Evaluation Report of the fourth Chinese Parsing Evaluation: CIPS-SIGHAN-ParsEval-2014

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 fourth Chinese parsing evaluation: CIPS-SIGHAN-ParsEval-2014, including its parsing, 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

For Chinese parsing evaluations, we have successfully held three times in 2009, 2010 and 2012. They are the CIPS-ParsEval-2009 (Zhou and Li, 2009), CIPS-SIGHAN-ParsEval-2010 (Zhou and Zhu, 2010) and CIPS-SIGHan-ParsEval-2012 (Zhou, 2012) respectively. Each evaluation has its different theme and goal.

The first ParsEval-2009 focused on Chinese chunk parsing. Three kinds of chunking tasks were designed for the Chinese chunks with different descriptive complexities. The evaluation results showed that as the increasing of the word number and descriptive complexity of the chunks from base chunks (BC) to functional chunks (FC) and event descriptive chunks (EDC), the final F1-value will also decrease about 6 points from 92% to 86% and 80%.

The second ParsEval-2010 and third ParsEval-2012 focused on Chinese syntactic parsing. They had different points of emphasis for parse tree evaluation.

In ParsEval-2010, we compared the parsing performance differences in two kinds of Chinese sentences. One is the EDC clauses with about 10 words averagely. The other is the complete sentences with about 23 words averagely. Evaluation results showed that there were about 8% drops for the complete sentence in the labelled F1-score measure.

In ParsEval-2012, we compared the parsing performance differences in two kinds of syntactic constituent in Chinese complete sentences. One is the syntactic constituents with complex internal compound relationships, including event combination and concept composition relations. The other is the syntactic constituents with ordinary internal relations, such as subject-predicate, predicate-object, modifier-head, etc. Evaluation results showed that there were 20% drops for the syntactic constituents with complex internal relations in the labelled F1-score measure.

The above 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 deeper parsing evaluation in the Predicate-Argument Structure (PAS) level to test whether the parser can deal with different syntactic alternatives with same event contents. We will introduce a new lexicon-based Combinatory Categorical Grammar (CCG) (Steedman 1996, 2000) annotation scheme in the evaluation, and propose a new implicit predicate argument (IPA) relation annotation method to build a large scale CCG bank with detailed PAS annotations. The special lexical dependency pairs automatically extracted from the CCG bank will be used as the final gold-standard data for evaluating parsers' IPA recognition capacity.

Same with previous ParsEval-2010 and ParsEval-2014, we also set two tracks in the ParsEval-2014. 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

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 146–152,
Wuhan, China, 20-21 October 2014

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 Task and Metrics

2.1 Parsing Evaluation Task

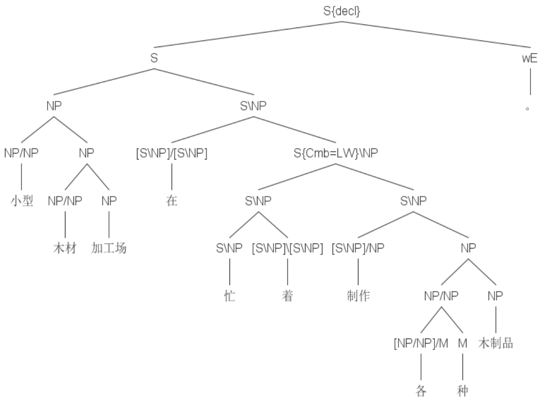
Input: A Chinese sentence with correct word segmentations. The following is an example:

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 CCG category tags to the words in the sentence and generate CCG derivation tree for the sentence.

Output: The CCG derivation tree with CCG category tags and feature annotations.

- (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 。))



(1)

2.2 Parsing Evaluation Metrics

There are two parsing stages for the CCG parsers. One is the syntactic category (CCG category) assignment stage. The other is the parse tree (CCG derivation tree) generation stage. So we design two different sets of metrics for them.

For the syntactic category (SC) parsing stage, basic metrics are SC 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 will 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%.

For the CCG derivation tree generation stage, the lexical dependency pairs (LDPs) automatically extracted from the CCG derivation trees will be used as the basic evaluation units. Basic metrics for them are LDP precision (LDP_P), recall (LDP_R) and F1-score (LDP_F1).

- LDP_P = (# of correctly labeled LDPs) / (# of automatically parsed LDPs) * 100%
- LDP_R= (# of correctly labeled LDPs) / (# of gold-standard LDPs) * 100%
- LDP_F1= 2*LDP_P*LDP_R / (LDP_P+LDP_R)

The correctly labeled LDPs must have the same annotation information with the gold-standard ones.

To obtain detailed evaluation results for different LDPs, we can classify them into 5 sets and compute different LDP_P, LDP_R and LDP_F1 for them respectively.

(1) LDPs with complex event relations in the sentence levels;

(2) LDPs with concept compound relations in the chunk levels;

(3) LDPs with predicate-argument relations in the clause levels, including head-complement and adjunct-head relations.

(4) LDPs with other non-PA relations in the chunk and clause levels, including modifier-head and operator-complement relations.

(5) All other LDPs.

We compute the weighted average of the F1-scores of the first four sets (Tot4_F1) to obtain the final ranked scores for different proposed parser systems. The computation formula is as follows: Tot5_F1=$\sum LDP\_F1_i * LDP\_Ratio_i$，$i \in [1,4]$.

LDP_Ratio_i is the distributional ratio for the i^{th} LDP set in the test set. It computation formula is: LDP_Ratio_i= (# of LDPs in i^{th} set) / (# of all LDPs) * 100%

For comparison analysis, we also compute the weighted average of F1-scores of all five sets for ranking reference.

3 Evaluation data

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

Firstly, we automatically transformed all the TCT parse trees into CCG derivation trees by using the TCT2CCG tool (Zhou, 2011), and built a CCG bank version 1.0 for the TCT data. In the bank, most of clauses can be obtained correct CCG derivation trees due to the direct application of the syntax-semantics linking (SSL) principles among the basic syntactic constructions in Chinese sentences. The above CCG derivation tree (1) in section 2.1 is a good example. But there are still many syntactic constructions consist of implicit predicate-argument (IPA) relations, such as the topicalization and relative clause constructions. They can’t be automatically transformed into correct CCG derivation trees through the explicit SSL mapping rules. To deal with the problem, we proposed to manually annotate the IPA relations in these special constructions and restructure the corresponding CCG derivation sub-trees according to these annotated PA tags.

The key for IPA annotation is to find the suitable construction examples that carry the IPA relations in Chinese sentences. So we classify all the event constructions (ECs) in the Chinese sentences into the following three sets:

1) Basic event constructions (BEC)

They are the typical subject-predicate-object constructions in Chinese clause level. The direct SSL can be found in the constructions. So the current TCT2CCG tool is OK for them. A simple example is as follows:

- 我(I) 读过(have read) 这本书(the book). (I have read the book.)

2) Derived event constructions (DEC)

They are the derived constructions in Chinese clause level due to some special pragmatics purposes or contexts. Most of them are the topicalization or argument-ellipsis constructions. The following is a topicalization example:

- 这本书(the book) 我(I) 读过(have read). (The book, I have read.)

The topicalized deep object “这本书(the book)” should be given special IPA tags to show the detailed SSL relations.

3) Transformed event constructions (TEC)

Most of them are the relative sub-clauses to describe the special event backgrounds for an ongoing main event predicate. The structural particle 的(*de*) is used as the relative marker for them. The following is a relative sub-clause example (underlined) in a complete clause:

- 我(I) 读过(have read) 的(*de*) 这本书(the book) 很有趣(very interesting). (The book that I have read is very interesting.)

It is a big challenge to identify whether the relative noun phrases are the real extracted arguments in TECs or not.

Based on the above event construction classification, we proposed an EC-based IPA annotation scheme. For each DEC or TEC example extracted from Chinese real sentences, two or three independent annotators were asked to select the suitable corresponding BEC menu for them on an IPA annotation platform. Some detailed information about the IPA annotation procedure can be found in (Qiu, 2014).

After manual IPA annotation, we can obtain the following ECs with IPA tags for the above two DEC and TEC examples:

- [T-np-Arg2 这本书(the book)] [S-np-Arg1 我(I)] [P-vp-Pred 读过(have read)][1]
- [S-np-Arg1 我(I)] [P-vp-Pred 读过(have read)] 的(*de*) [H-np-Arg2 这本书(the book)]

So, they show the same event contents with the following corresponding BEC annotation:

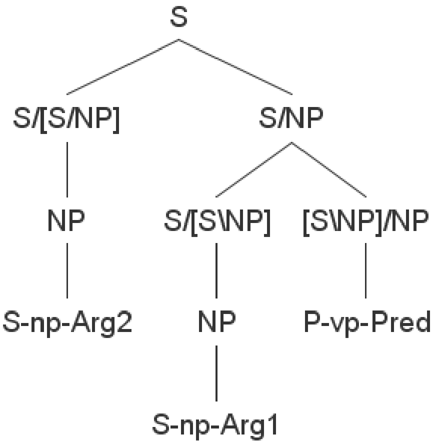
[1] Each event chunk will be given the following tag combinations: <Functional tag>-<Constituent tag>-<PA tag>. Some tags used in these examples are listed as follows: T-topic, S-subject, P-predicate, O-object, H-head; np-noun phrase, vp-verb phrase; ArgX-different argument position, Pred-predicate position

- [S-np-Arg1 我(I)] [P-vp-Pred 读过(have read)] [O-np-Arg2 这本书(the book)]

These detailed IPA tags provided us with enough indicators for further CCG derivation tree rebuilding. Some main CCG rebuilding principles are as follows:

1) The same CCG tags should be assigned to the event target predicates (ETP) in the corresponding BEC, DEC and TEC examples. So in the above three ECs, the ETP "读(read)" should be assigned the same CCG tag: (S\NP)/NP.
2) The deep arguments with same IPA tags should be linked to the same argument positions in the corresponding ETP's CCG tags. For example, the argument chunk with IPA tag "Arg1" should be linked to the first NP argument position in the corresponding ETP-读(read): (S\NP_1)/NP$_2$.

Based on the above principles, we proposed a CCG derivation tree rebuilding algorithm. Please refer (Qiu, 2014) for more details about the algorithm. Here, we will give some figures to show the key idea of rebuilding procedure for the DEC and TEC examples.



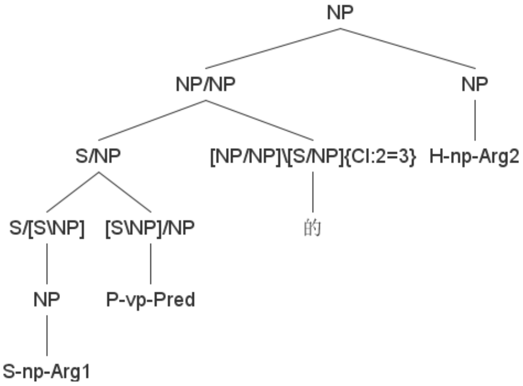
(2)

Figure (2) shows the rebuilt CCG derivation tree for a topicalized DEC. Two CCG type raising (TR) rules are used for locating two deep arguments:

- For deep subject: NP → S/(S\NP)
- For deep object: NP → S/(S/NP)

The CCG forward composition rule: S/(S\NP) (S\NP)/NP $\rightarrow_B$ S/NP, is used for the SSL of the deep subject. The special CCG forward application rule: S/(S\NP) S\NP → S, is used for the SSL of the topicalized deep object.

Figure (3) shows the rebuilt CCG derivation tree for a relative sub-clause TEC. The SSL of the deep subject is same with the above figure (2). The CCG co-indexing (CI) scheme is used for the SSL of the extracted deep object. It is assigned as a special feature in the CCG tag of the structure particle 的 (*de*): (NP$_1$/NP$_2$)\(S/NP$_3$) [CI:2=3], which means that the modified head (NP$_2$) of the relative clause is co-index with reduced deep object (NP$_3$) in the relative clause.



(3)

The rebuilt CCG derivation trees can provide consistent representations for different shallow syntactic alternatives with the same deep PA relations. Therefore, the same lexical dependency pairs for describing the PA relations in the above three different BEC, DEC and TEC examples can be automatically extracted (Hockenmaier et al., 2007) from the corresponding rebuilt CCG derivation trees:

- 读(read), (S\NP)/NP, 1, 我(I)
- 读(read), (S\NP)/NP, 2, 书(book)

They describe the same event contents consist in the above three EC examples. So we used these LDPs as the benchmark data for CCG parse tree evaluation.

4 Evaluation Results

4.1 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. After the TCT2CCG transformation, EC-based IPA annotation and CCG derivation tree rebuilding, all the training and test data have been annotated with suitable CCG format tags and derivation trees.

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	24108	34079	24.11

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 characteristics. Their peaks are 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 25. All these data show the complexity of the syntactic parsing task in the Chinese real world texts.

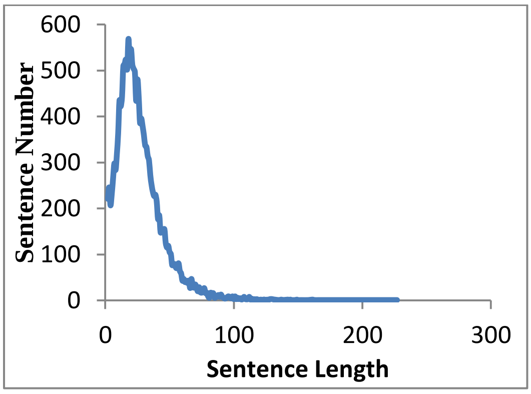


Figure 1 Sentence Length Distribution of the Training Set

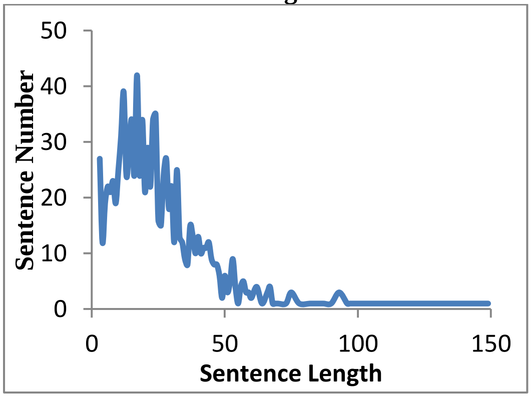


Figure 2 Sentence Length Distribution of the Test Set

4.2 General results

9 participants proposed the registration forms. Among them, only 1 participant proposed the final evaluation result. Table 2 lists the basic information of these participants.

错误!未找到引用源。 shows the ranked results of the proposed systems in the only Open track. Due to the difficulty of Chinese CCG parsing, the proposed system didn't show good parsing performance: SC_F1=71.81%, Tot5_LDP_F1=41.95%. Compared with the state-of-the-art English CCG parsers (Clark et al., 2004), the syntactic category tagging (supertagging) performance has about 20% drops in the Chinese CCG parser. It may indicate that the unknown word supertagging may be a big challenge for the Chinese language.

Table 4 lists the parsing performances of the LDPs with different internal dependency relations. As we have expected, the parsing performances of the LDPs with other non-PA relations (class 4) are the highest ones among them. The LDP-F1 score of them is about 5% better than the overall Tot4-LDP-F1 score. The second ones are the LDPs with PA relations. They show about 6% drops compared with the LDP with non-PA relations. It indicates that some outside lexical semantic resources may need for efficient PAS analysis. The parsing performances of the LDPs with complex event relations (class 1) and concept compound relations (class 2) are much lower than the overall LDP-F1 score with about 10-30% drops. Between them, the F1 score of the LDPs in class 1 is about 19% lower than that of class 2. 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. These performance changing trends are very similar with that were found in ParsEval-2012.

Table 2 Participant information for ParsEval-2014

ID	Participants	Systems (Open/Close)
1	NLP Labortory, Zhengzhou University	/
2	Brandeis University, USA	/
3	Beijing University of Posts and Telecommunications	/
4	Institute of Automation, CAS	1/0
5	Harbin Institute of Technology	/
6	Singapore Univ. of Technology and Design	/
7	Institut national des langues et civilisation Orientales(INALCO)	/
8	Zhejian Institute of Marine	/
9	Yahoo Corp.	/

Table 3 Ranked results in the Open Track of the CCG parsing task

ID	Models	SC_F1	LDP_P	LDP_R	LDP_F1	Tot4_LDP_P	Tot4_LDP_R	Tot4_LDP_F1	Rank
4	Single	71.81%	42.32%	42.27%	42.29%	41.83%	42.07%	41.95%	1

Table 4 Evaluation results of the different classes of LDPs in the Open Track

	Class 1			Class 2			Class 3			Class 4			Class 5		
ID	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
4	12.99%	11.92%	12.43%	26.80%	36.87%	31.04%	40.69%	40.47%	40.58%	47.60%	46.71%	47.15%	45.81%	43.62%	44.69%

5 Conclusions

Combinatory categorical grammar can provide strong platform for describing the deep PAS of different shallow syntactic alternatives with same event contents. So we introduced CCG into the 4th Chinese parsing evaluation (ParsEval-2014) and proposed an EC-based IPA annotation method to build a new CCG-based evaluation benchmark data. Although the number of the proposed systems was not enough to show the real application potential of CCG parsing for the Chinese language, we still think CCG parsing is a good direction need to be explored 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.: 61373075). Thanks Mr. Zhou Xiaocong to develop the LDP evaluation tools for the evaluation task.

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*

Clark, Stephen and James R. Curran. Parsing the WSJ using CCG and log-linear models. 2004. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 103–110, Barcelona, Spain.

Hockenmaier, J., Steedman, M.: 2007. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics* 33(3), 355--396

Han Qiu. 2014. Research on Chinese Predicate-Argument Structure Analysis and Annotation. Master thesis. Dept. of computer science and technology, Tsinghua University.

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

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

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.

Qiang Zhou. 2012. Evaluation Report of the third Chinese Parsing Evaluation: CIPS-SIGHAN-ParsEval-2012. Proc. of *CIPS-SIGHAN Joint Conference on Chinese Language Processing* (CLP-2012), Tianjin.