Improving Chinese Part-of-Speech Tagging with Regularized Global Information

Yang Yang, Xu Sun
MOE Key Laboratory of Computational Linguistics, Peking University

Abstract

Chinese part-of-speech (POS) tagging is a core Chinese NLP task, which is critical for higher level tasks. First, we investigate global information for improving Chinese POS tagging, including syntactical information and word clustering information from large-scale data. Second, we find that the global information is easy to suffer from overfitting, due to the complex structures and long range dependencies of the global information. Since this type of overfitting can hardly be solved by traditional regularization methods, we propose two solutions, structure regularization and regularized importance weighting, for effectively reducing the overfitting from the global information. By regularizing the global information, we can simplify the structures, further improve the accuracy, and at the same time substantially speed up the training speed of the tagger (better scalability).

1 Introduction

Parts of speech (POS) are linguistic categories of words, which are generally defined by the syntactic or morphological behavior of the word in question. POS tagging is important for many higher level tasks because of the large amount of information a POS tag gives about a word and its neighbors. The Chinese language has a number of characteristics that make Chinese language processing particularly challenging. While state-of-the-art tagging systems have achieved accuracies above 97% on English, Chinese POS tagging has proven to be more difficult and obtained accuracies about 94-95% (Hatori et al., 2011; Sun and Uszkoreit, 2012).

Global information plays an important role in boosting the tagging performance for Chinese. In this paper, we investigate the usefulness of intra-sentence global information by extending the domain of contextual features for a sequential labelling tagger. This strategy significantly increases the number of parameters, and a naive implementation thus decreases the prediction accuracy. To deal with this problem, we incorporate pseudo training data that are automatically generated by a parser-guided model. We also investigate the the effectiveness of inter-sentence global information by word clutsering. The word clusters generated from extra raw data (without annotation) enhance the consistency of prediction of distributionally similar words.

While the global information that we employed is helpful to our task, in preliminary experiments we also find the global information is easy to suffer from overfitting. The global information is based on complicated structures as well as long range dependencies, which are quite easy to be overfitting on the training data set. Thus, when we test the trained tagger on the new data set that are not observed in the limited training data, the trained tagger based on global information is going to have low accuracy due to the overfitting on the training data set.

The regularization technique is a good choice to deal with the overfitting problem. A typical choice of the regularization technique is the well known regularization techniques focused on model weights (i.e., so called weight-based regularization), such as L_2 (a Gaussian prior) and L_1 (a sparsity favored prior) regularization over the model weights. Unfortunately, we find those traditional weight-

based regularization methods perform poorly in our task. As we will show in the experiment section, using the traditional weight-based regularization has only very marginal improvements in our task. The major reason is that the overfitting risk is mainly from the employment of the global information. Thus, we need to find effective solutions to reduce the overfitting risk from the global information. Otherwise the full power of the global information can not be released.

To deal with this problem, we propose two different regularization methods for reducing the overfitting risk of the added global information. We will shown in experiments that by regularizing the global information, we can simplify the problem, can further improve the accuracy, and at the same time can substantially speed up the training speed of the tagger (thus we are able to use even larger training data for gaining further improvement of the accuracy).

2 Sequential Tagging with Global Information

State-of-the-art statistical POS taggers are usually built upon sequence labeling models, which is lack of both intra- and inter-sentence global information. First, as a part of syntactic analysis, non-local dependencies among words widely exist in a given sentence. It is very hard to capture such intra-sentence information by a sequence model with very limited factorization. This problem is more serious for Chinese, which, as an isolating language, leverages very little explicit morphological information for classification. Second, most models are defined at the sentence-level, and similarity between words appearing in different sentences is not well explored. Such inter-sentence global information is also important, especially for unknown words.

We use the recently proposed structured prediction method, the Search-based Probabilistic Online Learning Algorithm (SAPO) (Sun, 2015), for the algorithmic implementation of sequential tagging in our task. The SAPO method is a probabilistic extension of the traditional perceptron model (Collins, 2002). The SAPO method searches the top-n output candidates, derives probabilities based on the searched candidates, and conduct fast online learning by updating the model weights. We choose SAPO because it is very easy to implement – the implementation is almost as simple as the Collins' perceptron. The training speed of SAPO is also as fast as Collins' perceptron. On the other hand, because it has probabilistic information, in many real-world tasks it is as accurate as the heavy probabilistic models like conditional random fields (CRF), as has been shown in (Sun, 2015).

2.1 Syntactically Generated Pseudo Training Data

Syntactic parsers usually employ complementary factorization models to taggers. For example, a graph-based dependency parser evaluates every word pair no matter how many words are in between them. Such factorization is more expressive for encoding inter-sentence global information. Previous work indicates that parser is better at tag prediction involves long-range dependencies than sequence models (Sun and Uszkoreit, 2012). Moreover, some ensemble learning techniques are able to capture the local information that can be well predicted by a tagger and the global information that can be obtained by a syntactic parser. However, it is unreasonable to employ a parser to produce POS tags if only POS information is needed.

It is a very general case that a complex model that is able to yield better prediction is very slow. To build NLP systems that are accurate and efficient, Petrov et al. (2010) introduced the uptraining technique. The same idea is also explored by Liang et al. (2008), but named as structure compilation. The main idea behind is produce large scale pseudo training data by applying a complex model. Such pseudo training data can be utilized to train a simple yet efficient model with better accuracy.

In our problem, we build a stacked model which combines the predictive power of a parser and a tagger. This complex model is employed as the pseudo corpus annotater. The goal is to train a better SAPO sequential labelling tagger which is efficient at test time. Given that the automatic annotation of words involving intra-sentence dependencies are better processed, the pseudo training data is able to enhance a sequential labelling tagger. Especially, we

extend the window size of the contextual features to capture such global information.

2.2 Word Clustering from Raw Data

Most POS tagging models are defined at the sentence-level, which makes inter-sentence global information sharing difficult. To capture distributional similarities among words appearing in different sentences, we employ the word clustering technique. A SAPO sequential labelling tagger is easy to be extended with arbitrary features and therefore suitable to explore additional features derived from other sources. Following (Sun and Uszkoreit, 2012), we use word clusters as substitutes for word forms to assist the POS tagger. We are relying on the ability of discriminative learning to explore informative features, which play a central role in enhancing the tagging model.

3 Regularization of Global Information

The global information is helpful to the improvement of the Chinese POS tagging task. On the other hand, in preliminary experiments we also find the global information is easy to suffer from overfitting. As we have discussed, this is because the global information is based on relatively complicated structures and long range dependencies. Such information from relatively complicated structures and long range dependencies are easier to be overfitting on the training data set.

The traditional regularization methods are focused on regularizing the model weights (weight-based regularization), such as L_2 and L_1 regularization over the model weights/parameters. However, in practice we find those traditional regularization methods performs poorly in our Chinese POS tagging task. The reason is that the overfitting is mainly from the global information and the complicated structures, and this type of overfitting can hardly be solved by regularizing the model weights (as we will show in experiments). To deal with this problem, we propose two different regularization methods for reducing the overfitting risk from the global information. We will shown in experiments that by regularizing the global information, we can simplify the problem, can further improve the accuracy, and at the same time can substantially speed up the training speed of the tagger (thus we are able to use even larger training data for gaining further improvement of the accuracy).

3.1 Structure Regularization of Global Information

To deal with the overfitting of the global information, we first employ a structure regularization technique to simplify the structures in Chinese POS tagging. 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. As we know, in the modeling of sequential labelling (e.g., like SAPO, CRF, and HMM), the tags of a sentence is structurally dependent, because the optimal sequence of tags are determined by optimizing the tags via modeling the dependencies among tags (usually by assuming a Markov dependence among the neighbouring tags). As a result, all the tags in a sentence are actually inter-dependent because of the local dependence connected one by one.

In our Chinese POS tagging task, we find this structural dependence is quite easy to suffer from overfitting, especially because we have employed global information by integrating the syntactic structural dependencies. Our system is built upon a sequential labelling setting with global information based on the syntactic structural dependencies. The global information represents an increase of the structural complexity, and this increase of the structural complexity can make the POS tagging system more easier to suffer from overfitting upon the training data. The theoretical effectiveness of structure regularization has been discussed in the pioneer work of (Sun, 2014a; Sun, 2014b). We follow this general idea to develop specific structure regularization algorithms for our Chinese POS tagging task, and our detailed implementation is different and novel compared with the prior work of structure regularization (Sun, 2014a; Sun, 2014b).

We cast the structure regularization as a very simple preprocessing step during the training of our SAPO sequential labelling tagger. 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 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. Since we cast this problem as a sequential labelling problem with global information, 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 **??**.

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 averaged new size (complexity of tags) of the produced new training samples. The exact value of α should be automatically tuned and determined by testing via cross validation or simply using the development data.

3.2 Regularized Importance Weighting of Global Information

For modeling global information based on long range syntactic dependencies, we automatically built the syntactically generated pseudo training data by employing syntactic parsers. For example, a graph-based dependency parser evaluates every word pair no matter how many words are in between them. Such factorization is more expressive for encoding inter-sentence global information. While this type of global information has the advantage of capturing long range dependencies, it is also quite easy to suffer from overfitting. Except the structure regularization method that we have discussed, an additional regularization method for reducing the overfitting risk of the global information is to find a balanced combination of the "local information" and the "global information". Here, "local information" means the original training data (i.e., the standard PTB training data), and the "global information" means our newly built syntactically generated pseudo training data.

In other words, by default the tagger treats all training samples with equal importance in the training phase. That means a sample in the original training set is as important as a sample in the syntactically generated pseudo training data. This has a potential problem because we have only 20K training samples in the original training set (local information), and we have more than 250K or even 2000K training samples in the syntactically generated pseudo training data (global information). In this sense, if we treat all the training samples equally important, the model is largely controlled by the global information. To deal with this problem, we need to regularize the global information by regularizing the importance weighting of the global information related training samples.

To realize this purpose, we attach a smaller importance factor of the global information related training samples. This can be done by modifying the objective function of the tagger, i.e., we can attach a "importance factor" for the syntactically generated pseudo training data in the objective function. In algorithmic implementation of this idea, it is even more simple. First, given a training sample, we need to calculate an update term following the algorithmic definition of the given sequential tagger.[1] Then, if the current training

[1]For example, for training a CRF model with stochastic gradient descent, an update term is a gradient of the given

sample is from the original training data (local information), we multiply the update term with a fixed factor of 1 (i.e., do nothing). Otherwise, if the current training sample is from the syntactically generated pseudo training data (global information), we multiply the update term with a importance factor of β. The importance factor β represents the importance regularization strength over the global information. The β has a value that is smaller than 1, and the exact value should be automatically tuned and determined by testing via cross validation or simply using the development data.

4 Conclusions

We investigate global information for improving Chinese POS tagging, including syntactically generated pseudo training data and word class information generated from large-scale raw data. We confirmed in experiments that those global information is very helpful to the Chinese POS tagging task.

Moreover, we find that the global information is very easy to suffer from overfitting, and this type of overfitting based on the global information can hardly be solved by using traditional regularization methods. We propose two simple regularization methods, structure regularization and regularized importance weighting, for regularizing the global information. By regularizing the global information, we can simplify the problem, can further improve the accuracy, and at the same time can substantially speed up the training speed of the tagger (thus our tagger can be scaled up to larger-scale data for further improvement). With this solution, we can control the overfitting risk from the global information and can fully release the power of the global information.

References

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–8.

training sample; for training a SAPO model, an update term is the probabilistic extension of the perceptron additive update term.

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*, pages 1216–1224, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Percy Liang, Hal Daumé, III, and Dan Klein. 2008. Structure compilation: trading structure for features. In *Proceedings of the 25th international conference on Machine learning*, pages 592–599, New York, NY, USA. ACM.

Slav Petrov, Pi-Chuan Chang, Michael Ringgaard, and Hiyan Alshawi. 2010. Uptraining for accurate deterministic question parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 705–713, Cambridge, MA, October. Association for Computational Linguistics.

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)*, pages 242–252, Jeju Island, Korea, July. Association for Computational Linguistics.

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

Xu Sun. 2014b. Structure regularization for structured prediction: Theories and experiments. In *Technical report, arXiv 1411.6243*.

Xu Sun. 2015. Towards shockingly easy structured classification: A search-based probabilistic online learning framework. *Technical report, arXiv:1503.08381*.

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.