Improved Chinese Parsing Using Named Entity Cue

Dongchen Li, Xiantao Zhang and Xihong Wu
Key Laboratory of Machine Perception and Intelligence,
Speech and Hearing Research Center,
Peking University, China
{lidc,Zhangxt,wxh}@cis.pku.edu.cn

Abstract

Parsing and named entity recognition are two standalone techniques in natural language processing community. We expect that these two types of annotations should provide useful information to each other, and that modeling them jointly should improve performance and produce consistent outputs. Employing more fine-grained named entity annotations helps to parse complex named entity structures correctly. Thus, we integrate parsing and named entity recognition in a unified framework: 1. Through a joint representation of syntactic and named entity structures, we annotate named entity information to Penn Chinese Treebank5.0 (CTB5.0); 2. We annotate the nested structures for all nested named entities; 3. A latent annotation probabilistic context-free grammar (PCFGLA) model is trained on the data with joint representation. Experiment results demonstrate the mutual benefits for both Chinese parsing and named entities recognition tasks.

1 Why Exploit Named Entity Cue for Chinese Parsing?

Chinese parsing and named entity recognition are two basic Chinese NLP technologies. They play an important role in the Chinese information extraction, machine translation and question answering systems.

However, to the best of our knowledge, previous researches generally regard them as two standalone processes. One of the reasons is that the Treebank for training a parser has not been annotated with adequate named entity information. We argue that it will be beneficial to utilize named entity cue in parsing. Because one of the main difficulties in parsing Chinese is bracketing phrases with complex structure, and many complex phrases are named entities.

In Chinese there are a large number of named entities. Named entities (NEs) can be generally divided into three types: entity names, temporal expressions, and number expressions. They are “unique identifiers” of entities (organizations, persons, locations), time (date, times), and quantities (monetary values, percentages). According to Chinese Treebank fifth edition (CTB5.0) (Xue et al., 2002), every sentence contains over 1.5 entity names. Table 1 shows the distribution of these named entities in CTB5.0.

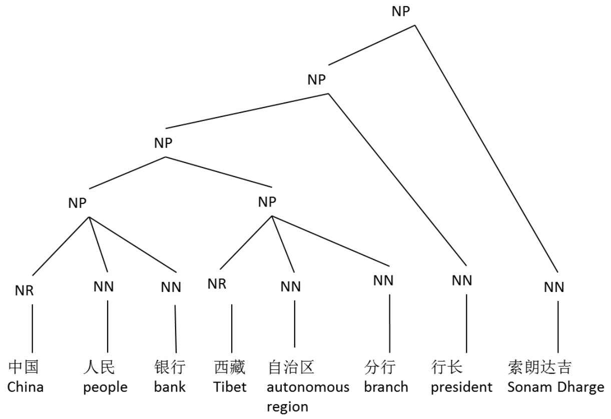


Figure 1: A named entity example with complex structures.

Different types of named entity phrases have their distinct structure patterns. However, all noun phrases including named entities get an identical label, say, noun phrase (NP). Computational processing of Chinese is typically based on the coarse syntactic tags. For example, in Figure 1, the structures of 中国人民银行“People’s Bank of China” and 中国人民银行西藏自治区分行行长索朗达吉“Sonam Dharge, the president of the Tibet Autonomous Region branch of

NE-Types	SubTypes	Description	Percent	Example
Entity Names	GPE	geographical / social / political entities	44.95	上海“Shanghai”, 山东省“Shandong Province”
	PERSON	person	29.34	理查德・尼克松 “Richard Nixon”
	ORG	organization	21.62	深圳市教育局 “Shenzhen Education Bureau”
	LOC	location(non-GPE locations)	8.67	淮河“Huai River”
Temporal expressions	DATE	date	14.74	一九九九年“the year of 1999”
	TIME	time	0.44	12时“12:00”
Number Expressions	NUM	number	52.58	172, 1.5
	ORD	ordinal number	6.78	第一“first”
	FRACTION	traction	4.06	百分之七十“70%”
	CODE	code	2.28	AK-47
	EVENT	event	0.83	五四“May Fourth Movement”
	TEMPERATURE	temperature	0.10	十二度“12 ℃”
	RATIO	ratio, score	0.08	0:05
	TEL	telephone number	0.05	23482192
	MONEY	money	0.02	三百万元“three million Yuan”

Table 1: The distribution of the named entity types (#sentences = 18789)

the People’s Bank of China” is quite different, but they get the identical label NP in CTB. A parser trained on these annotations is messy and hard to discriminate these complex structures correctly. Much work has illustrated that training the parser with manually annotated fine-grained labels and structures could help disambiguate parsing structure and improve parsing accuracy (Li, 2011; Li and Wu, 2012).

Thus, it is necessary to introduce these named entities in syntactic structure and integrate their recognition in the parsing process.

We integrate syntactic and named entity information in a unified framework through a joint representation. We add these named entity annotations into the syntactic structures in CTB5.0, with special care for nested named entity. Then we validate our annotations in parsing and named entity recognition tasks. This joint representation improves Chinese parsing accuracy significantly. Furthermore, the accuracies of the named entity recognition of our joint model outperform CRF-based NER system.

The rest of this paper is organized as follows. Section 2.1 reviews previously established Chinese Treebank (Penn Chinese Treebank) and Chinese corpus annotated with named entities (OntoNotes). Section 3 represents our joint representation of syntactic structures and named entities. In section 4 we perform experiments to illustrate the effectiveness of our joint representation.

2 Related Work

Penn Chinese Treebank (CTB) is the most widely used treebank for parsing Chinese. OntoNotes is a corpus annotated with both syntactic structure and named entities. We first review the annotations in these two corpora. Then, a brief introduction of Chinese parsing on character-level is given. Finally, we reviews the previous work on utilizing named entity cue in parsing.

Length of word	#NEs	#All	Percent
1	10276	166881	6.16
2	21843	222539	9.82
3	13588	30436	44.64
4	2532	6287	40.27
5	2300	2454	93.72
6	704	772	91.19
7	283	325	87.08
8	283	307	92.18
9	83	103	80.58
10	32	38	84.21
11	14	16	87.5
12	2	4	50
13	5	6	83.33

Table 2: Statistics of NEs' percent in different words' length

2.1 Penn Chinese Treebank and OntoNotes

CTB is a segmented, part-of-speech tagged, and fully bracketed corpus that currently has 500 thousand words (over 824K Chinese characters). There are totally 890 files in CTB5.0.

Parsing of Chinese is typically based on coarse part-of-speech tags and syntactic tags in CTB. In CTB, named entity phrase is simply labeled as a noun phrase (NP) without distinction of their diverse types (some of them may be labeled with an extra function tag P-N). Similarly, named entity words are simply labeled as a proper noun (NR), cardinal number (CD), ordinal number (OD) or temporal noun (NT), and they correspond to words in the parse trees without annotation of their internal word structure.

OntoNotes Release 4.0 (LDC2011T03) is a large, manually annotated corpus that contains various text genres and annotations (Hovy et al., 2006). It is also a corpus with annotation of entity names in Chinese. It contains 403 files which are also in CTB5.0, including the test set and development set in the standard parsing evaluation setup. Entity names in OntoNotes4.0 are annotated with 18 types of entity names, including PERSON, ORGANIZATION, GPE, LOC, PRODUCT and so on.

Many named entities contain other named entities inside them. However, works on named entity recognition (NER) and the annotation of OntoNotes have almost entirely ignored nested entities and instead chosen to focus on the outermost entities.

2.2 Parsing

Most high-performance parsers is based on probabilistic context-free grammars (PCFGs). They all refine grammar labels to capture more syntactic characteristic, ranging from full lexicalization and intricate smoothing (Collins, 1999; Charniak, 2000) to category refinement (Johnson, 1998; Klein and Manning, 2003). Latent annotation probabilistic context-free grammar (PCFG-LA) method in Matsuzaki et al. (2005) and Petrov and Klein (2007) automatically refines syntactic and lexical tags in an unsupervised manner, and has achieved state-of-the-art performance on both English and Chinese.

In recent years, there has been much work on character-level Chinese parsing. Qian and Liu (2012) trained three individual models of Chinese segmentation, POS tagging and Parsing separately during training, and incorporated them together in a discriminative framework. Zhang et al. (2013) integrated character-structure features in the joint model based on the discriminative shift-reduce parser of Zhang and Clark (2009) and Zhang and Clark (2011)Zhang and Clark (2009; 2011).

In spite of the convenience of its totally automatic learning process, the main defect of the latent factor models lies in that the training process is completely data-driven and suffers from data sparseness. To alleviate this problem, we leverage named entity cue, in the form of explicit annotation.

2.3 Named Entity Cue in Parsing

There is a large body of work on parsing and named entity recognition (Bikel and Chiang, 2000; Sekine and Nobata, 2004; Klementiev and Roth, 2006; Singh et al., 2010) separately. The sequence labeling approach has been shown to perform well on the task of Chinese NER (Chen et al., 2006; Yu et al., 2008). Finkel and Manning (2009a) and Finkel and Manning (2009b) paid special attention to the entity names in paring English. They gave a joint NER and parsing model with a discriminative parser, and improved accuracy for both tasks. We take advantage of named entity cue in character-level Chinese parsing, and further exploiting nested named entities in parsing.

Some existing work investigates the number expressions in parsing. Harper and Huang (2009) addressed this issue for achieving better parsing performance. Our work is not to verbalize sequences of digits; we annotate the entire constituent with fine label, such as DATE, NUM, TIME, FRACTION.

3 Our Approach

However, the completely data-driven state-split approach is prone to overfit the training data. Because the training data is always extremely sparse, and the automatically split categories might not be adequate. To improve parsing accuracy, Li (2011) manually annotated the internal structure of words, /citeli2012conjuncting manually annotated fine-grained labels for function words.

In our approach, all these types of named entity information are annotated to CTB5.0 through a joint representation in both word-level and character-level. Then we train a PCFG-LA parser on the corpus, and validate that named entity cue helps to improve parsing and NER accuracy simultaneously.

3.1 Named Entity Representation in Syntactic Tree

We argue that syntactic information and named entity information are mutual beneficial, so we enrich the annotations of the parse tree with fine-grained named entity labels to achieve the joint representation.

It is an important issue of how to define the types of Named entities. OntoNotes Release 4.0 (LDC2011T03) has annotated eighteen types of entity names. Some of these entity types do not occur frequently and are not always useful in practice, such as `works of art`, `product` and `law`, so we discard them in this study. In addition, we annotate the types of code, ratio and tel. All the named entity types are explained in Table 1.

There are totally 890 files in CTB5.0, and 403 of them have already been annotated with entity names in OntoNotes4.0. The test set and development set are setup as in the standard parsing evaluation. We annotated the left 487 files with previously mentioned types of named entities following the guideline of OntoNotes4.0.

3.2 Nested Named Entities Annotations

One of the main challenges for named entity recognition task is dealing with nested named entities. For example, Figure 1 contains nested named entities 中国人民银行西藏自治区分行“the Tibet Autonomous Region branch of the People’s Bank of China”, 中国人民银行“the People’s Bank of China”, and 索朗达吉“Sonam Dharge”. Tradition sequence labeling methods, such as CRF, treat the text as a linear sequence and have great difficulty in handling nested named entities, if not impossible.

We adopt a novel solution to explicitly represent nested named entities naturally in the syntactic structure. Nested named entities are exhaustively labelled in the syntactic tree structure, and each corresponds to one node in the tree.

Next, we will discuss the annotation process in detail. We refine the label of named entities’ components. As shown in Figure 1, 中国人民银行西藏自治区分行“People’s Bank of China branch of the Tibet Autonomous Region” is labeled as “NP_ORG”, and its two children in the tree are also labeled as “NP_ORG”. All the words’ structures are not changed; we just add a finer label to replace the original coarse label.

Further, we annotate the internal structure of a word that represents a nested named entity. There are three types of nested named entities: GPE, PERSON and temporal expression. We handle them respectively as follows.

For GPE, we split the GPE name and its geographical unit apart in a tree structure. This annotation style has the advantage of generalizing the common GPE composition structure. For example, 深圳市教育局“Shenzhen Education Bureau” is a ORG, but 深圳“Shenzhen” and 深圳市“Shenzhen city” are both GPE. The character 市“city” will obtain a special label. [1] In this case, we get a derivation which includes GPE → GPE GPEend. The experiment results in the next section show that the parser benefits a lot from this derivation. This example is shown in Figure 2.

We also distinguish the Chinese and foreign name by the entity name labels NR_PERSONF (Foreign

[1] When annotating the internal word structure, We do not need to distinguish an original word (e.g., Shenzhen City: NR_GPE) from an internal sub-word (e.g., Shenzhen: NR_GPE) explicitly. Because the internal sub-word can always be located by the geographical unit which is tagged by ”end”.

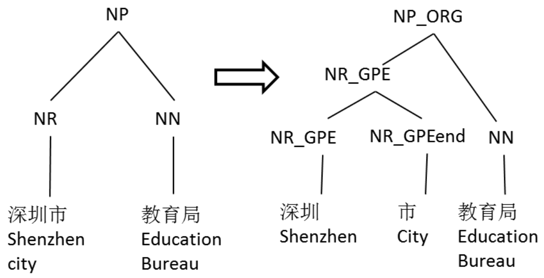


Figure 2: An example annotation for the phrase 深圳市教育局“Shenzhen Education Bureau”

Person Name) and NR_PERSONC (Chinese Person Name). It is obvious that a name containing the character ‘ • ’ is a foreign name. Using this cue, it is easy to recognize the foreign names. See Figure 3 for an illustration.

For temporal expressions, the nested structure is bracketed into number expressions and temporal unit. For instance, the word 十五日“the 15th day in a month” will be split with 十五-NUM and 日-Day. Figure 4 gives a detailed example.

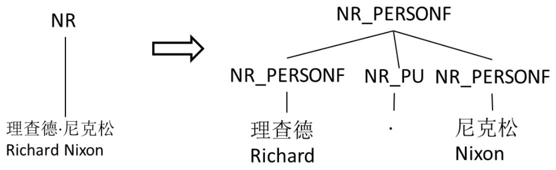


Figure 3: An example of annotation for the foreign name 理查德・尼克松“Richard Nixon”

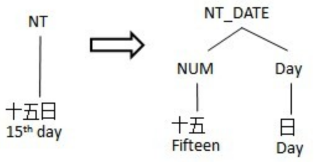


Figure 4: An example of nested annotation for the temporal expression 十五日“the 15th day in a month”

3.3 Our Annotation Method

The process of annotating named entity labels is as follows: Firstly, sentences also in OntoNotes (with file number from 1 to 325 and 1001 to 1078) will be selected, resulting in a small treebank with named entity annotations. A PCFG-LA parser is trained on the small treebank. Then the parser is used to label the rest of the sentences (with file number from 400 to 931 and 1100 to 1151). After that, the parsed sentences are manually corrected. Two persons marked the correct tags to each named entity independently. Manual correction is necessary, so can we avoid the danger of low-recall. Both persons should agree on a single tag when differences occurred.

The size of our new corpus is shown in Table 3.

CTB files	#Files	#Sens.	#NE	#NestedNE
1-325 1001-1078	403	8971	28344	1754
400-931 1100-1151	487	9778	28149	1144

Table 3: Statistics of the annotated corpus

3.4 Parsing Model

PCFG-LA in Petrov et al. (2006) used a hierarchical state-split approach to refine the original grammars. Starting with the basic non-terminal nodes, this method repeats the split-merge (SM) cycle to increase the complexity of grammars. Specifically, it splits every symbol into two, and then re-merge some new subcategories which cause little or less loss in likelihood incurred when removing it. In other words, the parser introduces latent annotations to refine the syntactic categories.

We employ Berkeley parser[2] in this study. We have re-implemented and enhance the Berkeley parser to handle Chinese character involved in nested named entity words efficiently and robustly. Especially, when the input is character not the word, we will change the strategy to deal the unknown character accordingly .

4 Experiments

In this section, we examine the effect of named entity cue in parsing Chinese. At the same time, the parser output an NER result. For the sake of comparison, here we also train a CRF model for NER as a baseline.

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

4.1 Experimental Setup

We present experimental results on Chinese Treebank (CTB) 5.0 with annotation of the named entity information. We adapted the standard data allocation and split the corpus as follows: files from CHTB_001.fid to CHTB_270.fid, and files from CHTB_400.fid to CHTB_1151.fid were used as training set. The development set includes files from CHTB_301.fid to CHTB_325.fid, and the test set includes files CHTB_271.fid to CHTB_300.fid. All traces and functional tags were stripped.

For comparison, we also trained a baseline BerkeleyParser without the cue, and a CRF model for named entity recognition. Our CRFs were implemented based on the CRF++ package [3], and the features used were mentioned in (Wan et al., 2011).

With regard to the parser from (Petrov et al., 2006), all the experiments were carried out after six cycles of split-merge.

4.2 Evaluation Metric

Three metrics were used for the evaluation of syntactic parsing: precision (P), recall (R) and F1-measure (F1) which is defined as 2PR/(P+R).

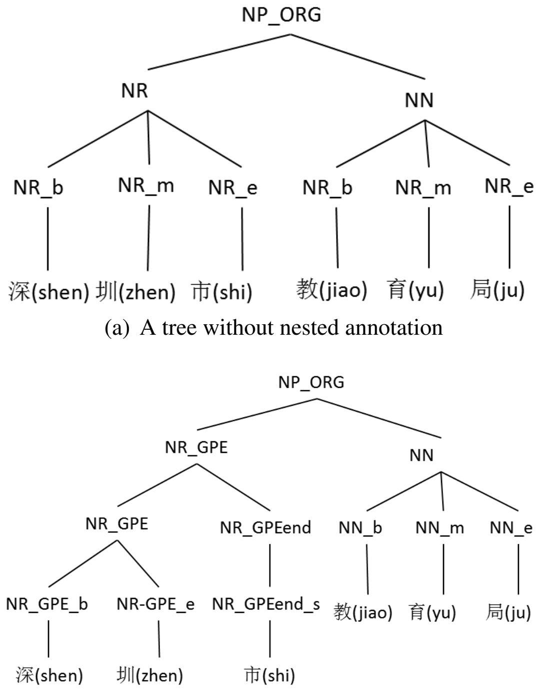
In the evaluation using the EVALB parseval, the additional named entity labels are also ignored. For instance, the label ‘NP_ORG’ and ‘NR_ORG’ will be replaced as ’NP’ and ’NR’ separately. The internal structure of nested named entity words are discarded by rules to make the results comparable to previous work.

We tested the significance of our results using Dan Bikel’s randomized parsing evaluation comparator[4], and validate the improvement in F1-measure is statistically significant.

4.3 Results on Parsing

In this section, we examine the effect of joint learning of syntactic structure and named entity cues for parsing.

Using the same data set setup and evaluation metric as the previous experiments, our parser achieves performance of 84.43 in F1-measure on the test data. Table 4 lists a few state-of-the-art word-level parser performance, showing that our system is competitive with all the others.



(a) A tree without nested annotation

(b) A tree with nested annotation

Figure 5: Not nested and Nested named entity annotation in the character-level tree for 深(shen) 圳(zhen) 市(shi) 教(jiao) 育(yu) 局(ju) “Shenzhen Education Bureau”

Experiment results show that named entity cue is useful for parsing. PCFG-LA method refines the syntactic categories by latent annotations, whereas, we introduce the fine-grained subcategorizations in the form of explicit annotations. The completely data-driven approach is prone to overfit, and the introduction of named entity cue by manual annotations is a more reliable way than unsupervised clustering.

System	**P**	**R**	**F1**
Petrov ’07	84.8	81.9	83.3
Qian’12	84.57	**83.68**	84.13
This paper	**85.53**	83.34	**84.43**

Table 4: Comparisons of our word-level parsing results with state-of-the-art systems

[3] http://crfpp.sourceforge.net/

[4] http://www.cis.upenn.edu/dbikel/software.html

4.4 Examining the Effectiveness of These Annotations for NER

The above experiments demonstrate that syntactic parsing benefits from our integrated approach. In this section, we exploit the effect on named entity recognition of joint learning.

For comparison to previous work, we convert word-level trees into character-level trees according to some rules. Then, the trained grammar has the ability to parse on characters and output syntactic structure and named entity labels. The simple rules used in this conversion are as follows:

- All part-of-speech tags in Word-level become constituent labels in character-level trees. Then a new node for each character if cerated, and we assign a new label for each new node. The new label consists of the POS tag of its word and its position in its word('b' for starting position, 'e' for end position, and 'm' for others). For example , the character 教"Jiao" in NN-教育局"Jiao Yu Ju", will be labeled as ‘NNb’. [5]

- All the characters underlying the NUM node will replace with “#NUM#”.

In Table 5, we show the NER result of our joint model. In the named entity evaluation, only the named entities with the correct boundaries and the correct categories are regarded as a correct recognition.

Model	GPE	PER	ORG	LOC
CRF	86.98	88.56	48.79	67.28
Parsing+NotNested	85.61	85.63	40.63	54.73
Parsing+NestedNR	**89.64**	**89.97**	**63.44**	**73.07**

Table 5: NER F1 results using different models

There is a great performance improvement on named entity recognition, especially on the recognition for ORG. On one hand, the internal structure of the named entity helps to determine the boundary of the entity. For instance, the organization phrase 中国华侨国际文化交流促进会"China International Cultural Exchange Association of the overseas Chinese" can be recognized . But the CRF model cannot capture the long-distance structure. On the other hand, the structural context in which it appears can help determine the type of the entity. As illustrated in Figure 6, the structure "NP_ORG CC NP_ORG" is a pattern, and the noun phrases on both sides of the 与"and" should be of the same type.

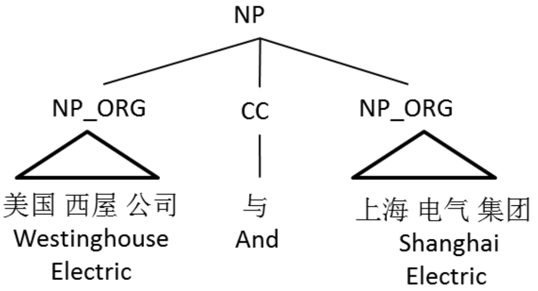


Figure 6: An example parsing result on the phrase 美国西屋公司与上海电气集团"Westinghouse Electric and Shanghai Electric"

5 Conclusion and Future Work

In this paper, we exploit the named entity cue in a unified framework for parsing. We annotate this cue in CTB5.0 through a joint representation of syntactic and named entity structures. Furthermore, we annotate nested named entity structure for all entity names, temporal expressions and number expressions. A PCFG-LA parser is then trained on the corpus. The evaluation shows that, introducing the named entity cue when training a parser help to recognize the complex named entity structures.

This preliminary investigation could be extended in several ways. First, it is natural to introduce other cues together, such as verbal subcategories and function word subcategories. Second, we would like to adopt discriminative parsing to integrate named entity cue into parsing.

Acknowledgment

This research was supported in part by the National Basic Research Program of China (973 Program) under grant 2013CB329304, the Key National Social Science Foundation of China under grant 12&ZD119, the National Natural Science Foundation of China under grant 91120001.

[5] This rule is the same as in Luo (2003) and Li (2011)

References

Daniel M Bikel and David Chiang. 2000. Two statistical parsing models applied to the chinese treebank. In *Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 12*, pages 1–6. Association for Computational Linguistics.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Association for Computational Linguistics.

Aitao Chen, Fuchun Peng, Roy Shan, and Gordon Sun. 2006. Chinese named entity recognition with conditional probabilistic models. In *5th SIGHAN Workshop on Chinese Language Processing, Australia*.

Michael Collins. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. thesis, University of Pennsylvania.

Jenny Rose Finkel and Christopher D Manning. 2009a. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 326–334. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D Manning. 2009b. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 141–150. Association for Computational Linguistics.

Mary Harper and Zhongqiang Huang. 2009. Chinese statistical parsing. *Gale Book*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Mark Johnson. 1998. Pcfg models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.

Dan Klein and Christopher D Manning. 2003. A parsing: fast exact viterbi parse selection. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 40–47. Association for Computational Linguistics.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 817–824. Association for Computational Linguistics.

Dongchen Li and Xihong Wu. 2012. Parsing tct with split conjunction categories. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 174–178. Association for Computational Linguistics and Chinese Information Processing Society of China.

Zhongguo Li. 2011. Parsing the internal structure of words: A new paradigm for chinese word segmentation. In *ACL*, pages 1405–1414.

Xiaoqiang Luo. 2003. A maximum entropy chinese character-based parser. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 192–199. Association for Computational Linguistics.

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*, pages 75–82. Association for Computational Linguistics.

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*, pages 404–411.

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 the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.

Xian Qian and Yang Liu. 2012. Joint chinese word segmentation, pos tagging and parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 501–511. Association for Computational Linguistics.

Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *LREC*.

Sameer Singh, Dustin Hillard, and Chris Leggetter. 2010. Minimally-supervised extraction of entities from text advertisements. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 73–81. Association for Computational Linguistics.

Xiaojun Wan, Liang Zong, Xiaojiang Huang, Tengfei Ma, Houping Jia, Yuqian Wu, and Jianguo Xiao. 2011. Named entity recognition in chinese news comments on the web. In *IJCNLP*, pages 856–864.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated chinese corpus. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–8. Association for Computational Linguistics.

Xiaofeng Yu, Wai Lam, Shing-Kit Chan, Yiu Kei Wu, and Bo Chen. 2008. Chinese ner using crfs and logic for the fourth sighan bakeoff. In *IJCNLP*, pages 102–105.

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*, pages 162–171. Association for Computational Linguistics.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese parsing exploiting characters. *51st Annual Meeting of the Association for Computational Linguistics*.