Towards Very Fast and Accurate Part-of-Speech Tagging: A Decomposed and Parallelized Learning Method

Xu Sun
MOE Key Laboratory of Computational Linguistics, Peking University

Abstract

A POS tagger is either based on non-structured or structured classification solutions. Surprisingly, we find that both of them are problematic on POS tagging – non-structured classification is "under-fitting" and structured classification is "over-fitting" on structures. We propose a decomposed learning method, which is in the middle of structured and non-structured classification. In this way, we can easily control the complexity of structures, and find a balance between over-fitting and under-fitting. Moreover, the decomposed mini-samples can be naturally used for parallel learning, and we propose a parallel learning scheme for speeding up the learning. Our experiments are on POS tagging benchmark tasks in English and Chinese. Results demonstrate that our simple method can achieve record-breaking accuracies in both English and Chinese tasks, with the error rate reductions of 2.3% and 2.4% over the existing best systems on English and Chinese, respectively. At the same time, our method is more than 15 times faster than existing methods, accomplishing the training in 12 minutes for English and in 5 minutes for Chinese POS tagging.

1 Introduction

Automatic part-of-speech (POS) tagging is a very traditional and fundamental task in natural language processing. In general, a part-of-speech represents a linguistic class of words, which contains the syntactic and morphological information of the word. Automatic tagging of POS for words plays an important role in higher level NLP applications, such as syntactic parsing, named entity recognition, and statistical machine translation.

There have been many methods proposed for solving the POS tagging problem. The early work on POS tagging is using non-structured classification methods such as the maximum entropy (ME) method. More recently, the structured classification methods, e.g., sequential labelling methods such as conditional random fields (CRF), are dominating on POS tagging, by treating sentences as linear chain structures for structured classification. We argue that this trend could have been misdirected, because our study suggests that complex structures are actually harmful to model accuracy. While it is obvious that intensive structural dependencies can effectively incorporate structural information, it is less obvious that intensive structural dependencies have a drawback of increasing the generalization risk, because more complex structures are easier to suffer from overfitting. Since this type of overfitting is caused by structure 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.

To solve this problem, we propose a decomposed learning method for POS tagging, which decomposes training samples into mini-samples with simpler structures, deriving a model with better generalization power. We show that the proposed decomposed learning method has good theoretical justification – the "stability" can be improved and many prior work has shown that improvement on stability can lead to improvement on generalization power. The proposed method can be interpreted as a back-off method from structured classification towards non-structured classification, i.e., a model in the middle of structured and non-structured classification. In this way, we can easily control the complexity of structures, and can find a balance between over-fitting and under-fitting.

Moreover, the decomposed mini-samples can be naturally used for parallel learning. We propose an efficient parallel learning scheme, which can improve the training speed of by more than 20 times. We perform experiments on well-known POS tagging benchmark tasks in different languages, including English and Chinese. Experimental results demonstrate that our simple method can easily beat the best existing systems, achieving record-breaking accuracies on both English and Chinese POS tagging tasks, and is 20 times faster than existing methods (e.g., can finish the training in 300 seconds with CRF).

The contributions of this work are three-fold:

- On the methodology side, we propose a decomposed learning method for POS tagging, which is in the middle of structured and non-structured classification. Moreover, the decomposed mini-samples can be naturally used for parallel learning, and we propose an efficient parallel learning scheme.
- On the application side, our simple method can achieve record-breaking accuracies in both English and Chinese tasks, with the error rate reductions of 2.3% and 2.4% over the existing best systems on English and Chinese, respectively. At the same time, our method is more than 15 times faster than existing methods, accomplishing the training in 12 minutes for English and in 5 minutes for Chinese POS tagging.
- On the theoretical side, we show that the proposed method can effectively improve the stability of the model, and the improvement of stability can lead to the improvement of generalization power. This explains why the proposed method can achieve record-breaking accuracies.

2 Related Work

First, we review the related work of POS tagging in English and Chinese. Then, we introduce the related work of the proposed decomposed and parallelized learning method.

2.1 English POS Tagging

Many method have been studied for English POS tagging, including the non-structured classification and the structured classification methods. The non-structured classification POS tagger include for example the maximum entropy taggers (Ratnaparkhi, 1996; Toutanova and Manning, 2000) and the SVM based tagger (GimÍęnez and MÍd’rquez, 2004). The structured classification methods in POS tagging include the hidden Markov model tagger (Brants, 2000), the structured perceptron (Collins, 2002), the perceptron training with lookahead (Tsuruoka et al., 2011), the bidirectional perceptron learning algorithm (Shen et al., 2007), and the maximum entropy cyclic dependency network (Toutanova et al., 2003).

2.2 Chinese POS Tagging

As a representative agglutinative language, Chinese has little morphology information, thus a number of changes are necessary in dealing with Chinese POS tagging. While the English POS tagging has relatively high accuracies about 97%, Chinese POS tagging is more difficult and obtains relatively low accuracies, ranging from 93% to 94% (Tseng et al., 2005; Huang et al., 2007; Huang et al., 2009; Li et al., 2011; Sun and Uszkoreit, 2012).

Both non-structured and structured prediction models have been studied for Chinese POS tagging (Tseng et al., 2005; Huang et al., 2007; Huang et al., 2009; Li et al., 2011; Sun and Uszkoreit, 2012). In Tseng et al. (2005), a maximum entropy model with morphological features are used for unknown word recognition. In Huang et al. (2007) and Huang et al. (2009), generative HMM models are used for Chinese POS tagging. Huang et al. (2007) proposed an HMM model with a re-ranking scheme and additional morphological and syntactic features for Chinese POS tagging. Huang et al. (2009) proposed an HMM model enhanced with latent variables for learning complex dependencies. Their experimental results on the Chinese Treebank are about 93% to 94% in terms of accuracy. More recently, Sun and Uszkoreit (2012) proposed a method for Chinese POS tagging by incorporating additional syntactic structure and word clustering information, which are extracted from additional large-scale unlabelled data (Chinese Gigaword).

2.3 Related Work of the Proposed Method

The related work on decomposed learning is relatively few, including the studies of (Sutton and McCallum, 2007) and (Samdani and Roth, 2012) on piecewise/decomposed training methods, the study of (Tsuruoka et al., 2011) on a “lookahead" learning method, and the study of structured regularization in (Sun, 2014). Our work differs from the prior work mainly because our work is built on a decomposed and parallelized learning framework, with theoretical arguments and justifications on improving stability for structured classification, and the detailed algorithm is quite different.

As for stochastic/online learning, stochastic gradient descent (SGD) is a popular training algorithm with rapid learning rates (Bertsekas, 1999; Bottou and Bousquet, 2008; Shalev-Shwartz and Srebro, 2008). Recently, a variety of parallelized and distributed versions has been proposed, including the synchronous parallel SGD training method (Langford et al., 2009) and the asynchronous (lock-free) parallel SGD training algorithm known as HOGWILD (Niu et al., 2011). The novelty of our work is that our parallel learning scheme is a decomposed one, which is very natural for parallel online learning and with very fast training speed.

On theoretical analysis on stability and generalization risk, related studies include (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2009) on non-structured classification and (Taskar et al., 2003; London et al., 2013a; London et al., 2013b) on structured classification.

3 Decomposed and Parallelized Learning

We first introduce the problem setting and definitions. Then, we described the proposed decomposed and parallelized learning method.

3.1 Problem Setting and Definitions

The observations can be indexed and be denoted by using an indexed sequence of observations $\boldsymbol{O} = \{o_1, \ldots, o_n\}$. We use the term *sample* to call $\boldsymbol{O} = \{o_1, \ldots, o_n\}$. In POS tagging, a sample corresponds to a sentence of n words with dependencies of linear chain structures. For simplicity in description and analysis, here we assume all samples have n observations (thus n tags). In a typical setting of structured prediction, all the n tags have inter-dependencies via connecting each Markov dependency between neighboring tags. Thus, we call n as *structure complexity* below.

A sample is converted to an indexed sequence of feature vectors $\boldsymbol{x} = \{\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}\}$, where $\boldsymbol{x}_{(k)} \in \mathcal{X}$ is of the dimension d and corresponds to the local features extracted from the position/index k. We can use an $n \times d$ matrix to represent $\boldsymbol{x} \in \mathcal{X}^n$. Let $\mathcal{Z} = (\mathcal{X}^n, \mathcal{Y}^n)$ and let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}$ denote a sample in the training data.

Suppose a training set is

$$S = \{\boldsymbol{z}_1 = (\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, \boldsymbol{z}_m = (\boldsymbol{x}_m, \boldsymbol{y}_m)\}$$

with size m, and the samples are drawn i.i.d. from a distribution D which is unknown. A learning algorithm is a function $G : \mathcal{Z}^m \mapsto \mathcal{F}$ with the function space $\mathcal{F} \subset \{\mathcal{X}^n \mapsto \mathcal{Y}^n\}$.

For structured prediction, a local classification on a position depends on the whole input $\boldsymbol{x} = \{\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}\}$ rather than a local window, due to the structural dependencies. To simplify the notation, we define

$$g(\boldsymbol{x}, k) \triangleq g(\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}, k)$$

We define *point-wise cost function* $c : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ as $c[G_S(\boldsymbol{x}, k), \boldsymbol{y}_{(k)}]$, which measures the cost on a position k by comparing $G_S(\boldsymbol{x}, k)$ and the gold-standard tag $\boldsymbol{y}_{(k)}$, and we introduce the point-wise loss as

$$\ell(G_S, \boldsymbol{z}, k) \triangleq c[G_S(\boldsymbol{x}, k), \boldsymbol{y}_{(k)}]$$

Then, we define *sample-wise cost function* $C : \mathcal{Y}^n \times \mathcal{Y}^n \mapsto \mathbb{R}^+$, which is the cost function with respect to a whole sample, and we introduce the sample-wise loss as

$$\mathcal{L}(G_S, \boldsymbol{z}) \triangleq C[G_S(\boldsymbol{x}), \boldsymbol{y}] = \sum_{k=1}^{n} \ell(G_S, \boldsymbol{z}, k)$$

Given G and a training set S, what we are most interested in is the *generalization risk* in structured prediction (i.e., expected average loss) (Taskar et al., 2003; London et al., 2013a; London et al., 2013b):

$$R(G_S) = \mathbb{E}_{\boldsymbol{z}}\Big[\frac{\mathcal{L}(G_S, \boldsymbol{z})}{n}\Big]$$

Since the distribution D is unknown, we have to estimate $R(G_S)$ by using the *empirical risk*:

$$R_e(G_S) = \frac{1}{mn} \sum_{i=1}^{m} \mathcal{L}(G_S, \boldsymbol{z}_i)$$

To train a structured prediction model, the target is to find the minimizer of the empirical risk $R_e(G_S)$, and typically with an additional regularizer for controlling weight complexity (i.e., weight regularization). That is,

$$\text{minimize}_{G_S} \frac{1}{mn} \sum_{i=1}^{m} \sum_{k=1}^{n} \ell(G_S, \boldsymbol{z}_i, k) + R(G_S) \quad (1)$$

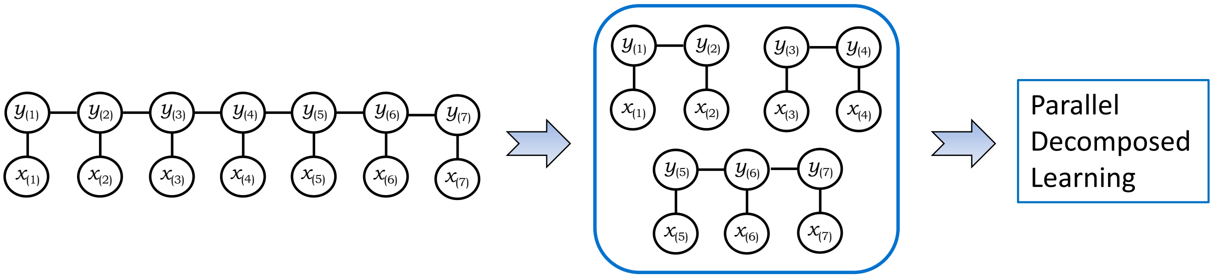

Figure 1: An illustration of decomposed and parallelized learning, which randomly decompose a training sample $\boldsymbol{z}$ with structure complexity 7 into three mini-samples with structure complexities of 2 and 3 (i.e., with expected structure complexity of $\frac{7}{3}$). The decomposed mini-samples are naturally used for parallel learning.

Algorithm 1 Decomposed and Parallelized Learning

```
Input: model weights $\boldsymbol{w}$, training set $S$ of $m$ samples and with expected sample size $n$, decomposition strength $\alpha$
repeat
    Sample $\boldsymbol{z}$ uniformly at random from $S$
    Randomly decompose $\boldsymbol{z}$ into mini-samples, each with expected size $n/\alpha$
    Update in parallel for each mini-sample $\boldsymbol{z}'$ such that $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla g_{\boldsymbol{z}'}(\boldsymbol{w})$
until Convergence
return $\boldsymbol{w}$
```

The formula (1) is a general representation of the objective function of structured prediction models. Our proposed is a general-purpose method, rather than depending on a specific structured prediction model. Thus, the denotation of our method and the theoretical analysis will be based on those general definitions of structured prediction models. Below, we describe the proposal of decomposed and parallelized learning method.

3.2 Decomposed and Parallelized Learning

Let $g(\boldsymbol{w})$ be the structured prediction objective function and $\boldsymbol{w} \in \mathcal{W}$ is the weight vector. Recall that the SGD update with fixed learning rate η has a form like this:

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta \nabla g_{\boldsymbol{z}_t}(\boldsymbol{w}_t) \tag{2}$$

where $g_{\boldsymbol{z}}(\boldsymbol{w}_t)$ is the stochastic estimation of the objective function based on $\boldsymbol{z}$ which is randomly drawn from the training set S.

Following prior work on asynchronous parallel training (Niu et al., 2011), we assume a shared memory machine with k processors, and a vector of variables $\boldsymbol{w}$ in the shared memory is accessible to all processors. Each processor can read and update $\boldsymbol{w}$, with an assumption that the component-wise addition operation is atomic, in other words, $\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i + v$ can be performed atomically.

The decomposed and parallelized learning method draws a training sample $\boldsymbol{z}$ at random from the training set. Recall that the training set S is of m samples and with expected sample size n. Assume we set the decomposition strength as α with $1 \leq \alpha \leq n$. Then, with a distribution (e.g., Gaussian distribution), the sample $\boldsymbol{z}$ is randomly decomposed into multiple mini-samples $N_\alpha(\boldsymbol{z}_i)$ with sub-structures, such that each mini-sample has expected size n/α.

In other words, $N_\alpha(\boldsymbol{z}_i)$ randomly splits $\boldsymbol{z}_i$ into α mini-samples $\{\boldsymbol{z}_{(i,1)}, \ldots, \boldsymbol{z}_{(i,\alpha)}\}$, so that the mini-samples have a distribution on their sizes (structure complexities) with the expected value $n' = n/\alpha$.

Then, based on the multiple mini-samples and the multicore computing machine, the algorithm update in parallel for each mini-sample $\boldsymbol{z}' \in N_\alpha(\boldsymbol{z}_i)$ with SGD update

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla g_{\boldsymbol{z}'}(\boldsymbol{w}) \tag{3}$$

As we can see, the proposed method is very simple. The algorithm is summarized in Algorithm 1.

4 Theoretical Analysis

Then, we analyze the stability and generalization risk of structured prediction based on our training algorithm. We show that a proper setting of the decomposition strength α can effective reduce the stability and generalization risk of structured prediction, thus giving a reasonable expectation that our algorithm can improve the structured prediction accuracy in testing on new samples.

4.1 Stability and Generalization

To state our theoretical results of overfitting risk, we must describe several quantities and assumptions following prior work (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2009).

We assume a simple real-valued structured prediction scheme such that the class predicted on position k of $\boldsymbol{x}$ is the sign of $G_S(\boldsymbol{x}, k) \in \mathcal{D}$.[1] Also, we assume the point-wise cost function c_τ is convex and τ-*smooth* such that $\forall y_1, y_2 \in \mathcal{D}, \forall y^* \in \mathcal{Y}$

$$|c_\tau(y_1, y^*) - c_\tau(y_2, y^*)| \le \tau |y_1 - y_2| \quad (4)$$

Also, we use a value ρ to quantify the bound of $|G_S(\boldsymbol{x}, k) - G_{S^{\setminus i}}(\boldsymbol{x}, k)|$ while changing a single sample (with size $n' \le n$) in the training set with respect to the structured input $\boldsymbol{x}$. This ρ-*admissible* assumption can be formulated as $\forall k$,

$$|G_S(\boldsymbol{x}, k) - G_{S^{\setminus i}}(\boldsymbol{x}, k)| \le \rho ||G_S - G_{S^{\setminus i}}||_2 \cdot ||\boldsymbol{x}||_2 \quad (5)$$

where $\rho \in \mathbb{R}^+$ is a value related to the design of G.

Theorem 1 (Stability and generalization) *With a training set S of size m, let the regularized objective function g have the minimizer f:*

$$\begin{aligned} f &= \operatorname*{argmin}_{g \in \mathcal{F}} R_{\alpha,\lambda}(g) \\ &= \operatorname*{argmin}_{g \in \mathcal{F}} \Big(\frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, \boldsymbol{z}'_j) + \frac{\lambda}{2} ||g||_2^2 \Big) \end{aligned} \quad (6)$$

where α is the decomposition strength with $1 \le \alpha \le n$. Let the point-wise loss function ℓ_τ is bounded by $\forall k, 0 \le \ell_\tau(G_S, \boldsymbol{z}, k) \le \gamma$. Let $R(f)$ and $R_e(f)$ be defined like before. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of the training set S, the generalization risk $R(f)$ is bounded by

$$R(f) \le R_e(f) + 2\tau\bar{\Delta} + \Big(4m\tau\bar{\Delta} + \gamma\Big)\sqrt{\frac{\ln \delta^{-1}}{2m}} \quad (7)$$

where $\bar{\Delta}$ denotes the function stability of f for $\forall \boldsymbol{z} \in \mathcal{Z}$ with $|\boldsymbol{z}| = n$, which is bounded by

$$\bar{\Delta} \le \frac{d\tau\rho^2 v^2 n^2}{m\lambda\alpha} \quad (8)$$

The proof can be extended from (Bousquet and Elisseeff, 2002) and (Sun, 2014). For the limit of space, we omit the full proof here. We can see from (8) that the structure-decomposition factor α can linearly improve (make it linearly smaller) the function stability term $\bar{\Delta}$. Furthermore, we can see from (7) that smaller function stability $\bar{\Delta}$ leads to smaller generalization risk between the empirical risk and the true expected risk.

Since the γ is typically with small value (in normalized loss, we have $\gamma = 1$), and the number of training samples m is typically with big value (especially with data-intensive tasks), the term $4m\tau\bar{\Delta}$ is dominating compared with the term γ. We can see that the number of training samples m and the regularization term λ also can reduce the generalization risk.

To summarize, by theoretical analysis we show that the proposed decomposed learning method is theoretically sound, because it can linearly improve the stability and generalization power (i.e., reduce the generalization risk) in structured prediction. In next section, we will show in experiments that our decomposed learning method can achieve much better accuracy than existing structured prediction methods, which empirically confirms our theoretical analysis.

References

D.P. Bertsekas. 1999. *Nonlinear Programming*. Athena Scientific.

Léon Bottou and Olivier Bousquet. 2008. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 161–168.

Olivier Bousquet and Andrĺę Elisseeff. 2002. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Thorsten Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing*, pages 224–231. Association for Computational Linguistics.

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*

[1] Many popular structured prediction models have a convex and real-valued cost function (e.g., CRFs).

natural language processing-Volume 10, pages 1–8. Association for Computational Linguistics.

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

Jianfeng Gao, Xiaolong Li, Daniel Micol, Chris Quirk, and Xu Sun. 2010. A large scale ranker-based system for search query spelling correction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING ’10, pages 358–366.

Jesĺšs Gimĺęnez and Lluĺłs Mĺd’rquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *LREC’04*. European Language Resources Association.

Zhongqiang Huang, Mary P. Harper, and Wen Wang. 2007. Mandarin part-of-speech tagging and discriminative reranking. In *EMNLP-CoNLL’07*, pages 1093–1102.

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

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.

John Langford, Alex J. Smola, and Martin Zinkevich. 2009. Slow learners are fast. In *NIPS’09*, pages 2331–2339.

Zhenghua Li, Min Zhang 0005, Wanxiang Che, Ting Liu, Wenliang Chen, and Haizhou Li. 2011. Joint models for chinese pos tagging and dependency parsing. In *EMNLP’11*, pages 1180–1191.

B. London, B. Huang, B. Taskar, and L. Getoor. 2013a. Pac-bayes generalization bounds for randomized structured prediction. In *NIPS Workshop on Perturbation, Optimization and Statistics*.

Ben London, Bert Huang, Ben Taskar, and Lise Getoor. 2013b. Collective stability in structured prediction: Generalization from one example. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 828–836.

Shuming Ma and Xu Sun. 2016. A new recurrent neural CRF for learning non-linear edge features. *CoRR*, 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)*.

Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 693–701.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing 1996*, pages 133–142.

Rajhans Samdani and Dan Roth. 2012. Efficient decomposed learning for structured prediction. In *ICML’12*.

Shai Shalev-Shwartz and Nathan Srebro. 2008. Svm optimization: inverse dependence on training set size. In *ICML’08*, pages 928–935. ACM.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. 2009. Learnability and stability in the general learning setting. In *Proceedings of COLT’09*.

Libin Shen, Giorgio Satta, and Aravind K. Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proceedings of ACL’07*.

Weiwei Sun and Hans Uszkoreit. 2012. Capturing paradigmatic and syntagmatic lexical relations: Towards accurate chinese part-of-speech tagging. In *Proceedings of ACL’12*, pages 242–252.

Xu Sun, Louis-Philippe Morency, Daisuke Okanohara, Yoshimasa Tsuruoka, and Jun’ichi Tsujii. 2008. Modeling latent-dynamic in shallow parsing: A latent conditional model with imrpoved inference. In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 841–848.

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.

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. 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.

Charles A. Sutton and Andrew McCallum. 2007. Piecewise pseudolikelihood for efficient training of conditional random fields. In *ICML'07*, pages 863–870. ACM.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2003. Max-margin markov networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of EMNLP 2000*, pages 63–70.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL'03*.

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005. Morphological features help pos tagging of unknown words across language varieties. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 32–39.

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)*.