Evaluation Reportof the third Chinese Parsing Evaluation: CIPS-SIGHAN-ParsEval-2012

Qiang Zhou
Center for Speech and Language Technology
Research Institute of Information Technology
Tsinghua National Laboratory for Information
Science and Technology
Tsinghua University, Beijing 100084, China.
zq-lxd@mail.tsinghua.edu.cn

Abstract

This paper gives the overview of the third Chinese parsing evaluation: CIPS-SIGHAN-ParsEval-2012, including its parsing sub-tasks, evaluation metrics, training and test data. The detailed evaluation results and simple discussions will be given to show the difficulties in Chinese syntactic parsing.

1 Introduction

The first and second Chinese parsing evaluations CIPS-ParsEval-2009(Zhou and Li, 2009) and CIPS-SIGHAN-ParsEval-2010 (Zhou and Zhu, 2010) were held successfully in 2009 and 2010 respectively. The evaluation results in the Chinese clause and sentence levels show that the complex sentence parsing is still a big challenge for the Chinese language.

This time we will focus on the sentence parsing task proposed by the second CIPS-SIGHAN-ParsEval-2010 to dig out the detailed difficulties of Chinese complex sentence parsing in the respect of two typical sentence complexity schemes: event combination in the sentence level and concept composition in the clausal level. We will introduce a new lexicon-based Combinatory Categorical Grammar (CCG) (Steedman1996, 2000) annotation scheme in the evaluation, and make a parallel comparison of the parser performance with the traditional Phrase Structure Grammar (PSG) used in the Tsinghua Chinese Treebank (TCT) (Zhou, 2004).

This evaluation includes two sub-tasks, i.e. PSG parsing evaluation and CCG parsing evaluation. For each sub-task, there are two tracks. One is the Close track in which model parameter estimation is conducted solely on the train data. The other is the Open track in which any datasets other than the given training data can be used to estimate model parameters. We will set separated evaluation ranks for these two tracks.

In addition, we will evaluate following two kinds of methods separately in each track.

1) Single system: parsers that use a single parsing model to finish the parsing task.

2) System combination: participants are allowed to combine multiple models to improve the performance. Collaborative decoding methods will be regarded as a combination method.

2 Evaluation Tasks

Task 1: PSG Parsing Evaluation

Input: A Chinese sentence with correct word segmentation annotation. The word number is more than 2. The following is an example:

- 小型(small) 木材(wood) 加工场(factory) 在(is) 忙(busy) 着(-modality) 制作(build) 各 (several) 种 (-classifier) 木制品 (woodwork) 。(period) (A small wood factory is busy to build several woodworks.)

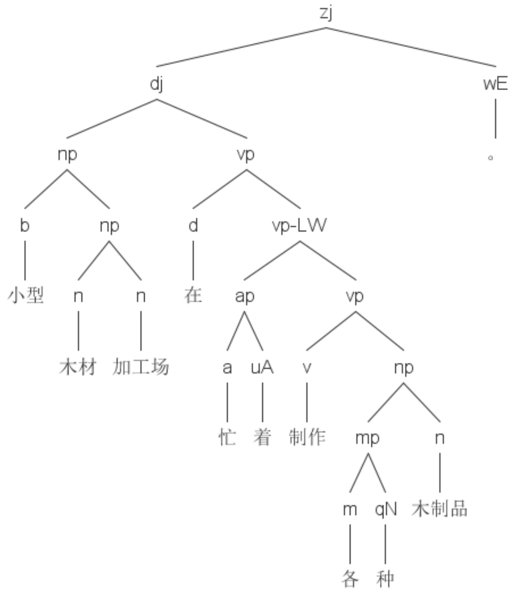
Parsing goal: Assign appropriate part-of-speech (POS) tags tothe words in the sentence and generate phrase structure tree for the sentence.

Output: The phrase structure tree with POS tags for the sentence.

- (zj (dj (np (b 小型) (np (n 木材) (n 加工场))) (vp (d 在) (vp-LW (ap (a 忙) (uA

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 159–167,
Tianjin, China, 20-21 DEC. 2012

着)) (vp (v 制作) (np (mp (m 各) (qN 种)) (n 木制品)))))) (wE 。))



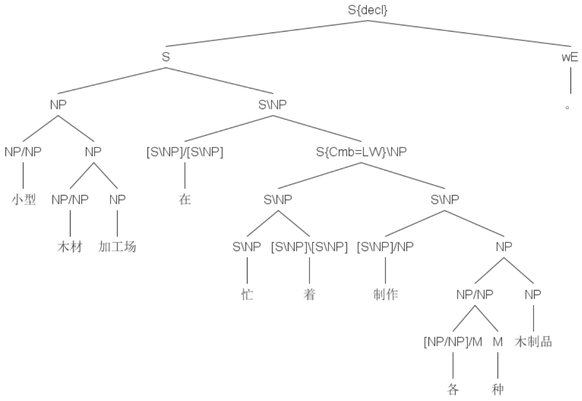
Task 2: CCG Parsing Evaluation

Input: Same with task 1.

Parsing goal: Assign appropriate CCG category tags tothe wordsin the sentence and generate CCG derivation tree for the sentence.

Output: The CCG derivation tree with CCG category tags for the sentence.

- (S{decl} (S (NP (NP/NP 小型) (NP (NP/NP 木材) (NP 加工场))) (S\NP ([S\NP]/[S\NP] 在) (S{Cmb=LW}\NP (S\NP (S\NP 忙) ([S\NP]\[S\NP] 着)) (S\NP ([S\NP]/NP 制作) (NP (NP/NP ([NP/NP]/M 各) (M 种)) (NP 木制品)))))) (wE 。))



3 Evaluation metrics

There are two parsing stages for the PSG and CCG parsers. One is the stage of syntactic category assignment, including POS tag and CCG category. The other is the stage of parse tree generation, including PSG parsing tree and CCG derivation tree. So we design two different sets of metrics for them.

3.1 Syntactic category evaluation metrics

Basic metrics are the syntactic category tagging precision (SC_P), recall (SC_R) and F1-score(SC_F1).

- SC_P= (#of correctly tagged words) /(# of automatically tagged words) * 100%
- SC_R= (#of correctly tagged words) /(# of gold-standard words) * 100%
- SC_F1= 2*SC_P*SC_R / (SC_P + SC_R)

The correctly tagged words must have the same syntactic categories with the gold-standard ones.

To obtain detailed evaluation results for different syntactic categories, we can classify all tagged words into different sets and compute different SC_P, SC_R and SC_F1 for them. The classification condition is as follows.

If (SC_Token_Ratio>=10%) then the syntactic tag will be one class with its SC tag, otherwise all other low-frequency SC-tagged words will be classified with a special class with Oth_SC tag. Where, SC_Token_Ratio= (word token # of one special SC in the test set) / (word token # in the test set) * 100%.

3.2 Parsing tree evaluation metrics

Basic metrics are the labeled constituent precision (LC_P), recall (LC_R) and F1-score (LC_F1).

- LC_P = (#of correctly labeled constituents) /(# of automatically parsed constituents) * 100%
- LC_R= (# of correctly labeled constituents) / (# of gold-standard constituents) * 100%
- LC_F1= 2*LC_P*LC_R / (LC_P+LC_R)

The correctly labeled constituents must have the same syntactic tags and left and right boundaries with the gold-standard ones.

To obtain detailed evaluation results for different syntactic constituents, we can classify them into 6 sets and compute different LC_P, LC_R and LC_F1 for them.

(1) Clausal and phrasal constituents
(2) Complex event constituents
(3) Concept compound constituents
(4) Single-node constituents
(5) Complementary parsing constituents
(6) All other constituents

The classification is based on the syntactic constituent tags annotated in the automatically parsed results. Please refer next section for more detailed information.

We compute the labeled F1-scores of the first four sets (Tot4_LC_F1) to obtain the final ranked scores for different proposed systems. For comparison analysis, we also list the F1-scores of all six sets for ranking reference.

To estimate the possible performance upper bound of the automatic parsers, we also design the following complementary metrics:

(1) Unlabeled constituent precision (ULC_P)= (# of constituents with correct boundaries) / (# of automatically parsed constituents) * 100%

(2) Unlabeled constituent recall (ULC_R)= (# of constituents with correct boundaries) / (# of gold standard constituents) * 100%

(3) Unlabeled constituent F1-score (ULC_F1)= 2*ULC_P*ULC_R / (ULC_P + ULC_R)

(4) Non-crossed constituent precision (NoCross_P)= (# of constituents non-crossed with the gold standard constituents) / (# of automatically parsed constituents) * 100%

4 Evaluation data

We used the annotated sentences in the TCT version 1.0 (Zhou, 2004) as the basic resources and designed the following automatic transformation procedures to obtain the final training and test data for the two parsing tasks.

Firstly, we make binary for all TCT annotation trees[1] and obtain a new binarizated TCT version. Two new grammatical relation tags RT and LT are added to describe the inserted dummy nodes with left and right punctuation combination structures. They can provide basic parsing tree structures for PSG and CCG parsing evaluations.

Secondly, we classify all TCT constituents into 6 sets, according to the syntactic constituent (SynC) and grammatical relation (GR) tags annotated in TCT[2].

1. Clausal and phrasal constituents, if all the following two conditions are matched
   a) TCT GR tag ∈{ZW, PO, DZ, ZZ, JY, FW, JB, AD}
   b) TCT Sync tag ∈{dj, np, sp, tp, mp, vp, ap, dp, pp, mbar, bp}
2. Complex event constituents, if one of the following conditions is matched.
   a) TCT SynC tag=fj and TCT GR tag ∈{BL, LG, DJ, YG, MD, TJ, JS, ZE, JZ, LS}
   b) TCT SynC tag=jq
3. Concept compound constituents, if all the following two conditions are matched
   a) TCT GR tag ∈{LH, LW, SX, CD, FZ, BC, SB}
   b) TCT Sync tag ∈{np, vp, ap, bp, dp, mp, sp, tp, pp}
4. Single-node constituents, if TCT SynC tag=dlc
5. Complementary parsing constituents, if TCT GR tag ∈{RT, LT, XX}
6. All other constituents

They will provide basic information for detailed parsing tree evaluation metrics computation.

Finally, we build the evaluation data sets for two parsing tasks through the following approaches:

1. For PSG parsing evaluation, we automatically transform the TCT annotation data through:
   a) For the syntactic constituents belong to the above class 2-3 and 5-6, we retain the original TCT two tags;
   b) For the syntactic constituent belong to the above class 1-4, we only retain the original TCT SynC tags.
2. For CCG parsing evaluation, we automatically transform the TCT annotation data into CCG format by using the TCT2CCG tool (Zhou, 2011).

5 Evaluation Results

5.1 Training and Test data

All the news and academic articles annotated in the TCT version 1.0 (Zhou, 2004) are selected as the basic training data for the evaluation. It consists of about 480,000 Chinese words. 1000 sentences extracted from the TCT-2010 version are used as the basic test data.

Table 1 shows the basic statistics of the training and test set. Figure 1 and Figure 2 list the distribution curve of the annotated sentences with different lengths (word sums) in the training and test set. They show very similar statistical

[1] TCT binarizationalgorithm and TCT2CCG tool were finished during the author visited Microsoft Research Asia (MSRA) in April, 2011. The visiting project was supported by the MSRA research foundation provided by Prof. Ming Zhou and Prof. Changning Huang.

[2] Please refer (Zhou, 2004) for more detailed descriptions of these syntactic constituent and grammatical relation tags.

characteristics. Their peaksare located in the region of 14 to 23. More than 75% annotated sentences have 15 or more Chinese words. The average sentence length is about 26. All these data show the complexity of the syntactic parsing task in the Chinese real world texts.

Table 1 Basic statistics of the training and test data: Average Sentence Length(ASL)=Word Sum/ Sent. Sum)

	Sent. Sum	Word Sum	Char. Sum	ASL
Training Set	17558	473587	762866	26.97
Test Set	1000	25226	39564	25.23

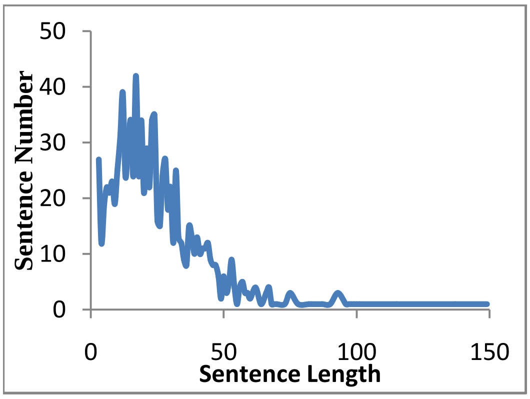


Figure 2 Sentence Length Distribution of the Test Set

Table 2 shows the statistics of different annotated constituents in the training and test set. We can find than about 68% constituents among them are clausal and phrasal constituents (class 1). They are the backbones of the syntactic parsing trees of Chinese sentences. About 20% constituents are complementary parsing constituents (class 5). It shows the importance of the punctuations in Chinese syntactic parsing. They can provide useful segmentation information to select suitable syntactic structures. About 12% constituents are complex event constituents (class 2) and concept compound constituents (class 3). They are the main points to determinate the parsing complexity of Chinese sentences. Few annotated examples in the training set will bring in more difficulties for feature extraction and parameter training in the statistics-based parsing models.

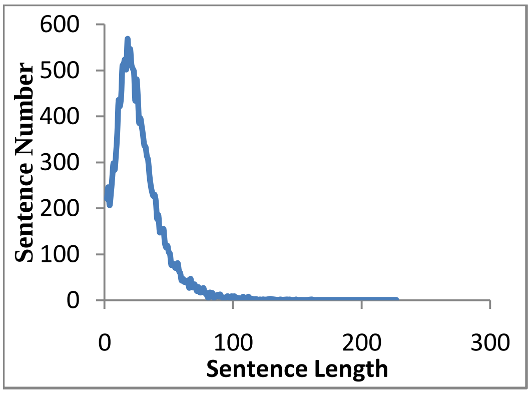


Figure 1 Sentence Length Distribution of the Training Set

Table 2 Different annotated constituents in the training and test set

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Total
Training set	310394	24239	30719	2735	89836	316	458239
Test set	16617	1578	1224	53	4746	50	24268

Table 3 Participant information for ParsEval-2012

ID	Participants	Registered Tasks	Proposed Tasks	Systems (Open/Close)
1	Institute of Automation, Chinese Academy of Science	PSG/CCG	/	/
2	Dalian University of Technology	PSG	/	/
3	Nanjing Normal University	PSG	/	/
4	Beijing Information Science and Technology University	PSG	PSG	1/0
5	Harbin Institute of Technology	PSG/CCG	PSG	3/0
6	Speech and Hearing Research Center, Peking University	PSG/CCG	PSG	1/1
7	University of Macau	PSG	PSG	0/1
8	Japan Patent Information Organization	PSG/CCG	PSG	0/1*

5.2 General results

8 participants proposed the registration forms, including 8 for PSG parsing and 4 for CCG parsing subtasks. Among them, 5 participants proposed the final evaluation results of 8 systems. All of them are for PSG parsing task. Table 3lists the basic information of these participants. Because the proposed result of the ID No. 8 gave very little standard binarizated parsing trees and lot of multiple-node constituents, after modifying current evaluation tool, we also include its result in the following evaluation performance tables.

Table4 and Table 5 show the ranked results of the proposed systems in the Open track and Close track respectively. We can find that the best parsing performances (Tot4_LC_F1) of the single model systems in the Open and Close track of the PSG parsing task is about 76-77%, which are similar with the best evaluation results in the task 2-2 of CIPS-SIGHAN-ParsEval-2010. In the respect of the unlabeled constituents, most single model systems can achieve about 87% F1 score, which are 10% better than that of the labeled constituents. After model combination, the F1 score of the best multiple model system can be improved to 90.3% (ID=05). We think it possibly reach the upper bound of boundary identification in the Chinese syntactic parsing task.

As we expected, the parsing performances of the clausal and phrasal constituents (class 1) and the complementary parsing constituents (class 5) are better than the overall results. The best labeled constituent F1 score of the single model system listed in Table 9 is 80.72%, which is about 4% better than the overall F1 score. Due to their simple internal structures, the complementary parsing constituents (class 5) obtain better parsing performances than that of the class 1 (+about 1-2%). The parsing performances of the complex event constituents (class 2) and the concept compound constituents (class 3) are much lower than the overall results with about 20-30% drops in the labeled constituent F1 score. Between them, the LC_F1 of constituents in class 2 is about 8-10% lower than that of class 3. A possible reason is that they may need more long-distance dependency features that are very difficult to be extracted through current statistical parsing model. The same trend can be also found in the performance data in the Open track listed in Table 7.

Unlike the labeled constituents, the parsing performances of the unlabeled constituents of different classes in the Open and Close Track didn’t show such larger differences (Table 6 and Table 8). Only the concept compound constituents (class 3) show lower F1 scores (-about 8-10% lower). The main reason is there are lots of crossed coordination constituents in the automatic parsing trees. It is still a big problem to identify the correct boundaries of the coordination constituents in the complex structures.

5.3 Detailed results

To evaluate the effect of different training corpus scale for parser performance, we divide all training data into *N* parts. In each training round, the *n* parts (n∈[1,10]) annotation corpora can be used to train *N* different parsing models with incremental training data. Based on them, *N* different test results can be obtained on the same test data set. Therefore, several variation trend diagrams of different kinds of evaluation metrics on different training corpus can be built. In the evaluation, we set *N*=10.

2 participants provided their incremental training test results, including 1 system in the Open track and 2 systems in the Close track.Figure 3, Figure 4 and Figure 5 show their general results. We list the following four main evaluation metrics in the figures for reference: syntactic category tagging F1 score (SC_F1), unlabeled constituent F1 score (ULC_F1), labeled constituent F1 score (LC_F1) and the labeled F1-scores of the first four constituent sets (Tot4_LC_F1).

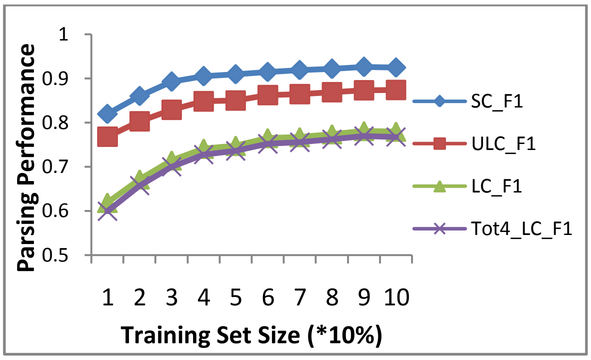


Figure 3 General performance improvement curve under different training data (ID=06, Open Track)

To find the performance improvement trend under different training data more clearly and detailed, we also collect the corresponding data of different class constituents. Figure 6, Figure 7 and Figure 8 show the results. In these figures, we select the labeled constituent F1 score (LC_F1) for reference.

From these figures, we can find that all the parsing performances are gradually improved after using more annotated data for training. It indicates the importance of large-scale annotated sentences for Chinese parser development. But the effects of the annotated sentences for different constituents and parsing stages are different and variable. We need to design new treebank building strategy to annotate more effective sentences with little manual workloads.

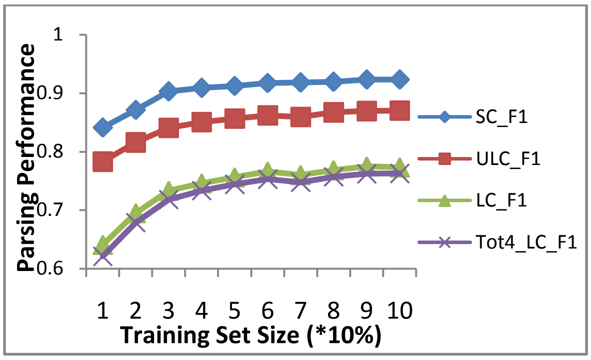


Figure 4General performance improvement curve under different training data (ID=06, Close Track)

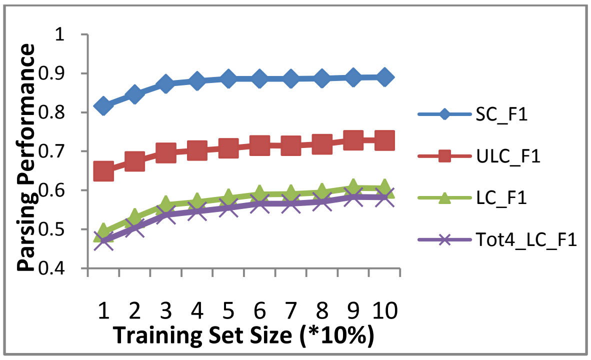


Figure 5General performance improvement curve under different training data (ID=07, Close Track)

For the syntactic category assignment stage (POS tagging in the PSG parsing subtask), after using all the training data, the SC_F1 still show some improvement trend. So we can expect to use more POS annotated sentences to obtain better POS tagging performance. 96% SC_F1 in the 4thSigHan bakeoff evaluation (Jin and Chen, 2008) under about 1M Chinese words training data proves the feasibility of this approach.

For the parse tree generation stage, we can find the different improvement effects of the training data for different kinds of constituents. For the clausal and phrasal constituents (class 1) and the complementary parsing constituents (class 5), more than 60% current training data may be enough to train a better parsing model. But for the complex event constituents (class 2) and the concept compound constituents (class 3), the fluctuated performance curves show the deficiency of current training data. How to select and annotated enough annotated sentences for them is still an open question need to be explored in the future.

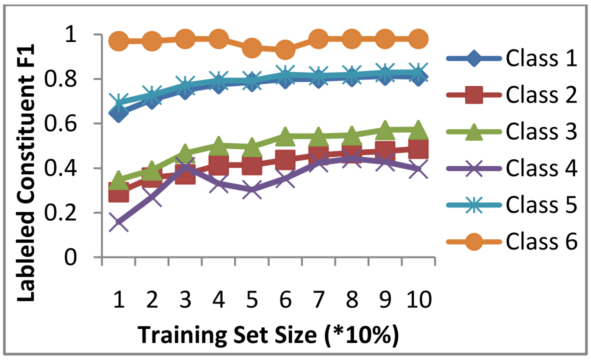


Figure 6 Performance improvement curve of different class of constituents under different training data (ID-06, Open Track)

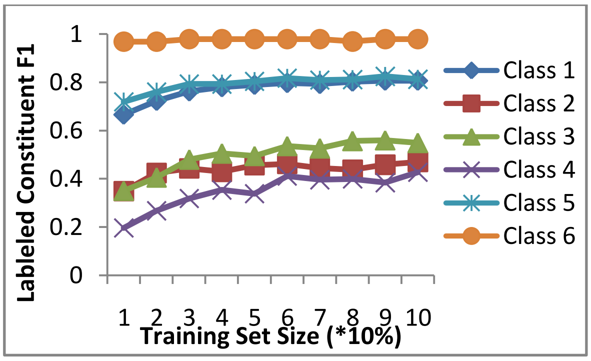


Figure 7 Performance improvement curve of different class of constituents under different training data (ID-06, Close Track)

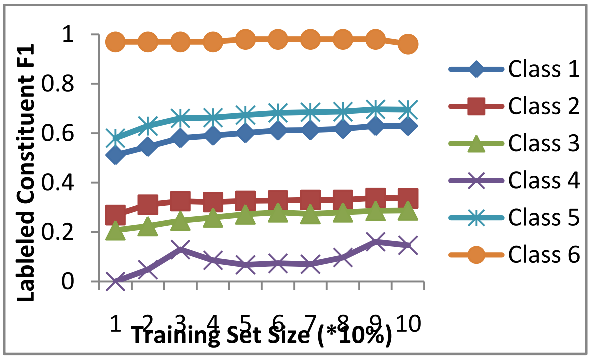


Figure 8 Performance improvement curve of different class of constituents under different training data (ID-07, Close Track)

5.4 Different parsing systems

4 participants proposed 5 technical reports to describe their parsing systems. In the section, we will briefly introduction some key techniques used in these systems.

(Zhang et al., 2012) proposed a bagging method to combine different parsers trained on different treebanks. They adopted Berkeley parser to train two different sub-models based on the TCT and CTB data, and then combined their outputs through CKY-parsing algorithm.

(Li and Wu, 2012) proposed a multilevel coarse-to-fine scheme for hierarchically split PCFGs. After automatically generating a sequence of nested partitions or equivalence classes of the PCFG non-terminals, the parsing model can start from a coarser level to prune the next finer level.

(Huang et. all, 2012) adopted a factored model to parse the Simplified Chinese. The factored model is one kind of combined structure between PCFG structure and dependency structure. It mainly uses an extremely effective A* parsing algorithm which enables to get a more optimal solution.

(Wang et al., 2012) presented a challenge to parse simplified Chinese and traditional Chinese with a same rule-based Chinese grammatical resource---Chinese Sentence Structure Grammar (CSSG).The experiments show that the CSSG that was developed for covering simplified Chinese constructions can also analyze most traditional Chinese constructions.

6 Conclusions

Parsing evaluation under standard benchmark can provide objective research platform for parsing model development and language resource construction. The expected theme of the 3rd Chinese parsing evaluation is to dig out the detailed difficulties of complex sentence parsing. So we design new tag set and propose two different parsing subtasks for performance comparison.

Although there are not any CCG evaluation results proposed, more than 5 PSG parsing results still give us enough evaluation data to verify our preliminary assumptions. Due to their complex internal structure, long-distance dependency and little annotation examples in real world annotated texts, the concept compound constituents and complex event constituents show extremely lower parsing performance than that of most clausal and phrasal constituents. How to collect enough annotated examples for them and explore new feature extraction method will be new research topic in the future.

Acknowledgments

The research was supported by National Basic Research Program of China (Grant No.: 2013CB329304) and National Science Foundation of China (Grant No.: 60873173). Thanks Mr. Qiu Han to develop the evaluation tools and manage all the evaluation results.

References

Clark, S., Copestake, A., Curran, J.R., Zhang, Y., Herbelot, A., Haggerty, J., Ahn, B.G., Wyk, C.V., Roesner, J., Kummerfeld, J., Dawborn, T.: 2009 Large-scale syntactic processing: Parsing the web. *Final Report of the 2009 JHU CLSP Workshop*

QiupingHuang, Liangye He, Derek F. Wong and Lidia S. Chao. 2012. A Simplified Chinese Parser with Factored Model. In *Proc. of CLP-2012*.

Guangjin Jin and Xiao Chen.2008.The Fourth International Chinese Language Processing Bakeoff: ChineseWord Segmentation, Named Entity Recognition and Chinese POS Tagging. In *Proc. of Sixth SIGHAN Workshop on Chinese Language Processing*, P69-81

Dongchen Li and Xihong Wu. 2012. Parsing TCT with a Coarse-to-fine Approach. In *Proc. of CLP-2012*.

Mark Steedman. 1996. *Surface Structure and Interpretation*. MIT Press, Cambridge, MA.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA.

Xiangli Wang, TerumasaEhara and Yuan Li. 2012. Parsing Simplified Chinese and Traditional Chinese with Sentence Structure Grammar. In *Proc. of CLP-2012*.

Meishan Zhang, WanxiangChe and Ting Liu. 2012. Multiple TreeBanks Integration for Chinese Phrase Structure. In *Proc. of CLP-2012*.

Qiang Zhou. 2004. Chinese Treebank Annotation Scheme. *Journal of Chinese Information*, 18(4), p1-8.

Qiang Zhou, Yuemei Li. 2009. Evaluation report of CIPS-ParsEval-2009.In Proc. of First Workshop on Chinese Syntactic Parsing Evaluation, Beijing China, Nov. 2009.pIII—XIII.

Qiang Zhou, Jingbo Zhu. 2010. Chinese Syntactic Parsing Evaluation. *Proc. of CIPS-SIGHAN Joint Conference on Chinese Language Processing* (CLP-2010), Beijing, August 2010, pp 286-295.

Qiang Zhou. 2011. Automatically transform the TCT data into a CCG bank: designation specification Ver 3.0. Technical Report CSLT-20110512, Center for speech and language technology, Research Institute of Information Technology, Tsinghua University.

Table 4 Ranked results in the Open Track of the PSG parsing task

ID	Sys_ID	Models	SC_F1	ULC_P	ULC_R	ULC_F1	NoCross_P	LC_P	LC_R	LC_F1	Tot4_LC_P	Tot4_LC_R	Tot4_LC_F1	Rank
5	CPBag	Multiple	93.97%	90.30%	90.24%	90.27%	90.30%	82.19%	82.14%	82.16%	81.34%	81.26%	81.30%	1
5	Cbag	Multiple	93.29%	90.35%	90.29%	90.32%	90.35%	82.08%	82.03%	82.05%	81.20%	81.12%	81.16%	2
5	Bbag	Multiple	93.06%	89.57%	89.51%	89.54%	89.57%	81.12%	81.07%	81.10%	80.23%	80.11%	80.17%	3
6		Single	92.50%	87.44%	87.43%	87.44%	87.44%	78.01%	78.00%	78.01%	76.81%	76.66%	76.74%	1
4		Single	92.73%	87.11%	87.13%	87.12%	87.11%	63.95%	63.96%	63.95%	70.10%	68.08%	69.08%	2
8*		Single	59.00%	38.57%	23.07%	28.87%	38.72%	29.21%	17.48%	21.87%	27.75%	18.76%	22.39%	3

Table 5 Ranked results in the Close Track of the PSG parsing task

ID	Models	SC_F1	ULC_P	ULC_R	ULC_F1	NoCross_P	LC_P	LC_R	LC_F1	Tot4_LC_P	Tot4_LC_R	Tot4_LC_F1	Rank
6	Single	92.29%	87.02%	87.04%	87.03%	87.02%	77.29%	77.32%	77.30%	76.35%	76.20%	76.27%	1
7	Single	89.01%	72.74%	72.86%	72.80%	72.74%	60.45%	60.55%	60.50%	58.26%	58.15%	58.20%	2

Table 6 Evaluation results of the different classes in the Open Track (unlabeled constituents)

	Class 1			Class 2			Class 3			Class 4			Class 5			Class 6		
ID	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
4	87.20%	90.21%	88.68%	82.27%	82.64%	82.45%	91.55%	5.31%	10.04%	81.54%	100.00%	89.83%	84.69%	53.27%	65.40%	92.68%	4408.00%	181.55%
5-b	89.63%	90.41%	90.01%	87.02%	87.52%	87.27%	84.56%	72.96%	78.33%	89.19%	62.26%	73.33%	91.22%	91.55%	91.39%	100.00%	96.00%	97.96%
5-c	90.53%	91.50%	91.02%	87.19%	87.14%	87.16%	84.51%	72.22%	77.89%	94.12%	60.38%	73.56%	91.90%	92.01%	91.96%	100.00%	96.00%	97.96%
5-cp	90.51%	91.54%	91.02%	87.04%	86.82%	86.93%	84.47%	71.57%	77.49%	91.43%	60.38%	72.73%	91.79%	91.93%	91.86%	100.00%	96.00%	97.96%
6	87.35%	87.30%	87.33%	85.51%	87.52%	86.50%	80.24%	76.31%	78.22%	75.00%	67.92%	71.29%	90.15%	90.83%	90.49%	100.00%	96.00%	97.96%

Table 7 Evaluation results of the different classes in the Open Track (labeled constituents)

	Class 1			Class 2			Class 3			Class 4			Class 5			Class 6		
ID	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
4	74.42%	76.98%	75.68%	25.68%	25.79%	25.74%	39.44%	2.29%	4.32%	46.15%	56.60%	50.85%	75.54%	47.51%	58.34%	0.42%	20.00%	0.82%
5-b	83.77%	84.50%	84.13%	51.04%	51.33%	51.18%	68.47%	59.07%	63.42%	67.57%	47.17%	55.56%	84.57%	84.87%	84.72%	100.00%	96.00%	97.96%
5-c	84.79%	85.70%	85.24%	51.30%	51.27%	51.28%	68.74%	58.74%	63.35%	76.47%	49.06%	59.77%	85.50%	85.61%	85.55%	100.00%	96.00%	97.96%

5-cp	84.93%	85.89%	85.41%	51.40%	51.27%	51.33%	68.76%	58.25%	63.07%	77.14%	50.94%	61.36%	85.48%	85.61%	85.55%	100.00%	96.00%	97.96%
6	80.97%	80.92%	80.94%	48.17%	49.30%	48.73%	58.76%	55.88%	57.29%	41.67%	37.74%	39.60%	82.66%	83.29%	82.98%	100.00%	96.00%	97.96%

Table 8 Evaluation results of the different classes in the Closed Track (Unlabeled constituents)

	Class 1			Class 2			Class 3			Class 4			Class 5			Class 6		
ID	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
6	87.26%	87.17%	87.21%	84.69%	83.78%	84.23%	77.92%	76.96%	77.44%	76.56%	92.45%	83.76%	89.23%	90.12%	89.67%	100.00%	96.00%	97.96%
7	71.42%	71.31%	71.36%	80.81%	76.87%	78.79%	52.64%	52.94%	52.79%	46.85%	98.11%	63.41%	80.22%	81.54%	80.88%	100.00%	100.00%	100.00%

Table 9 Evaluation results of the different classes in the Closed Track (labeled constituents)

	Class 1			Class 2			Class 3			Class 4			Class 5			Class 6		
ID	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
6	80.76%	80.68%	80.72%	47.28%	46.77%	47.02%	55.25%	54.58%	54.91%	39.06%	47.17%	42.74%	80.91%	81.71%	81.31%	100.00%	96.00%	97.96%
7	62.93%	62.83%	62.88%	34.44%	32.76%	33.58%	28.68%	28.84%	28.76%	10.81%	22.64%	14.63%	68.91%	70.04%	69.47%	96.00%	96.00%	96.00%