Chinese Syntactic Parsing Evaluation

Qiang Zhou
Center for Speech and Language Tech.
Research Institute of Information Tech.
Tsinghua University
zq-lxd@tsinghua.edu.cn

Jingbo Zhu
Natural Language Processing Lab.
Northeastern University
zhujingbo@mail.neu.edu.cn

Abstract

The paper introduced the task designing ideas, data preparation methods, evaluation metrics and results of the second Chinese syntactic parsing evaluation (CIPS-Bakeoff-ParsEval-2010) jointed with SIGHAN Bakeoff tasks.

1 Introduction

Syntactic parsing is an important technique in the research area of natural language processing. The evaluation-driven methodology is a good way to spur the its development. Two main parts of the method are a benchmark database and several well-designed evaluation metrics. Its feasibility has been proven in the English language.

After the release of the Penn Treebank (PTB) (Marcus et al., 1993) and the PARSEVAL metrics (Black et al., 1991), some new corpus-based syntactic parsing techniques were explored in the English language. Based on them, many state-of-art English parser were built, including the well-known Collins parser (Collins, 2003), Charniak parser (Charniak and Johnson, 2005) and Berkeley parser (Petrov and Klein, 2007). By automatically transforming the constituent structure trees annotated in PTB to other linguistic formalisms, such as dependency grammar, and combinatory categorical grammar (Hockenmaier and Steedman, 2007), many syntactic parser other than the CFG formalism were also developed. These include Malt Parser (Nivre et al., 2007), MSTParser (McDonald et al., 2005), Stanford Parser (Klein and Manning, 2003) and C&C Parser (Clark and Curran, 2007).

Based on the Penn Chinese Treebank (CTB) (Xue et al., 2002) developed on the similar annotation scheme of PTB, these parsing techniques were also transferred to the Chinese language. (Levy and Manning, 2003) explored the feasibility of applying lexicalized PCFG in Chinese. (Li et al., 2010) proposed a joint syntactic and semantic model for parsing Chinese. But till now, there is not a good Chinese parser whose performance can approach the state-of-art English parser. It is still an open challenge for parsing Chinese sentences due to some special characteristics of the Chinese language. We need to find a suitable benchmark database and evaluation metrics for the Chinese language.

Last year, we organized the first Chinese syntactic parsing evaluation --- CIPS-ParsEval-2009 (Zhou and Zhu, 2009). Five Chinese parsing tasks were designed as follows:

- Task 1: Part-of-speech (POS) tagging;
- Task 2: Base chunk (BC) parsing
- Task 3: Functional chunk (FC) parsing
- Task 4: Event description clause (EDC) recognition
- Task 5: Constituent parsing in EDCs

They cover different levels of Chinese syntactic parsing, including POS tagging (Task 1), shallow parsing (Task 2 & 3), complex sentence splitting (Task 4) and constituent tree parsing (Task 5). The news and academic articles annotated in the Tsinghua Chinese Treebank (TCT ver1.0) were used to build different gold-standard data for them. Some detailed information about CIPS-ParsEval-2009 can be found in (Zhou and Li, 2009).

This evaluation found the following difficult points for Chinese syntactic parsing.

1) There are two difficulties in Chinese POS tagging. One is the nominal verbs. The POS accuracy of them is about 17% lower than the overall accuracy. The other is the unknown

words. The POS accuracy of them is about 40-10% lower than the overall accuracy.

2) The chunks with complex internal structures show poor performance in two chunking tasks. How to recognize them correctly needs more lexical semantic knowledge.

3) The joint recognition of constituent tag and head position show poor performance in the constituent parsing task of EDCs.

Therefore, the second Chinese syntactic parsing evaluation (CIPS-Bakeoff-ParsEval-2010) jointed with SIGHAN Bakeoff tasks was proposed to deal with these problems. Some new designing ideas are as follows:

1) We use the segments sentences as the input of the syntactic parser to test the effects of POS tagging for Chinese parsing.

2) We design a new metric to evaluate performance of event construction recognition in a constituent parser of EDCs.

3) We try to evaluate the performance of event relation recognition in Chinese complex sentence.

In the following sections, we will introduce the task designing ideas, data preparation methods, evaluation metrics and results of the evaluation.

2 Task description

For the syntactic parsing task (Task 2) of the CIPS-Bakeoff-2010, we designed two sub-tasks:

Task 2-1: Parsing the syntactic trees in Chinese event description clauses

Task 2-2: Parsing the syntactic trees in Chinese sentences.

Each subtask is separated as close and open track. In the close track, only the provided training data can be used to build the parsing model. In the open track, other outside language resources can be freely used.

We will give two examples to show the detailed goals of these two sub-tasks:

1) Task 2-1

Input: a Chinese event description clause with correct word segmentation annotations

- 沿途，我们 不时 见到 因 更新 而 伐倒 的 树木， 因 修 路 需 伐倒 的 树木

Ouput: a syntactic parsing tree of the EDC with appropriate constitutent tag, head position and POS tag annotations.

- [dj-2 沿途/s， /wP [dj-1 我们/rNP [vp-1 不时/d [vp-0 见到/v [np-0-2 [np-2 [vp-1 [pp-1 因/p 更新/v] [vp-1 而/cC 伐倒/v]] 的/uJDE 树木/n]，/wP [np-2 [vp-1 [pp-1 因/p [vp-0 修/v 路/n]] [vp-1 需/vM 伐倒/v]] 的/uJDE 树木/n]]]]]][1]

2) Task 2-2

Input: a Chinese sentence with correct word segmentation annotations

- 沿途 ， 我们 不时 见到 因 更新 而 伐倒 的 树木 ， 因 修 路 需 伐倒 的 树木 ， 都 是 有用 之 材 ； 运送 树木 的 货车 、 拖拉机 ， 南来北往 。

Output: a syntactic parsing tree of the sentence with appropriate constitute tag and POS tag annotations.

- [zj [fj [fj [dj 沿途/s ，/wP [dj 我们/rNP [vp 不时/d [vp 见到/v [np [np [vp [pp 因/p 更新/v] [vp 而/cC 伐倒/v]] 的/uJDE 树木/n]，/wP [np [vp [pp 因/p [vp 修/v 路/n]] [vp 需/vM 伐倒/v]] 的/uJDE 树木/n]]]]]]，/wP [vp 都/d [vp 是/v [np 有用/a 之/uJDE 材/n]]]] ；/wP [dj [np [vp 运送/v 树木/n] 的/uJDE [np 货车/n 、/wD 拖拉机/n]]，/wP 南来北往/v]]。/wE]

We define a Chinese sentence as the Chinese word serials ending with period, question mark or exclamation mark in the Chinese text. Usually, a Chinese sentence can describe a complex situation with several inter-related events. It consists of several clauses separated by commas or semicolons to describe one or more detailed event content. We call these clauses as event description clauses.

We use the following example to explain the relationship between a Chinese sentence and

[1] Each bracketed constituent is annotated with constituent tag and head positions separated by '-'.
Constituent tags used in the sentence are: dj-simple sentence, vp-verb phrase, np-noun phrase, pp-preposition phrase.
POS tags used are: s-space noun, wP-comma, rNP-personal pronoun, d-adverb, v-verb, p-preposition, cC-conjunction, uJDE-particle, n-noun, vM-modality verb;

event description clauses.

- [沿途，我们见到因为更新而伐倒的树木，因为建筑需伐倒的树木]，[都是有用之材]；[运送树木的货车、拖拉机，南来北往]。 (1)
- [Along the way, we see the trees have been cut down for regeneration, and the trees needed to be cut for building]. [All of them are useful building material]. [We also see several freight trucks and tractors for carry away trees going south and north].

The sentence gives us several sequential situations through the vision changing along the author's journey way: Firstly, we see the trees that have been cut down. They are useful building material. Then, we see several trucks and tractors to carry away these trees. They are going south and north busily. All the above situations are described through three EDCs annotated with bracket pairs in the sentence.

Interestingly, in the corresponding English translation, the same situation is described through three English sentences with complete subject and predicate structures. They show difference event description characteristics of these two languages.

The Chinese author tends to describe a complex situation through a sentence. Many complex event relations are implicit in the structural sequences or semantic connections among the EDCs of the sentence. So many subjects or objects of an EDC can be easily omitted based on the adjacent contexts.

The English author tends to describe a complex situation through several sentences. Each sentence can give a complete description of an event through the subject and predicate structure. The event relations are directly set through the paragraph structures and conjunctions.

The distinction between Chinese sentence and EDC can make us focus on different evaluation emphasis in the CIPS-Bakeoff-2010 section.

For an EDC, we can focus on the parsing performance of event content recognition. So we design a special metric to evaluate the recall of the event recognition based on the syntactic parsing results.

For a sentence, we can focus on the parsing performance of event relation recognition. So we separate the simple and complex sentence constitutes and give different evaluation metrics for them.

Some detailed designations of the evaluation metrics can be found in section 4.

3 Data preparation

The evaluation data were extracted from Tsinghua Chinese Treebank (TCT) and PKU Chinese Treebank (PKU-CTB).

TCT (Zhou, 2004) adopted a new annotation scheme for Chinese Treebank. Under this scheme, every Chinese sentence will be annotated with a complete parse tree, where each non-terminal constituent is assigned with two tags. One is the syntactic constituent tag, such as noun phrase(np), verb phrase(vp), simple sentence(dj), complex sentence(fj), etc., which describes basic syntactic characteristics of a constituent in the parse tree. The other is the grammatical relation tag, which describes the internal structural relation of its sub-components, including the grammatical relations among different phrases and the event relations among different clauses. These two tag sets consist of 16 and 27 tags respectively.

Now we have two Chinese treebanks annotated under above scheme: (1) TCT version 1.0, which is a 1M words Chinese treebank covering a balanced collection of journalistic, literary, academic, and other documents; (2) TCT-2010, which consists of 100 journalistic annotated articles. The following is an annotated sentence under TCT scheme:

- [zj-XX [fj-LS [dj-ZW 我们/rN [vp-PO 问/v [dj-ZW [np-DZ 他/rN 自己/rN] [vp-PO 买/v 多少/m]]]]，/， [dj-ZW 他/rN [vp-LW [vp-PO 凑近/v [sp-DZ 记者/n 面前/s]] [vp-PO 伸出/v [np-DZ [mp-DZ 4/m 个/qN] 指头/n]]]]] 。/。] [2] (2)

PKU-CTB (Zhan et al., 2006) adopted a traditional syntactic annotation scheme. They annotated Chinese sentences with syntactic constitu-

[2] Some grammatical relation tags used in the sentence are as follows: LS—complex timing event relation, ZW—subject-predicate relation, DZ—modifier-head relation, PO—predicate-object relation.

ent and head position tags in a complete parse tree. The tag set consists of 22 constituent tags. Because every content word is directly annotated with suitable constituent tag, there are many unary phrases in PKU-CTB annotated sentences. Its current annotation scale is 881,771 Chinese words, 55264 sentences. The following is an annotated sentence under PKU-CTB scheme:

- (zj (!fj (!fj (!dj (np (vp (!v (建筑)) !np (!n (公司))) !vp (!vp (!v (进)) np (!n (区)))) wco (，) dj (np (ap (!b (有关)) !np (!n (部门))) !vp (dp (!d (先)) !vp (!vp (!vp (!v (送)) v (上)) np (qp (mp (!rm (这)) !q (些)) !np (np (!n (法规性)) !np (!n (文件)))))))) wco (，) vp (c (然后) !vp (!v (有) np (ap (!b (专门)) !np (!n (队伍))) vp (!vp (!v (进行)) vp (!vp (!v (监督)) vp (!v (检查))))))) wfs (。)))[3] (3)

Due to the different annotation schemes and formats used in these two treebanks, we proposed the following strategies to build the gold-standard data set for Task 2-1 and Task 2-2:

1) Unify POS tag set

The PKU-CTB has 97 POS tags, and TCT has 70 POS tags. After analyzing these POS tags, we found most of them have same meanings. So we designed a unified POS tag set with 58 intersected tags. All the POS tags used in PKU-CTB and TCT can be automatically mapped to this unified tag set.

2) Transform PKU-CTB annotations

Firstly, we mapped the POS tags into the unified tag set, and transformed the word and POS tag format into TCT's format. Then, we deleted all unary constituents in PKU-CTB parse trees and transferred the constituent structures and tags into TCT's constituent tags. Finally, we manually proofread the transformed parse trees to modify some constituent structures that are inconsistent with TCT annotation scheme. About 5% constituents are modified.

3) Extract EDCs and event annotations from TCT

Based on the detailed grammatical relation tags annotated in TCT, we can easily extract each EDC for a TCT sentence (Zhou and Zhu, 2009). Then, we proposed an algorithm to extract different event constructions in each EDC and build a large scale Chinese event bank. It can be used as a gold-standard data to evaluation the event recognition performance of an automatic syntactic parser in Task 2-1.

An event construction is an event chunk serial controlled by an event target verb. It is a basic unit to describe event content. For example, for the first EDC extracted from the above sentence (1), we can obtain the follow four event constructions for the event target verb '见到', '伐倒', '修', and '伐倒' .

- [D-sp 沿途/s-@] ， /wP [S-np 我们/rNP-@] [D-dp 不时/d-@] [P-vp-Tgt 见到/v-@] [O-np 因/p 更新/v 而/cC 伐倒/v 的/uJDE 树木/n-@ ，/wP 因/p 修/v 路/n 需/vM 伐倒/v 的/uJDE 树木/n-@][4]
- [D-pp 因/p 更新/v-@] [P-vp-Tgt 需/vM 伐倒/v-@] 的/uJDE [H-np 树木/n-@] …
- … 因/p [P-vp-Tgt 修/v-@] [O-np 路/n-@] 需/vM 伐倒/v 的/uJDE 树木/n
- … [D-pp 因/p 修/v-@ 路/n] [P-vp-Tgt 需/vM 伐倒/v-@]的/uJDE [H-np 树木/n-@]

4) Obtain TCT constituent structure trees

We can easily select all syntactic constituent tags annotated in TCT sentences to build the gold-standard parsing trees for Task 2-2.

We mainly used the journalistic and academic texts annotated in TCT and PKU-CTB to build different training and test set for task 2-1 and 2-2. Table 1 summarizes current building status of these gold-standard data sets.

[3] The PKU-CTB uses the similar POS and constituent tags with TCT scheme. The exclamation symbol '!' is used to annotate the head of each constituent in the parse tree.

[4] Each event chunk is annotated with bracket pairs with functional and constituent tags. Some functional tags used in the EDCs are as follows: D—adverbial, S—subject, P—predicate, O—object. The constituent tags are same with that ones used in above parse tree. The head of each chunk is indicated through '-@'.

Data set	Source	Genre	Methods
2-1, TR	TCT ver1.0	News, Academy	POS unification, EDC and event extraction
2-1, TS	TCT-2010	News	POS unification, EDC and event extraction
2-2, TR	TCT ver1.0	News, Academy	POS unification, Parse tree extraction
2-2, TS	PKU-CTB	Academy	POS unification, annotation transformation

Table 1 Gold-standard data building status (TR=Training data, TS=Test data)

We selected all news and academic texts annotated in TCT ver1.0 to form the training set of Task 2-1 and 2-2. 1000 EDCs extracted from TCT-2010 were selected as the test set of Task 2-1. These sentences are extracted from the People's Daily corpus with the same source of TCT ver1.0. 1000 sentences extracted from PKU-CTB were selected as the test set of Task 2-2. Most of them are extracted from the technical reports or popular science articles. They have much more technical terms than the encyclopedic articles used in TCT ver1.0. Table 2 shows the basic statistics of all the training and test sets in Task 2.

Data set	Word Sum	Sent. Sum	Average Length
2-1, TR	425619	37219	11.44
2-1, TS	9182	1000	9.18
2-2, TR	481061	17529	27.44
2-2, TS	26381	1000	26.38

Table 2 Basic statistics of Task 2

4 Evaluation metrics

For Task 2-1, we designed three kinds of evaluation metrics:

1) POS accuracy (POS-A)

This metri is used to evaluate the performance of automatic POS tagging. Its computation formula is as follows:

- POS accuracy = (sum of words with correct POS tags) / (sum of words in gold-standard sentences) * 100%

The correctness criteria of POS tagging is as follows:

- ✧ The automatically assigned POS tag is same with the gold-standard one.

2) Constituent parsing evaluation

We selected three commonly-used metrics to evaluation the performance of constituent parsing: labeled precision, recall, and F1-score. Their computation formulas are as follows:

- Precision = (sum of correctly labeled constituents) / (sum of parsed constituents) * 100%
- Recall = (sum of correctly labeled constituents) / (sum of gold-standard constituents) *100%
- F1-score = 2*P*R / (P+R)

Two correctness criteria are used for constituent parsing evaluation:

- ✧ 'B+C' criteria: the boundaries and syntactic tags of the automatically parsed constituents must be same with the gold-standard ones.
- ✧ 'B+C+H' criteria: the boundaries, syntactic tags and head positions of the automatically parsed constituents must be same with the gold-standard ones.

3) Event recognition evaluation

We only considered the recognition recall of each event construction annotated in the event bank, due to the current parsing status of Task 2-1 output. For each event target verb annotated in the event bank, we computed their Micro and Macro average recognition recall. The computation formulas are as follows:

- Micro Recall = (sum of all correctly recognized event constructions) / (sum of all gold standard event constructions) * 100%
- Macro Recall = (sum of Micro-R of each event target verb) / (sum of event target verbs in gold-standard set)

The correctness criteria of event recognition should consider following two matching conditions:

Condition 1: Each event chunk in a gold-standard event construction should have a corresponding constituent in the automatic parse tree. For the single-word chunk, the automatically assigned POS tag should be same with the gold standard one. For the multiword chunk, the

boundary, syntactic tag and head positions of the automatically parsed constituent should be same with the gold-standard ones. Meanwhile, the corresponding constituents should have the same layout sequences with the gold standard event construction.

Condition 2: All event-chunk-corresponding constituents should have a common ancestor node in the parse tree. One of the left and right boundaries of the ancestor node should be same with the left and right boundaries of the corresponding event construction.

For Task 2-2, we design two kinds of evaluation metrics:

1) POS accuracy (POS-A)

This index is used to evaluate the performance of automatic POS tagging. Its formula and correctness criteria are same with the above definitions of Task 2-1.

2) Constituent parsing evaluation

To evaluate the parsing performance of event relation recognition in complex Chinese sentences, we firstly divided all parsed constituents into following two parts:

- Constituent of complex sentence (C_S), whose tag is 'fj';
- Constituents in simple sentence (S_S), whose tags are belong to the tag set {dj, vp, ap, np, sp, tp, mp, mbar, dp, pp, bp}.

Then we computed the labeled precision, recall and F1-socre of these two parts and obtain the arithmetic mean of these two F1-score as the final ranking index. Their computation formulas of each part are as follows:

- Precision = (sum of correctly labeled constituents in one part) / (sum of parsed constituents in the part) * 100%
- Recall = (sum of correctly labeled constituents in one part) / (sum of gold-standard constituents in the part) *100%
- F1-score = 2*P*R / (P+R)
- Total F1-Score = (C_S F1 + S_S F1) / 2

We use the above 'B+C' correctness criteria for constituent evaluation in Task 2-2.

ID	Participants	Task 2-1			Task 2-2		
		TPI	Open	close	TPI	open	Close
01	School of Computer Sci. and Tech., Harbin Institute of Technology	Y			Y		1
02	Knowledge Engineering Research Center, Shenyang Aerospace Univ.	Y		3	Y		2
03	Dalian University of Technology	Y		1	Y		1
04	National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Science	Y	2	2	Y	4	2
05	Beijing University of Posts and Telecommunications	Y		2	Y		
06	University of Science and Technology of China	Y			Y		
07	Dept. of Computer Science and Technology, Shanghai Jiao Tong University,	Y		3	Y		3
08	Soochow University	Y			Y		
09	Harbin Institute of Technology	Y		1	Y		
10	German Research Center for Artificial Intelligence	Y	1	1	Y	1	
11	China Center for Information Industry Development	N			Y	1	
12	City University of Hong Kong	Y			Y		
13	National Central University	Y			Y		
Total		12	3	13	13	6	9

Table 3 Result submission data of all participants in Task 2. (TPI=Take Part In)

5 Evaluation results

The Task 2 of CIPS-Bakeoff-2010 attracted 13 participants. Almost all of them took part in the two subtasks: Task 2-1 and 2-2. Only one participant took part in the Task 2-2 subtask alone.

Among them, 9 participants submitted parsing results. In Task 2-1, we received 16 parsing results, including 13 close track systems and 3 open track systems. In Task 2-2, we received 15 parsing results, including 9 close track systems and 6 open track systems. Table 3 shows the submission information of all participants of Task 2.

5.1 Task 2-1 analysis

We evaluated the parsing performance of EDC on the constituent and event level respectively. The constituent parsing evaluation only considers the parsing performance of one single constituent. The event recognition evaluation will consider the recognition performance of a complete event construction. So it can provide more useful reference information for event extraction application.

Table 5 and Table 6 show the evaluation results of constituent parsing in the close and open tracks respectively. In the close track, the best F1-score under 'B+C' criteria is 85.39%, while the best F1 score under 'B+C+H' criteria is 83.66%. Compared with the evaluation results of the task 5 in CIPS-ParEval-2009 under the similar training and test conditions (Zhou and Li, 2009), the performance of head identification is improved about 2%. Table 4 shows the detailed comparison data.

Rank	ID	'B+C'	'B+C+H'	POS-A
09-1	08	87.22	83.70	Gold
09-2	15	86.25	81.75	Gold
10-1	02	85.39	83.66	93.96
10-2	04	84.36	82.51	91.84

Table 4 F1 scores of the Top-2 single-model close-track systems in the ParsEval-2009 and ParsEval-2010.

Table 7 and Table 8 show the evaluation results of event recognition in the close and open tracks respectively. When we consider the complete event constructions contained in a parse tree, the best Macro-Recall is only about 71%. There are still lots of room to improve in the future.

ID	Sys-ID	Model	'B+C'			'B+C+H'			POS-A	Rank
			P	R	F1	P	R	F1		
02	SAU01	Single	85.42	85.35	85.39	83.69	83.63	83.66	93.96	1
02	SAU02	Single	85.02	85.11	85.06	83.21	83.31	83.26	93.96	2
04	a	Single	84.40	84.32	84.36	82.55	82.47	82.51	91.84	3
04	b	Single	83.79	83.74	83.76	81.82	81.78	81.80	91.67	4
10	DFKI_C	Single	82.93	82.85	82.89	80.54	80.46	80.50	81.99	5
02	SAU03	Single	80.28	79.31	79.79	78.55	77.61	78.08	93.93	6
07	b	Single	78.61	78.76	78.69	76.61	76.75	76.68	92.77	7
07	c	Single	77.78	78.13	77.96	75.78	76.13	75.95	92.77	8
05	BUPT	Single	74.86	76.05	75.45	71.06	72.20	71.63	87.00	9
05	BUPT	Multiple	74.48	75.64	75.05	70.72	71.81	71.26	87.00	10
03	DLUT	Single	71.42	71.19	71.30	69.22	69.00	69.11	86.69	11
09	InsunP	Single	70.69	70.48	70.58	67.07	66.87	66.97	77.87	12
07	a	Single	9.09	12.51	10.53	7.17	9.88	8.31	7.02	13

Table 5 Constituent parsing evaluation results of Task 2-1 (Close Track), ranked with 'B+C+H'- F1

ID	Sys-ID	Model	'B+C'			'B+C+H'			POS-A	Rank
			P	R	F1	P	R	F1		
04	a	Single	86.07	86.08	86.08	84.27	84.28	84.27	92.51	1
04	b	Single	83.79	83.74	83.76	81.82	81.78	81.80	91.67	2
10	DFKI_C	Single	82.37	83.05	82.71	79.99	80.65	80.32	81.87	3

Table 6 Constituent parsing evaluation results of Task 2-1 (Open Track), ranked with 'B+C+H'- F1

ID	Sys-ID	Model	Micro-R	Macro-R	POS-A	Rank
02	SAU01	Single	72.47	71.53	93.96	1
02	SAU02	Single	72.93	70.71	93.96	2
04	a	Single	67.37	65.05	91.84	3
04	b	Single	67.17	64.23	91.67	4
02	SAU03	Single	63.73	63.54	93.93	5
07	c	Single	63.14	62.48	92.77	6
07	b	Single	62.74	62.47	92.77	7
10	DFKI_C	Single	55.99	53.58	81.99	8
03	DLUT	Single	51.75	53.33	86.69	9
05	BUPT	Single	53.08	48.82	87.00	10
05	BUPT	Multiple	52.88	48.75	87.00	11
09	InsunP	Single	43.15	43.14	77.87	12
07	a	Single	1.13	0.79	7.02	13

Table 7 Event recognition evaluation results of Task 2-1 (Close Track), ranked with Macro-R

ID	Sys-ID	Model	Micro-R	Macro-R	POS-A	Rank
04	a	Single	70.62	69.33	92.51	1
04	b	Single	67.17	64.23	91.67	2
10	DFKI_C	Single	54.47	52.25	81.87	3

Table 8 Event recognition evaluation results of Task 2-1 (Open Track), ranked with Macro-R

5.2 Task 2-2 analysis

Table 9 and Table 10 show the evaluation results of constituent parsing in the close and open tracks of Task 2-2 respectively. In each track, the F1-score of the complex sentence recognition is about 5-6% lower than that of the constituents in simple sentences. It indicates the difficultness of event relation recognition in real world Chinese sentences. Some new features need to be explored for them.

Almost all the parsing performances of the systems in the open track are better than that ones in the close track. It indicates some outside language resources may useful for parsing performance improvement. Compared with the commonly-used English Treebank PTB with about 1M words, our current annotated data may be not enough to train a good Chinese parser. We may need to collect more useful treebank data in the future evaluation tasks.

The F1-scores of constituent parsing in simple sentences of Task 2-2 are still about 5-6% lower than that of EDC constituents under 'B+C' criteria in Task 2-1. It indicates some lower level errors may be propagated to up-level constituents during complex sentence parsing. How to restrict the error propagation chains is an interesting issue need to be explored.

5.3 POS tagging analysis

The best POS accuracy in Task 2-1 is 93.96%, approaching to the state-of-art performance of the Task 1 in CIPS-ParsEval-2009, under similar training and test conditions. But the POS accuracy in Task 2-2 is about 3-4% lower than it. A possible reason is that there are lots of unknown words in the test data of Task 2-2. Most of them are technical terms outside the training data lexicon. How to deal with the unknown words is still an open challenge for POS tagging.

6 Conclusions

The paper introduced the task designing ideas, data preparation methods, evaluation metrics and results of the second Chinese syntactic parsing evaluation jointed with SIGHAN Bakeoff tasks.

Some new contributions of the evaluation are as follows:

1) Set a new metric to evaluate the event construction recognition performance in the constituent parsing tree;

ID	Sys-ID	Model	Constituents in S_S			C_S constituent			Total	POS-A	Rank
			P	R	F1	P	R	F1	F1		
04	b	Single	77.79	77.47	77.63	69.55	76.50	72.86	75.24	88.79	1
04	a	Single	77.91	77.54	77.73	68.47	76.90	72.44	75.08	88.95	2
O2	SAU01	Single	78.64	78.73	78.69	70.22	71.62	70.91	74.80	91.05	3
O2	SAU02	Single	78.46	78.34	78.40	69.48	72.42	70.92	74.66	91.03	4
03	DLUT	Single	61.67	59.75	60.69	65.27	67.31	66.27	63.48	79.67	5
01	CHP	Single	70.20	69.64	69.92	53.95	59.47	56.58	63.25	89.62	6
07	b	Single	55.33	59.57	57.37	6.25	0.64	1.16	29.26	89.01	7
07	c	Single	52.57	57.69	55.01	7.47	1.68	2.74	28.88	89.01	8
07	a	Single	0.71	1.00	0.83	0.00	0.00	0.00	0.42	1.39	9

Table 9 Constituent parsing evaluation results of Task 2-2 (Close Track), ranked with Tot-F1 (S_S=simple sentence, C_S=complex sentence)

ID	Sys-ID	Model	Constituents in S_S			C_S constituent			Total	POS-A	Rank
			P	R	F1	P	R	F1	F1		
04	d	Single	80.04	79.68	79.86	70.11	76.50	73.17	76.51	89.59	1
04	a	Single	80.27	79.99	80.13	70.36	75.54	72.86	76.50	89.69	2
04	c	Single	80.25	79.95	80.10	70.40	75.30	72.77	76.44	89.78	3
04	b	Single	80.02	79.68	79.85	69.82	75.62	72.60	76.22	89.75	4
10	DFKI_C	Single	79.37	79.27	79.32	71.06	73.22	72.13	75.72	81.23	5
11*	CCID	Single	/	/	/	/	/	/	/	/	/

Table 10 Constituent parsing evaluation results of Task 2-2 (Open Track), ranked with Tot-F1 (S_S=simple sentence, C_S=complex sentence) There are some data format errors in the submitted results of CCID system (ID=11)

2) Set a separated metric to evaluate the event relation recognition performance in complex Chinese sentence.

Through this evaluation, we found:

1) The event construction recognition in a Chinese EDC is still a challenge. Some new techniques and machine learning models need to be explored for this task.
2) Compared with about 90% F1-score of the state-of-art English parser, the 75% F1-score of current Chinese parser is still on its primitive stage. There is a long way to go in the future.
3) The event relation recognition in real world complex Chinese sentences is a difficult problem. Some new features and methods need to be explored for it.

They lay good foundations for the new task designation in the future evaluation round.

Acknowledgements

Thanks Li Yemei for her hard work to organize the evaluation. Thanks Li Yanjiao and Li Yumei for their hard work to prepare the test data for the evaluation. Thanks Zhu Muhua for making the evaluation tools and processing all the submitted data. Thanks all participants of the evaluation.

The work was also supported by the research projects of National Science Foundation of China (Grant No. 60573185, 60873173) and National 863 High-Tech research projects (Grant No. 2007AA01Z173).

References

E. Black, S. Abney, et al. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Speech and natural language: proceedings of a workshop, held at Pacific Grove, California*, page 306.

E. Charniak and M. Johnson. 2005. Coarse-to-fine nbest parsing and MaxEnt discriminative reranking. In *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics*, page 180.

S. Clark and J.R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

D. Klein and C. Manning. 2003. Accurate Unlexicalized Parsing. In *Proc. of ACL-03*.

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

J. Hockenmaier and M. Steedman. 2007. CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

R. Levy and C. Manning. (2003). Is it harder to parse Chinese, or the Chinese Treebank? In *Proc. of ACL-03*.

J. Li, G. Zhou, and H.T. Ng. 2010. Joint Syntactic and Semantic Parsing of Chinese. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1108–1117.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proc. of HLT/EMNLP*, pages 523–530.

Mitchell P.Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, 19(2): 313-330

J. Nivre, J. Hall, J. Nilsson, el.al. 2007. Malt-Parser: A language-independent system for data driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

S. Petrov and D. Klein. 2007. Improved inference for unlexicalized parsing. In *Proc. of NAACL HLT 2007*, pages 404–411.

N. Xue, F. Chiou, and M. Palmer. 2002. Building a large-scale annotated Chinese corpus. In *Proc. of COLING-2002*.

Zhan Weidong, Chang Baobao, Dui Huiming, Zhang Huarui. 2006. Recent Developments in Chinese Corpus Research. *Presented in The 13th NIJL International Symposium, Language Corpora: Their Compliation and Application. Tokyo, Japan*.

Zhou Qiang, 2004. Chinese Treebank Annotation Scheme. *Journal of Chinese Information*, 18(4):1-8.

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

Zhou Qiang, Zhu Jingbo. 2009. Evaluation tasks and data preparation of CIPS-ParsEval-2009, http://www.ncmmsc.org/ CIPS-ParsEval-20