Capturing Paradigmatic and Syntagmatic Lexical Relations: Towards Accurate Chinese Part-of-Speech Tagging

Weiwei Sun[†*] and **Hans Uszkoreit**[‡]
[†]Institute of Computer Science and Technology, Peking University
[†]Saarbrücken Graduate School of Computer Science
[†‡]Department of Computational Linguistics, Saarland University
[†‡]Language Technology Lab, DFKI GmbH
ws@pku.edu.cn, uszkoreit@dfki.de

Abstract

From the perspective of structural linguistics, we explore paradigmatic and syntagmatic lexical relations for Chinese POS tagging, an important and challenging task for Chinese language processing. Paradigmatic lexical relations are explicitly captured by word clustering on large-scale unlabeled data and are used to design new features to enhance a discriminative tagger. Syntagmatic lexical relations are implicitly captured by constituent parsing and are utilized via system combination. Experiments on the Penn Chinese Treebank demonstrate the importance of both paradigmatic and syntagmatic relations. Our linguistically motivated approaches yield a relative error reduction of 18% in total over a state-of-the-art baseline.

1 Introduction

In grammar, a part-of-speech (POS) is a linguistic category of words, which is generally defined by the syntactic or morphological behavior of the word in question. Automatically assigning POS tags to words plays an important role in parsing, word sense disambiguation, as well as many other NLP applications. Many successful tagging algorithms developed for English have been applied to many other languages as well. In some cases, the methods work well without large modifications, such as for German. But a number of augmentations and changes become necessary when dealing with highly inflected or agglutinative languages, as well as analytic languages, of which Chinese is the focus of this paper. The Chinese language is characterized by the lack of formal devices such as morphological tense and number that often provide important clues for syntactic processing tasks. While state-of-the-art tagging systems have achieved accuracies above 97% on English, Chinese POS tagging has proven to be more challenging and obtained accuracies about 93-94% (Tseng et al., 2005b; Huang et al., 2007, 2009; Li et al., 2011).

It is generally accepted that Chinese POS tagging often requires more sophisticated language processing techniques that are capable of drawing inferences from more subtle linguistic knowledge. From a linguistic point of view, meaning arises from the differences between linguistic units, including words, phrases and so on, and these differences are of two kinds: paradigmatic (concerning substitution) and syntagmatic (concerning positioning). The distinction is a key one in structuralist semiotic analysis. Both paradigmatic and syntagmatic lexical relations have a great impact on POS tagging, because the *value* of a word is determined by the two relations. Our error analysis of a state-of-the-art Chinese POS tagger shows that the lack of both paradigmatic and syntagmatic lexical knowledge accounts for a large part of tagging errors.

This paper is concerned with capturing paradigmatic and syntagmatic lexical relations to advance the state-of-the-art of Chinese POS tagging. First, we employ unsupervised word clustering to explore paradigmatic relations that are encoded in large-scale unlabeled data. The word clusters are then explicitly utilized to design new features for POS tagging. Second, we study the possible impact of syntagmatic relations on POS tagging by comparatively analyzing a (syntax-free) sequential tagging model

*This work is mainly finished when this author (corresponding author) was in Saarland University and DFKI.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 242–252,
Jeju, Republic of Korea, 8-14 July 2012.

and a (syntax-based) chart parsing model. Inspired by the analysis, we employ a full parser to implicitly capture syntagmatic relations and propose a *Bootstrap Aggregating* (Bagging) model to combine the complementary strengths of a sequential tagger and a parser.

We conduct experiments on the Penn Chinese Treebank and Chinese Gigaword. We implement a discriminative sequential classification model for POS tagging which achieves the state-of-the-art accuracy. Experiments show that this model are significantly improved by word cluster features in accuracy across a wide range of conditions. This confirms the importance of the paradigmatic relations. We then present a comparative study of our tagger and the Berkeley parser, and show that the combination of the two models can significantly improve tagging accuracy. This demonstrates the importance of the syntagmatic relations. Cluster-based features and the Bagging model result in a relative error reduction of 18% in terms of the word classification accuracy.

2 State-of-the-Art

2.1 Previous Work

Many algorithms have been applied to computationally assigning POS labels to English words, including hand-written rules, generative HMM tagging and discriminative sequence labeling. Such methods have been applied to many other languages as well. In some cases, the methods work well without large modifications, such as German POS tagging. But a number of augmentations and changes became necessary when dealing with Chinese that has little, if any, inflectional morphology. 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 93-94% (Tseng et al., 2005b; Huang et al., 2007, 2009; Li et al., 2011).

Both discriminative and generative models have been explored for Chinese POS tagging (Tseng et al., 2005b; Huang et al., 2007, 2009). Tseng et al. (2005a) introduced a maximum entropy based model, which includes morphological features for unknown word recognition. Huang et al. (2007) and Huang et al. (2009) mainly focused on the generative HMM models. To enhance a HMM model, Huang et al. (2007) proposed a re-ranking procedure to include extra morphological and syntactic features, while Huang et al. (2009) proposed a latent variable inducing model. Their evaluations on the Chinese Treebank show that Chinese POS tagging obtains an accuracy of about 93-94%.

2.2 Our Discriminative Sequential Model

According to the *ACL Wiki*, all state-of-the-art English POS taggers are based on discriminative sequence labeling models, including structure perceptron (Collins, 2002; Shen et al., 2007), maximum entropy (Toutanova et al., 2003) and SVM (Gimnez and Mrquez, 2004). A discriminative learner is easy to be extended with arbitrary features and therefore suitable to recognize more new words. Moreover, a majority of the POS tags are locally dependent on each other, so the Markov assumption can well captures the syntactic relations among words. Discriminative learning is also an appropriate solution for Chinese POS tagging, due to its flexibility to include knowledge from multiple linguistic sources.

To deeply analyze the POS tagging problem for Chinese, we implement a discriminative sequential model. A first order linear-chain CRF model is used to resolve the sequential classification problem. We choose the CRF learning toolkit *wapiti*[1] (Lavergne et al., 2010) to train models. In our experiments, we employ a feature set which draws upon information sources such as word forms and characters that constitute words. To conveniently illustrate, we denote a word in focus with a fixed window $w_{-2}w_{-1}ww_{+1}w_{+2}$, where w is the current token. Our features includes:

Word unigrams: w_{-2}, w_{-1}, w, w_{+1}, w_{+2};
Word bigrams: w_{-2}-w_{-1}, w_{-1}-w, w-w_{+1}, w_{+1}-w_{+2};
In order to better handle unknown words, we extract morphological features: character n-gram prefixes and suffixes for n up to 3.

2.3 Evaluation

2.3.1 Setting

Penn Chinese Treebank (CTB) (Xue et al., 2005) is a popular data set to evaluate a number of Chinese NLP tasks, including word segmentation (Sun and

[1] http://wapiti.limsi.fr/

Xu, 2011), POS tagging (Huang et al., 2007, 2009), constituency parsing (Zhang and Clark, 2009; Wang et al., 2006) and dependency parsing (Zhang and Clark, 2008; Huang and Sagae, 2010; Li et al., 2011). In this paper, we use CTB 6.0 as the labeled data for the study. The corpus was collected during different time periods from different sources with a diversity of topics. In order to obtain a representative split of data sets, we define the training, development and test sets following two settings. To compare our tagger with the state-of-the-art, we conduct an experiment using the data setting of (Huang et al., 2009). For detailed analysis and evaluation, we conduct further experiments following the setting of the CoNLL 2009 shared task. The setting is provided by the principal organizer of the CTB project, and considers many annotation details. This setting is more robust for evaluating Chinese language processing algorithms.

2.3.2 Overall Performance

Table 1 summarizes the per token classification accuracy (Acc.) of our tagger and results reported in (Huang et al., 2009). Huang et al. (2009) introduced a bigram HMM model with latent variables (*Bigram HMM-LA* in the table) for Chinese tagging. Compared to earlier work (Tseng et al., 2005a; Huang et al., 2007), this model achieves the state-of-the-art accuracy. Despite of simplicity, our discriminative POS tagging model achieves a state-of-the-art performance, even better.

System	Acc.
Trigram HMM (Huang et al., 2009)	93.99%
Bigram HMM-LA (Huang et al., 2009)	94.53%
Our tagger	94.69%

Table 1: Tagging accuracies on the test data (setting 1).

2.4 Motivating Analysis

For the following experiments, we only report results on the development data of the CoNLL setting.

2.4.1 Correlating Tagging Accuracy with Word Frequency

Table 2 summarizes the prediction accuracy on the development data with respect to the word frequency on the training data. To avoid overestimating the tagging accuracy, these statistics exclude all punctuations. From this table, we can see that words with low frequency, especially the out-of-vocabulary (OOV) words, are hard to label. However, when a word is very frequently used, its behavior is very complicated and therefore hard to predict. A typical example of such words is the language-specific function word "的." This analysis suggests that a main topic to enhance Chinese POS tagging is to bridge the gap between the infrequent words and frequent words.

Freq.	Acc.
0	83.55%
1-5	89.31%
6-10	90.20%
11-100	94.88%
101-1000	96.26%
1001-	93.65%

Table 2: Tagging accuracies relative to word frequency.

2.4.2 Correlating Tagging Accuracy with Span Length

A word projects its grammatical property to its maximal projection and it syntactically governs all words under the span of its maximal projection. The words under the span of current token thus reflect its syntactic behavior and good clues for POS tagging. Table 3 shows the tagging accuracies relative to the length of the spans. We can see that with the increase of the number of words governed by the token, the difficulty of its POS prediction increase. This analysis suggests that syntagmatic lexical relations plays a significant role in POS tagging, and sometimes words located far from the current token affect its tagging much.

Len.	Acc.
1-2	93.79%
3-4	93.39%
5-6	92.19%
7-	94.18%

Table 3: Tagging accuracies relative to span length.

3 Capturing Paradigmatic Relations via Word Clustering

To bridge the gap between high and low frequency words, we employ word clustering to acquire

the knowledge about paradigmatic lexical relations from large-scale texts. Our work is also inspired by the successful application of word clustering to named entity recognition (Miller et al., 2004) and dependency parsing (Koo et al., 2008).

3.1 Word Clustering

Word clustering is a technique for partitioning sets of words into subsets of syntactically or semantically similar words. It is a very useful technique to capture paradigmatic or substitutional similarity among words.

3.1.1 Clustering Algorithms

Various clustering techniques have been proposed, some of which, for example, perform automatic word clustering optimizing a maximum-likelihood criterion with iterative clustering algorithms. In this paper, we focus on distributional word clustering that is based on the assumption that words that appear in similar contexts (especially surrounding words) tend to have similar meanings. They have been successfully applied to many NLP problems, such as language modeling.

Brown Clustering Our first choice is the bottom-up agglomerative word clustering algorithm of (Brown et al., 1992) which derives a hierarchical clustering of words from unlabeled data. This algorithm generates a hard clustering – each word belongs to exactly one cluster. The input to the algorithm is sequences of words $w_1, ..., w_n$. Initially, the algorithm starts with each word in its own cluster. As long as there are at least two clusters left, the algorithm merges the two clusters that maximizes the quality of the resulting clustering. The quality is defined based on a class-based bigram language model as follows.

$$P(w_i|w_1, ...w_{i-1}) \approx p(C(w_i)|C(w_{i-1}))p(w_i|C(w_i))$$

where the function C maps a word w to its class $C(w)$. We use a publicly available package[2] (Liang et al., 2005) to train this model.

MKCLS Clustering We also do experiments by using another popular clustering method based on the exchange algorithm (Kneser and Ney, 1993). The objective function is maximizing the likelihood $\prod_{i=1}^{n} P(w_i|w_1, ..., w_{i-1})$ of the training data given a partially class-based bigram model of the form

$$P(w_i|w_1, ...w_{i-1}) \approx p(C(w_i)|w_{i-1})p(w_i|C(w_i))$$

We use the publicly available implementation MKCLS[3] (Och, 1999) to train this model.

We choose to work with these two algorithms considering their prior success in other NLP applications. However, we expect that our approach can function with other clustering algorithms.

3.1.2 Data

Chinese Gigaword is a comprehensive archive of newswire text data that has been acquired over several years by the Linguistic Data Consortium (LDC). The large-scale unlabeled data we use in our experiments comes from the Chinese Gigaword (LDC2005T14). We choose the Mandarin news text, i.e. Xinhua newswire. This data covers all news published by Xinhua News Agency (the largest news agency in China) from 1991 to 2004, which contains over 473 million characters.

3.1.3 Pre-processing: Word Segmentation

Different from English and other Western languages, Chinese is written without explicit word delimiters such as space characters. To find the basic language units, i.e. words, segmentation is a necessary pre-processing step for word clustering. Previous research shows that character-based segmentation models trained on labeled data are reasonably accurate (Sun, 2010). Furthermore, as shown in (Sun and Xu, 2011), appropriate string knowledge acquired from large-scale unlabeled data can significantly enhance a supervised model, especially for the prediction of out-of-vocabulary (OOV) words. In this paper, we employ such supervised and semi-supervised segmenters[4] to process raw texts.

3.2 Improving Tagging with Cluster Features

Our discriminative sequential tagger is easy to be extended with arbitrary features and therefore suitable to explore additional features derived from other

[2] http://cs.stanford.edu/~pliang/software/brown-cluster-1.2.zip

[3] http://code.google.com/p/giza-pp/

[4] http://www.coli.uni-saarland.de/~wsun/ccws.tgz

sources. We propose to use of word clusters as substitutes for word forms to assist the POS tagger. We are relying on the ability of the discriminative learning method to explore informative features, which play a central role in boosting the tagging performance. 5 clustering-based uni/bi-gram features are added: w_{-1}, w, w_{+1}, w_{-1}_w, w_w_{+1}.

3.3 Evaluation

Features	Data	Brown	MKCLS
Baseline	CoNLL	94.48%	
+c100	+1991-1995(S)	94.77%	94.83%
+c500	+1991-1995(S)	94.84%	94.93%
+c1000	+1991-1995(S)	- -	94.95%
+c100	+1991-1995(SS)	94.90%	94.97%
+c500	+1991-1995(SS)	94.94%	94.88%
+c1000	+1991-1995(SS)	94.89%	94.94%
+c100	+1991-2000(SS)	94.82%	94.93%
+c500	+1991-2000(SS)	94.92%	94.99%
+c1000	+1991-2000(SS)	94.90%	95.00%
+c100	+1991-2004(SS)	- -	94.87%
+c500	+1991-2004(SS)	- -	**95.02%**
+c1000	+1991-2004(SS)	- -	94.97%

Table 4: Tagging accuracies with different features. *S*: supervised segmentation; *SS*: semi-supervised segmentation.

Table 4 summarizes the tagging results on the development data with different feature configurations. In this table, the symbol "+" in the *Features* column means current configuration contains both the baseline features and new cluster-based features; the number is the total number of the clusters; the symbol "+" in the *Data* column means which portion of the Gigaword data is used to cluster words; the symbol "S" and "SS" in parentheses denote (s)upervised and (s)emi-(s)upervised word segmentation. For example, "+1991-2000(S)" means the data from 1991 to 2000 are processed by a supervised segmenter and used for clustering. From this table, we can clearly see the impact of word clustering features on POS tagging. The new features lead to substantial improvements over the strong supervised baseline. Moreover, these increases are consistent regardless of the clustering algorithms. Both clustering algorithms contributes to the overall performance equivalently. A natural strategy for extending current experiments is to include both clustering results together, or to include more than one cluster granularity. However, we find no further improvement. For each clustering algorithm, there are not much differences among different sizes of the total clustering numbers. When a comparable amount of unlabeled data (five years' data) is used, the further increase of the unlabeled data for clustering does not lead to much changes of the tagging performance.

3.4 Learning Curves

Size	Baseline	+Cluster
4.5K	90.10%	91.93%
9K	92.91%	93.94%
13.5K	93.88%	94.60%
18K	94.24%	94.77%

Table 5: Tagging accuracies relative to sizes of training data. Size=#sentences in the training corpus.

We do additional experiments to evaluate the effect of the derived features as the amount of labeled training data is varied. We also use the "+c500(MKCLS)+1991-2004(SS)" setting for these experiments. Table 5 summarizes the accuracies of the systems when trained on smaller portions of the labeled data. We can see that the new features obtain consistent gains regardless of the size of the training set. The error is reduced significantly on all data sets. In other words, the word cluster features can significantly reduce the amount of labeled data required by the learning algorithm. The relative reduction is greatest when smaller amounts of the labeled data are used, and the effect lessens as more labeled data is added.

3.5 Analysis

Word clustering derives paradigmatic relational information from unlabeled data by grouping words into different sets. As a result, the contribution of word clustering to POS tagging is two-fold. On the one hand, word clustering captures and abstracts context information. This new linguistic *knowledge* is thus helpful to better correlate a word in a certain context to its POS tag. On the other hand, the clustering of the OOV words to some extent fights the sparse data problem by correlating an OOV word with in-vocabulary (IV) words through their classes. To evaluate the two contributions of the word clustering, we limit entries of the clustering lexicon to only contain IV words, i.e. words appearing in the training corpus. Using this constrained lexicon,

we train a new "+c500(MKCLS)+1991-2004(SS)" model and report its prediction power in Table 6. The gap between the baseline and *+IV clustering* models can be viewed as the contribution of the first effect, while the gap between the *+IV clustering* and *+All clustering* models can be viewed as the second contribution. This result indicates that the improved predictive power partially comes from the new interpretation of a POS tag through a clustering, and partially comes from its memory of OOV words that appears in the unlabeled data.

	Baseline	+IV Clustering	+All clustering
Acc.	94.48%	94.70%(↑0.22)	95.02%(↑0.32)

Table 6: Tagging accuracies with IV clustering.

Table 7 shows the recall of OOV words on the development data set. Only the word types appearing more than 10 times are reported. The recall of all OOV words are improved, especially of proper nouns (NR) and common verbs (VV). Another interesting fact is that almost all of them are content words. This table is also helpful to understand the impact of the clustering information on the prediction of OOV words.

4 Capturing Syntagmatic Relations via Constituency Parsing

Syntactic analysis, especially the full and deep one, reflects syntagmatic relations of words and phrases of sentences. We present a series of empirical studies of the tagging results of our *syntax-free* sequential tagger and a *syntax-based* chart parser[5], aiming at illuminating more precisely the impact of information about phrase-structures on POS tagging. The analysis is helpful to understand the role of syntagmatic lexical relations in POS prediction.

4.1 Comparing Tagging and PCFG-LA Parsing

The majority of the state-of-the-art constituent parsers are based on generative PCFG learning, with lexicalized (Collins, 2003; Charniak, 2000) or latent annotation (PCFG-LA) (Matsuzaki et al., 2005; Petrov et al., 2006; Petrov and Klein, 2007) refinements. Compared to lexicalized parsers, the PCFG-LA parsers leverages on an automatic procedure to learn refined grammars and are therefore more robust to parse non-English languages that are not well studied. For Chinese, a PCFG-LA parser achieves the state-of-the-art performance and defeat many other types of parsers (Zhang and Clark, 2009). For full parsing, the Berkeley parser[6], an open source implementation of the PCFG-LA model, is used for experiments. Table 8 shows their overall and detailed performance.

	#Words	Baseline	+Clustering	Δ
AD	21	33.33%	42.86%	<
CD	249	97.99%	98.39%	<
JJ	86	3.49%	26.74%	<
NN	1028	91.05%	91.34%	<
NR	863	81.69%	88.76%	<
NT	25	60.00%	68.00%	<
VA	15	33.33%	53.33%	<
VV	402	67.66%	72.39%	<

Table 7: The tagging recall of OOV words.

4.1.1 Content Words vs. Function Words

Table 8 gives a detailed comparison regarding different word types. For each type of word, we report the accuracy of both solvers and compare the difference. The majority of the words that are better labeled by the tagger are content words, including nouns(NN, NR, NT), numbers (CD, OD), predicates (VA, VC, VE), adverbs (AD), nominal modifiers (JJ), and so on. In contrast, most of the words that are better predicted by the parser are function words, including most particles (DEC, DEG, DER, DEV, AS, MSP), prepositions (P, BA) and coordinating conjunction (CC).

4.1.2 Open Classes vs. Close Classes

POS can be divided into two broad supercategories: closed class types and open class types. Open classes accept the addition of new morphemes (words), through such processes as compounding, derivation, inflection, coining, and borrowing. On the other hand closed classes are those that have relatively fixed membership. For example, nouns and verbs are open classes because new nouns and verbs are continually coined or borrowed from other languages, while *DEC*/*DEG* are two closed classes because only the function word "的" is assigned to

[5] Both the tagger and the parser are trained on the same portion from CTB.

[6] http://code.google.com/p/berkeleyparser/

Parser<Tagger		Parser>Tagger	
♠ AD	94.15<94.71	♡ AS	98.54>98.44
♠ CD	94.66<97.52	♡ BA	96.15>92.52
CS	91.12<92.12	♡ CC	93.80>90.58
ETC	99.65<100.0	♡ DEC	85.78>81.22
♠ JJ	81.35<84.65	♡ DEG	88.94>85.96
LB	91.30<93.18	♡ DER	80.95>77.42
LC	96.29<97.08	♡ DEV	84.89>74.78
M	95.62<96.94	DT	98.28>98.05
♠ NN	93.56<94.95	♡ MSP	91.30>90.14
♠ NR	89.84<95.07	♡ P	96.26>94.56
♠ NT	96.70<97.26	VV	91.99>91.87
♠ OD	81.06<86.36		
PN	98.10<98.15		
SB	95.36<96.77		
SP	61.70<68.89		
♠ VA	81.27<84.25	Overall	
♠ VC	95.91<97.67	Tagger:	94.48%
♠ VE	97.12<98.48	Parser:	93.69%

Table 8: Tagging accuracies of relative to word classes.

them. The discriminative model can conveniently include many features, especially features related to the word formation, which are important to predict words of open classes. Table 9 summarizes the tagging accuracies relative to IV and OOV words. On the whole, the Berkeley parser processes IV words slightly better than our tagger, but processes OOV words significantly worse. The numbers in this table clearly shows the main weakness of the Berkeley parser is the the predictive power of the OOV words.

	IV	OOV
Tagger	95.22%	81.59%
Parser	95.38%	64.77%

Table 9: Tagging accuracies of the IV and OOV words.

4.1.3 Local Disambiguation vs. Global Disambiguation

Closed class words are generally function words that tend to occur frequently and often have structuring uses in grammar. These words have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with other words within a sentence. They signal the structural relationships that words have to one another and are the glue that holds sentences together. Thus, they serve as important elements to the structures of sentences. The disambiguation of these words normally require more syntactic clues, which is very hard and inappropriate for a sequential tagger to capture. Based on global grammatical inference of the whole sentence, the full parser is relatively good at dealing with structure related ambiguities.

We conclude that discriminative sequential tagging model can better capture local syntactic and morphological information, while the full parser can better capture global syntactic structural information. The discriminative tagging model are limited by the Markov assumption and inadequate to correctly label structure related words.

4.2 Enhancing POS Tagging via Bagging

The diversity analysis suggests that we may improve parsing by simply combining the tagger and the parser. Bootstrap aggregating (Bagging) is a machine learning ensemble meta-algorithm to improve classification and regression models in terms of stability and classification accuracy (Breiman, 1996). It also reduces variance and helps to avoid overfitting. We introduce a Bagging model to integrate different POS tagging models. In the training phase, given a training set D of size n, our model generates m new training sets D_i of size $63.2\% \times n$ by sampling examples from D without replacement. Namely no example will be repeated in each D_i. Each D_i is separately used to train a tagger and a parser. Using this strategy, we can get $2m$ *weak* solvers. In the tagging phase, the $2m$ models outputs $2m$ tagging results, each word is assigned one POS label. The final tagging is the voting result of these $2m$ labels. There may be equal number of different tags. In this case, our system prefer the first label they met.

4.3 Evaluation

We evaluate our combination model on the same data set used above. Figure 1 shows the influence of m in the Bagging algorithm. Because each new data set D_i in bagging algorithm is generated by a random procedure, the performance of all Bagging experiments are not the same. To give a more stable evaluation, we repeat 5 experiments for each m and show the averaged accuracy. We can see that the Bagging model taking both sequential tagging and chart parsing models as basic systems outperform the baseline systems and the Bagging model taking either model in isolation as basic systems. An

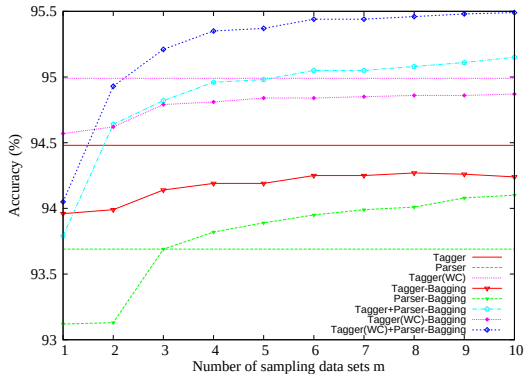

Figure 1: Tagging accuracies of Bagging models. *Tagger-Bagging* and *Tagger(WC)-Bagging* means that the Bagging system built on the tagger with and without word clusters. *Parser-Bagging* is named in the same way. *Tagger+Paser-Bagging* and *Tagger(WC)+Paser-Bagging* means that the Bagging systems are built on both tagger and parser.

interesting phenomenon is that the Bagging method can also improve the parsing model, but there is a decrease while only combining taggers.

5 Combining Both

We have introduced two separate improvements for Chinese POS tagging, which capture different types of lexical relations. We therefore expect further improvement by combining both enhancements, since their contributions to the task is different. We still use a Bagging model to integrate the discriminative tagger and the Berkeley parser. The only difference between current experiment and previous experiment is that the sub-tagging models are trained with help of word clustering features. Figure 1 also shows the performance of the new Bagging model on the development data set. We can see that the improvements that come from two ways, namely capturing syntagmatic and paradigmatic relations, are not much overlapping and the combination of them gives more.

Table 10 shows the performance of different systems evaluated on the test data. The final result is remarkable. The word clustering features and the Bagging model result in a relative error reduction of 18% in terms of the classification accuracy. The significant improvement of the POS tagging also help successive language processing. Results in Table 11 indicate that the parsing accuracy of the Berkeley parser can be simply improved by inputting the Berkeley parser with the POS Bagging results. Although the combination with a syntax-based tagger is very effective, there are two weaknesses: (1) a syntax-based model relies on linguistically rich syntactic annotations that are not easy to acquire; (2) a syntax-based model is computationally expensive which causes efficiency difficulties.

Systems	Acc.
Baseline	94.33%
Tagger(WC)	94.85%
Tagger+Parser($m = 15$)	94.96%
Tagger(WC)+Parser($m = 15$)	95.34%

Table 10: Tagging accuracies on the test data (CoNLL).

Tagger	LP	LR	F
Berkeley	82.71%	80.57%	81.63
Bagging($m = 15$)	82.96%	81.44%	82.19

Table 11: Parsing accuracies on the test data. (CoNLL)

6 Conclusion

We hold a view of structuralist linguistics and study the impact of paradigmatic and syntagmatic lexical relations on Chinese POS tagging. First, we harvest word partition information from large-scale raw texts to capture paradigmatic relations and use such knowledge to enhance a supervised tagger via feature engineering. Second, we comparatively analyze syntax-free and syntax-based models and employ a Bagging model to integrate a sequential tagger and a chart parser to capture syntagmatic relations that have a great impact on non-local disambiguation. Both enhancements significantly improve the state-of-the-art of Chinese POS tagging. The final model results in an error reduction of 18% over a state-of-the-art baseline.

Acknowledgement

This work is mainly finished when the first author was in Saarland University and DFKI. At that time, this author was funded by DFKI and German Academic Exchange Service (DAAD). While working in Peking University, the first author is supported by NSFC (61170166) and National High-Tech R&D Program (2012AA011101).

References

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479. URL http://portal.acm.org/citation.cfm?id=176313.176316.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W02-1001.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

Jesús Giménez and Lluís Màrquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 43–46.

Liang Huang and Kenji Sagae. 2010. Dynamic programming for linear-time incremental parsing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1077–1086. Association for Computational Linguistics, Uppsala, Sweden. URL http://www.aclweb.org/anthology/P10-1110.

Zhongqiang Huang, Vladimir Eidelman, and Mary Harper. 2009. 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*, pages 213–216. Association for Computational Linguistics, Boulder, Colorado. URL http://www.aclweb.org/anthology/N/N09/N09-2054.

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)*, pages 1093–1102. Association for Computational Linguistics, Prague, Czech Republic. URL http://www.aclweb.org/anthology/D/D07/D07-1117.

Reinhard Kneser and Hermann Ney. 1993. Improved clustering techniques for class-based statistical language modeling. In *In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603. Association for Computational Linguistics, Columbus, Ohio. URL http://www.aclweb.org/anthology/P/P08/P08-1068.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. pages 504–513. URL http://www.aclweb.org/anthology/P10-1052.

Zhenghua Li, Min Zhang, Wanxiang Che, Ting Liu, Wenliang Chen, and Haizhou Li. 2011. Joint models for Chinese pos tagging and dependency parsing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1180–1191. Association for Computational Linguistics, Edinburgh, Scotland, UK. URL http://www.aclweb.org/anthology/D11-1109.

Percy Liang, Michael Collins, and Percy Liang. 2005. Semi-supervised learning for natural language. In *Master' s thesis, MIT*.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Probabilistic cfg with latent annotations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 75–82. Association for Computational Linguistics, Stroudsburg,

PA, USA. URL http://dx.doi.org/10.3115/1219840.1219850.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 337–342. Association for Computational Linguistics, Boston, Massachusetts, USA.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 71–76. Association for Computational Linguistics, Stroudsburg, PA, USA. URL http://dx.doi.org/10.3115/977035.977046.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics, Sydney, Australia.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411. Association for Computational Linguistics, Rochester, New York.

Libin Shen, Giorgio Satta, and Aravind Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 760–767. Association for Computational Linguistics, Prague, Czech Republic. URL http://www.aclweb.org/anthology/P07-1096.

Weiwei Sun. 2010. Word-based and character-based word segmentation models: Comparison and combination. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1211–1219. Coling 2010 Organizing Committee, Beijing, China. URL http://www.aclweb.org/anthology/C10-2139.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese word segmentation using unlabeled data. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 970–979. Association for Computational Linguistics, Edinburgh, Scotland, UK. URL http://www.aclweb.org/anthology/D11-1090.

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 the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180. Association for Computational Linguistics, Stroudsburg, PA, USA. URL http://dx.doi.org/10.3115/1073445.1073478.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005a. A conditional random field word segmenter. In *In Fourth SIGHAN Workshop on Chinese Language Processing*.

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

Mengqiu Wang, Kenji Sagae, and Teruko Mitamura. 2006. A fast, accurate deterministic parser for Chinese. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 425–432. Association for Computational Linguistics, Sydney, Australia. URL http://www.aclweb.org/anthology/P06-1054.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The penn Chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Yue Zhang and Stephen Clark. 2008. A tale of two parsers: Investigating and combining graph-based

and transition-based dependency parsing. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 562–571. Association for Computational Linguistics, Honolulu, Hawaii. URL http://www.aclweb.org/anthology/D08-1059.

Yue Zhang and Stephen Clark. 2009. Transition-based parsing of the Chinese treebank using a global discriminative model. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 162–171. Association for Computational Linguistics, Paris, France. URL http://www.aclweb.org/anthology/W09-3825.