Improved Structure Regularization for Chinese Part-of-Speech Tagging

Xu Sun
MOE Key Laboratory of Computational Linguistics, Peking University

Abstract

Chinese part-of-speech (POS) tagging is important for Chinese NLP applications. While Chinese POS tagging is typically casted as a sequential tagging task, we find that it is easy to be structurally overfitting on the training data. We propose an "improved structure regularization (ISR)" method for Chinese POS tagging, which regularize the training sample into multiple mini-samples. By regularizing the training samples, we can simplify the structures, further improve the accuracy, and at the same time accelerate the training speed of the tagger. Experiments on standard Chinese POS tagging benchmark dataset show that our tagger achieves 95.56% on the test data, and at the same time with a by-product of 5$\times$ faster training speed. To our knowledge, this result is better than the existing best report.

1 Introduction

Chinese part-of-speech (POS) tagging is important for Chinese natural language processing applications, including named entity recognition, syntactic parsing, and statistical machine translation. Chinese POS tagging is typically casted as a sequential tagging task, and there have been many methods proposed for solving this tagging problem. A typical sequence tagger is implemented by using, for example, conditional random fields (CRF), via treating Chinese POS tagging as linear chain structured prediction problems.

However, as suggested by Sun (2014), structured learning is very easy to be overfitting. While it is obvious that structural dependencies can effectively incorporate structural information, it is less obvious that structural dependencies can also increase the risk of overfitting, and that more complex structures are easier to suffer from overfitting. Since this type of overfitting is caused by structure based complexity, it can hardly be solved by ordinary regularization methods such as L_2 and L_1 regularization schemes, which is only for controlling weight complexity. In our experiments we find that the traditional POS tagger based on linear chain structured prediction is easy to be (structurally) overfitting on the training data, which limits the accuracy of the trained tagger.

Inspired by Sun (2014), we propose an "improved structure regularization" method for Chinese POS tagging, which regularize the training sample into multiple mini-samples with simpler structures, deriving a POS tagger with less overfitting risk from structural dependencies. By regularizing the dependency structures of the training samples, we can simplify the structures, further improve the accuracy. In other words, the proposed method can be interpreted as a back-off method from "fully" structured prediction towards "partially" structured prediction, which is less easier to be overfitting and achieves better tagging accuracy. Moreover, as a by-product, "improved structure regularization" can accelerate the training speed of the tagger, simply because the structural dependencies are simplified.

The contribution of this work is that we propose an improved structure regularization method for Chinese POS tagging, which is easy to implement and can achieve substantially better results on Chinese POS tagging. As a by-product, the proposed method can also accelerate the training speed of the POS tagger. Experiments on standard Chinese POS tagging benchmark dataset show that our tagger achieves 95.56% on the test data. To our knowledge, this result is better than existing best report.

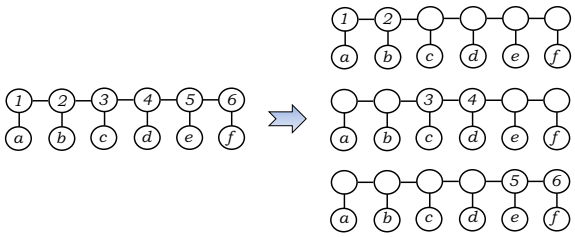

Figure 1: Illustration of improved structure regularization with $\alpha = 3$.

2 Improved Structure Regularization

In our Chinese POS tagging task, we find the structural dependencies are easy to suffer from overfitting. To deal with the overfitting from structural dependencies, we propose an improved structure regularization technique to simplify the structures in Chinese POS tagging, which is novel extension of the original structure regularization method described in Sun (2014). Here, the term "structure" indicates the the structure of tags, i.e., the structural dependencies among the tags. The Chinese POS tagging task is usually casted as a sequential labelling problem.

We follow the general idea of the pioneer work of Sun (2014) to develop the improved structure regularization algorithms for our Chinese POS tagging task, and the detailed method is different and novel compared with the prior work of structure regularization (Sun, 2014). Our improved structure regularization algorithm is described in Algorithm 1. We perform this sample-breaking operation as a simple preprocessing step before each training iteration. By forced breaking of the original training samples, we get new training samples with smaller complexity of tag interactions. We use a scaler α to denote the number of copied samples, with $1 \leq \alpha \leq n$, and n is the length of the sentence (number of words). As we can see, actually α represents the strength of regularization, because the structure is going to be simpler when the value of α is increased. The exact value of α should be automatically tuned and determined by testing via cross validation or simply using the development data.

As we can see, the major difference of the proposed method compared with the existing structure regularization method is that we do not directly split the original samples into mini-samples. Instead, we copy the original sample into multiple copied samples. After that, we mask the copied samples by randomly masking some tags of the copied samples into null tags. Masking of the tags into null tags is to reduce the structural dependencies. In this way, the masked null tags are not involved in the gradient calculation, and will not increase the computational cost. One advantage of our improved version compared with the original structure regularization method is that the features are not affected by decomposing structures anymore, so that our method is more stable and can achieve better results, without increase on the computational cost. The implementation is also simpler and flexible by using this improved version.

In other words, our implementation of structure regularization is simply forced breaking of the tags of the training samples, and at the same time still keep the same words (i.e., do not break the words). We do not break the words because we do not want to lose features – some features are extracted based on a large local window, and we may lose those large window features if we also break the words.

We use an additional example to illustrate our idea. Suppose we have a sentence of 6 words, and this sentence has 6 POS tags in our POS tagging task. We denote this sentence as (`abcdef`, `123456`), where `abcdef` represents the 6 words and `123456` represents the 6 corresponding tags. Those 6 tags are inter-dependent as far as they are in the same training sample (i.e, in the same sentence), and we can say that this sentence has the structure complexity of 6. Then, suppose we want to regularize this structure of the complexity 6 to simpler structures of the complexity 2, we can simply (forced) break the original sample (`abcdef`, `123456`) into three new samples: (`abcdef`, `12XXXX`), (`abcdef`, `XX34XX`), and (`abcdef`, `XXXX56`). In the new samples, the tag `X` simply means there is no tag at this position, or more precisely, there is no need to tag the corresponding word. Hence, for the new sample (`abcdef`, `12XXXX`), it means we only care about the tags of `ab`, and we still keep the words `cdef` because we simply do not want to lose the original features (for example, we can still use 3-gram or 4-gram features in this case). An illustration is shown in Figure 1.

3 Related Work

Many algorithms have been applied to computationally assigning POS labels to English words,

Algorithm 1 Improved Structure Regularization

Input: model weights $\boldsymbol{w}$, training set S, decomposition strength α
repeat
    Sample $\boldsymbol{z}$ uniformly at random from S
    Randomly copy $\boldsymbol{z}$ into α copied samples
    For each copied sample, randomly mask $(\alpha - 1)/\alpha$ continuous tags into null tags
    Update for each copied sample $\boldsymbol{z}'$ such that $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla g_{\boldsymbol{z}'}(\boldsymbol{w})$
until Convergence
return $\boldsymbol{w}$

and such methods have been widely applied to many other languages as well. Previous work, however, shows that tagging for Chinese is more challenging and a number of augmentations and changes became necessary. While state-of-the-art tagging systems have achieved accuracies above 97% on English, Chinese POS tagging has proven to be more challenging and obtains accuracies about 94-95% (Tseng et al., 2005; Huang et al., 2009a; Hatori et al., 2011; Sun and Uszkoreit, 2012; Sun et al., 2013).

Early work usually formulated Chinese POS tagging as a sequential classification problem and employed various sequence labeling models for solution. Experiments in Huang et al. (2007) indicated that a naive HMM is very hard to achieve good classification performance. Huang et al. (2009a) proposed to utilize latent annotations to enhance an HMM and obtained significantly better accuracy.

More recently, Hatori et al. (2011) introduced a joint model for POS tagging and dependency parsing and showed that the POS tagging is improving with joint modeling. Sun et al. (2013) designed and used paradigmatic information to improve Chinese POS tagging. Sun and Uszkoreit (2012) proposed to use word clusters that are automatically induced from large-scale raw texts to improve Chinese POS tagging.

References

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, pages 1–8.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research* 3:951–991.

Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2011. Incremental joint POS tagging and dependency parsing in Chinese. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, pages 1216–1224.

Zhongqiang Huang, Vladimir Eidelman, and Mary Harper. 2009a. Improving a simple bigram hmm part-of-speech tagger by latent annotation and self-training. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, Boulder, Colorado, pages 213–216.

Zhongqiang Huang, Vladimir Eidelman, and Mary P. Harper. 2009b. Improving a simple bigram hmm part-of-speech tagger by latent annotation and self-training. In *HLT-NAACL'09 (Short Papers)*. pages 213–216.

Zhongqiang Huang, Mary Harper, and Wen Wang. 2007. Mandarin part-of-speech tagging and discriminative reranking. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, pages 1093–1102.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*. pages 282–289.

Shuming Ma and Xu Sun. 2016. A new recurrent neural CRF for learning non-linear edge features. *CoRR* abs/1611.04233. http://arxiv.org/abs/1611.04233.

Ryan T. McDonald, Koby Crammer, and Fernando C. N. Pereira. 2005. Online large-margin training of dependency parsers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Weiwei Sun, Xiaochang Peng, and Xiaojun Wan. 2013. Capturing long-distance dependencies in sequence models: A case study o
In *Proceedings of the Sixth International Joint*

Conference on Natural Language Processing. Asian Federation of Natural Language Processing, Nagoya, Japan, pages 180–188. http://www.aclweb.org/anthology/I13-1021.

Weiwei Sun and Hans Uszkoreit. 2012. Capturing paradigmatic and syntagmatic lexical relations: Towards accurate chinese part-of-speech tagging. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, pages 242–252. http://www.aclweb.org/anthology/P12-1026.

Xu Sun. 2014. Structure regularization for structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2402–2410.

Xu Sun. 2016. Asynchronous parallel learning for neural networks and structured models with dense features. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. pages 192–202.

Xu Sun, Xuancheng Ren, Shuming Ma, and Houfeng Wang. 2017. meprop: Sparsified back propagation for accelerated deep learning with reduced overfitting. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. pages 3299–3308.

Xu Sun, Houfeng Wang, and Wenjie Li. 2012. Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 253–262.

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005. Morphological features help POS tagging of unknown words across language varieties. In *The Fourth SIGHAN Workshop on Chinese Language Processing*.

Yoshimasa Tsuruoka, Yusuke Miyao, and Junąŕichi Kazama. 2011. Learning with lookahead: Can history-based models rival globally optimized models? In *Conference on Computational Natural Language Learning (CoNLL)*.