finally, we find that our baseline WS system lags behind by large margin other systems in the same closed

track, which motivates us to investigate the issue and improve our final performance through enhancing our baseline system.

References

1. Jiang, W., Huang, L., Liu, Q.: Automatic adaptation of annotation standards: Chinese word segmentation and POS tagging - a case study. In: Proceedings of ACL, pp. 522–530 (2009)
2. Jiang, W., Huang, L., Liu, Q., Lü, Y.: A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In: Proceedings of ACL 2008: HLT, pp. 897–904 (2008)
3. Jiang, W., Sun, M., Lü, Y., Yang, Y., Liu, Q.: Discriminative learning with natural annotations: word segmentation as a case study. In: Proceedings of ACL, pp. 761–769 (2013)
4. Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., Isahara, H.: An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging. In: Proceedings of ACL-AFNLP 2009, pp. 513–521 (2009)
5. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning (ICML 2001), pp. 282–289 (2001)
6. Li, Z., Chao, J., Zhang, M., Chen, W.: Coupled sequence labeling on heterogeneous annotations: pos tagging as a case study. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 1: Long Papers, pp. 1783–1792. Association for Computational Linguistics, Beijing july, 2015
7. Li, Z., Che, W., Liu, T.: Exploiting multiple treebanks for parsing with quasisynchronous grammar. In: ACL, pp. 675–684 (2012)
8. Liu, Y., Zhang, Y., Che, W., Liu, T., Wu, F.: Domain adaptation for CRF-based Chinese word segmentation using free annotations. In: Proceedings of EMNLP, pp. 864–874 (2014)
9. Qiu, X., Huang, C., Huang, X.: Automatic corpus expansion for Chinese word segmentation by exploiting the redundancy of web information. In: Proceedings of COLING, pp. 1154–1164 (2014)
10. Qiu, X., Qian, P., Huang, X.: Overview of the nlpcc 2015 shared task: chinese word segmentation and pos tagging for micro-blog texts (2015). arXiv preprint arXiv:1505.07599
11. Qiu, X., Zhao, J., Huang, X.: Joint Chinese word segmentation and POS tagging on heterogeneous annotated corpora with multiple task learning. In: Proceedings of EMNLP, pp. 658–668 (2013)
12. Sun, W.: A stacked sub-word model for joint chinese word segmentation and part-of-speech tagging. In: Proceedings of ACL, pp. 1385–1394 (2011)
13. Sun, W., Wan, X.: Reducing approximation and estimation errors for Chinese lexical processing with heterogeneous annotations. In: Proceedings of ACL, pp. 232–241 (2012)
14. Sun, W., Xu, J.: Enhancing chinese word segmentation using unlabeled data. In: Proceedings of EMNLP, pp. 970–979 (2011)
15. Wang, A., Kan, M.Y.: Mining informal language from chinese microtext: joint word recognition and segmentation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 731–741. Association for Computational Linguistics, Sofia, August 2013

16. Xue, N., et al.: Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing **8**(1), 29–48 (2003)
17. Yang, F., Vozila, P.: Semi-supervised Chinese word segmentation using partial-label learning with conditional random fields. In: Proceedings of EMNLP, pp. 90–98 (2014)
18. Zeng, X., Wong, D.F., Chao, L.S., Trancoso, I.: Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 770–779. Association for Computational Linguistics, Sofia, August 2013
19. Zhang, L., Wang, H., Sun, X., Mansur, M.: Exploring representations from unlabeled data with co-training for Chinese word segmentation. In: Proceedings of EMNLP, pp. 311–321 (2013)
20. Zhang, L., Wang, H., Sun, X., Mansur, M.: Improving Chinese word segmentation on micro-blog using rich punctuations. In: Proceedings of ACL: Short Papers (2013)
21. Zhang, M., Zhang, Y., Che, W., Liu, T.: Character-level Chinese dependency parsing. In: Proceedings of ACL, pp. 1326–1336 (2014)
22. Zhang, M., Zhang, Y., Che, W., Liu, T.: Type-supervised domain adaptation for joint segmentation and POS-tagging. In: Proceedings of COLING, pp. 588–597 (2014)
23. Zhang, Y., Clark, S.: Joint word segmentation and POS tagging using a single perceptron. In: Proceedings of ACL 2008: HLT, pp. 888–896 (2008)