Parsing Simplified Chinese and Traditional Chinese with A Rule-based Parser

Xiangli Wang[1], Terumasa Ehara[2], Yuan Li[3]
[1] Japan Patent Information Organization, Tokyo, Japan
[2] Yamanashi Eiwa College, Yamanashi, Japan
[3] The University of Tokyo, Tokyo, Japan
{xiangli_wang@japio.or.jp, eharate@{yamanashi-eiwa, y-eiwa}.ac.jp, liyuan@is.s.u-tokyo.ac.jp}

1 Introduction

Chinese divides into simplified Chinese and traditional Chinese. Some treebank resources like Penn Chinese Treebank: CTB had been built for training simplified Chinese parser (Yu, et al. 2010) while Sinica Treebank was developed for parsing traditional Chinese (Chen et al., 1999). Limit to our knowledge, there are still not grammatical resources that analyze both simplified Chinese and traditional Chinese.

A rule-based Chinese grammatical resource --- Chinese Sentence Structure Grammar: CSSG had been developed based on the idea of Sentence Structure Grammar: SSG (Wang et al., 2012). We assume that a rule-based grammatical resource should analyze both simplified Chinese and traditional Chinese if there are no obvious differences between their grammatical constructions. Aiming at verifying this assumptions, we parse the test sentences from the simplified Chinese parsing task (task 3) and the traditional Chinese parsing task (task 4) of CLP 2012 with the same rule-based parser that was implemented the grammatical resource CSSG.

CSSG includes two parts of resources: the grammatical rules and a simplified Chinese morphological dictionary. We transfer the simplified Chinese characters of the dictionary to traditional Chinese characters for obtaining a traditional Chinese morphological dictionary. We parse the test sentences of task 3 and task 4 with the same CSSG rules but different morphological dictionaries (simplified or traditional Chinese characters). We convert CSSG parsing trees to TCT-style trees and Sinica-style trees to participate in the evaluations of the two tasks. The experiments show that the CSSG rules can parse both simplified Chinese and traditional Chinese, but the performance of the latter is lower than the former. We noticed that a few traditional Chinese constructions are different from simplified Chinese.

2 Chinese Sentence Structure Grammar

Chinese Sentence Structure Grammar: CSSG is a rule-based Chinese grammatical resource that was developed based on the idea of Sentence Structure Grammar: SSG. SSG is a new idea to formalize grammatical rules. Sentence Structure Grammar has 3 main ideas (Wang et al., 2012):

1) Treat the construction of a sentence as a whole, which consists of a predicate (or more) and its semantic-related constituents.
2) Classify predicate verbs according to their semantic properties.
3) Indicate the semantic relations between predicate and its semantic-related constituents directly on parsing tree.

SSG is a kind of context-free grammar, but it differs from Phrase Structure Grammar: PSG: 1) the latter describes a sentence with some context-free phrase rules, but the former treats a sentence as a whole sentential construction, which consists of a predicate (or more) and its semantically-related constituents; 2) the former classify predicate verbs according to their semantic properties. For instance, as shown in figure 1, "停/park" and "飞/fly" have different semantic properties. "停/park" is a kind of verb that needs an agent, an object and a location. In contrast, "飞/fly" is a kind of verb that needs an agent and a direction. Predicate verbs can be classified according to their semantic properties; 3) the latter does only syntactic analysis while the former do syntactic analysis and semantic analysis simultaneously. The semantic role set of SSG should be designed based on the idea of the deep cases in Case Grammar, which a linguistic theory proposed by Fillmore (1968).

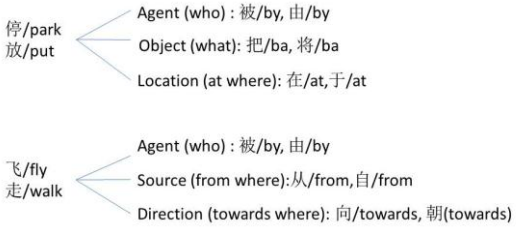

Figure 1: the semantic properties of two types of verbs

For instance, a) is a passive construction. b) is the PSG rule set while c) is the SSG rule set to analyze a). Figure 2 and figure 3 show the SSG parsing tree and the PSG parsing tree of a). As shown in figure 2, the SSG parsing tree provide not only syntactic information like "np" and "sp" but sematic roles, like "Agent", "Object" and "Location", which indicate the semantic relations between the predicate and its semantic-related constituents. Syntactic parsing and sematic parsing can be done simultaneously with the formal grammatical framework SSG.

a. 车/car 被/by 约翰/John 停/park 在/at 停车场/car-park
The car is parked at the car-park by John

b. Rule1:s→ np vp
Rule2: vp→ pp vp
Rule3: vp→ v pp
Rule4: pp→p np
Rule5: np→ n
Rule6: sp→ sq

c. Rule1: s → Object bei Agent Vaol at Location
Rule2: Object→ np
Rule3: Agent→ np
Rule4: Location→ sp
Rule5: np→ n
Rule6: sp→ sq

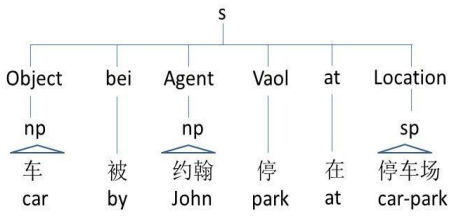

Figure 2: the SSG parsing tree of (a)

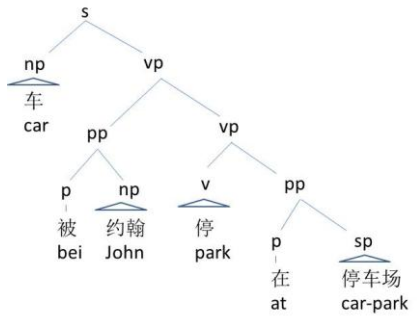

Figure 3: the PSG parsing tree of (a)

3 Comparison between TCT, Sinica Treebank and CSSG

3.1 Tsinghua Chinese Treebank and CSSG

Tsinghua Chinese Treebank: TCT (Zhou, 2004) is used as the training data for the simplified Chinese parsing task. TCT and CSSG are very different grammatical resources.

	CSSG	TCT
Formalism	SSG	PSG
Form	Grammatical rules	Treebank
Word segmentation criteria	Original	Original
POS tag set	Original	Original
Phrase tag set	Original	Original
Semantic role set	Original	none

Table 1: the differences between CSSG and TCT

Their main differences are: 1) they were developed based on different formal grammatical framework. As shown in figure 2 and 3, the former is based on Context-free Phrase Structure Grammar: PSG while the latter is based on another kind of Context-free grammar formalism idea---Sentence Structure Grammar: SSG. Since PSG parses sentences in syntactic level but SSG analyze sentences more deeply, CSSG provides both syntactic information and semantic roles while TCT shows only syntactic information; 2) CSSG is a rule-based grammatical resource while TCT is a Treebank. The designers and developers of the treebanks are usually different people. The designers draw up the annotation scheme first, then the developers annotate parsing trees according to the annotation scheme and their own intuition; in contrast, the designer and the developer of CSSG is the same person who designed and developed the CSSG rules introspectively to cover most simplified Chinese constructions; 3) both of them define the word segmentation criteria and POS tag set originally. For instance, as shown in figure 4 and figure 5, TCT treats "来自/come-from" as one verb while CSSG treats "来自/come-from" as two words: "来/come" is a predicate verb and "自/from" is treated as a case-marker that mark a source case; 4) they design the phrase tag set originally. As shown in figure 4 and 5, verb phrases appear in TCT while there are no verb phrases in CSSG; their definitions of prepositional phrase are different; as shown in figure 7: CSSG and 6: TCT, both e) and f) are treated as locative phrases in TCT while e) is treated as a locative phrase and f) is treated as a temporal phrase in CSSG. Table 4 shows the differences between TCT and CSSG briefly.

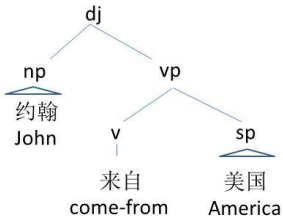

Figure 4: the TCT parsing tree of (d)

d. 约翰/John 来/come 自/from 美国/America
John comes from America

e. 桌子/table 后/behind
Behind the table

f. 回/go-back 家/home 后/after
After going back home

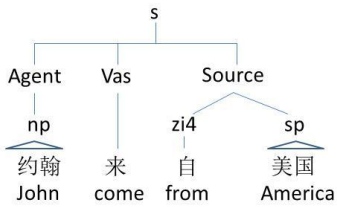

Figure 5: the CSSG parsing tree of (d)

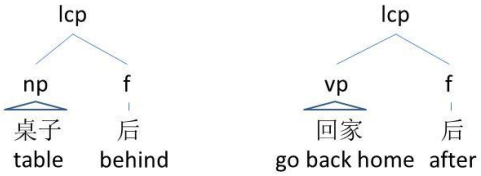

Figure 6: the TCT parsing trees of (e) and (f)

3.2 Sinica Treebank and CSSG

Sinica Treebank (Chen et al., 1999) is used as the training data for the traditional Chinese parsing task. CSSG are quite different from Sinica Treebank.

g. 那個/that 人/person 把/ba 老鼠/rat 帶/take 回/ back-to 茅屋/cottage
That man takes the rat back to the cottage

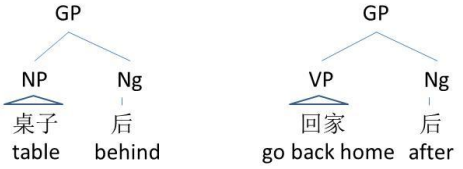

Figure 7: the Sinica parsing trees of (e) and (f)

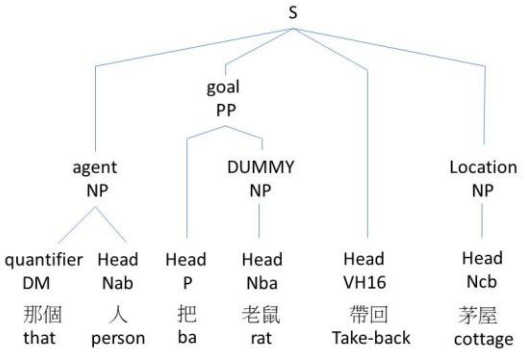

Figure 8: the Sinica parsing tree of (g)

They differ from each other in 6 respects: 1) Sinica Treebank consists of traditional Chinese parsing trees while CSSG is developed for covering simplified Chinese constructions; 2) the former is a rule-based grammatical resource while the latter is a Treebank; 3) both Sinica Treebank and CSSG represent syntactic and semantic information simultaneously, but their formal grammatical framework are different. Sinica Treebank is based on Information-based Case Grammar: ICG, which is a kind of unification-based formalism, and describe syntactic and semantic information in lexical entries (Chen and Huang, 1990); in contrast, CSSG is based on Sentence Structure Grammar: SSG, which is a kind of context-free grammar formalism that indicate both syntactic and semantic constraints in grammatical rules directly; 4) they define the word segmentation criteria and POS tag set originally. For instance, as figure 8 and 9 shown, "那個/that" is treated as one word in Sinica Treebank, but treated as two words in CSSG. "帶回/take-back" is one word in Sinica Treebank while it is split into a verb "帶/take" and a case-marker "回/back" that marks a goal case "茅屋/cottage" in CSSG; 5) they define the phrase tag set originally. For instance, the word "后" can lead not only a locative constituent like e) but a temporal constituent such as f). In Sinica Treebank, Both e) and f) are analyzed as a locative phrase (shown in figure 7); in contrast, the locative constituent is treated as a locative phrase while the temporal constituent is treat a temporal phrase in CSSG; 6) they define semantic role set originally. Their designs of the semantic role sets are very different. Figure 8 shows the Sinica-tree while figure 9 represents the CSSG tree of g). "老鼠/rat" is treated as a goal case and "茅屋/cottage" is analyzed as a location case in Sinica Treebank while "老鼠/rat" is regarded as an object case and "茅屋/cottage" is analyzed as a goal case in CSSG. Table 2 shows the differences between these two resources briefly.

	CSSG	Sinica Treebank
Character	Simplified	Traditional
Formalism	SSG	ICG
Form	Grammatical rules	Treebank
Word segmentation criteria	Original	Original
POS tag set	Original	Original
Phrase tag set	Original	Original
Semantic role set	Original	Original

Table 2: the differences between CSSG and Sinica Treebank

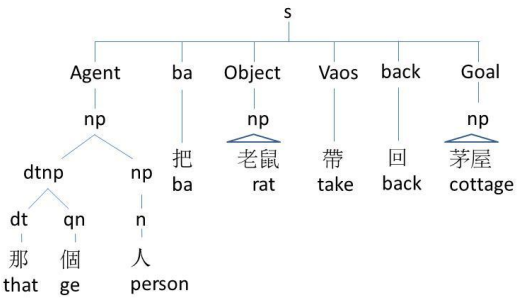

Figure 9: the CSSG parsing tree of (g)

4 Experimental Results

4.1 Experimental Setting

CSSG includes the grammatical rules and a simplified Chinese morphological dictionary. For parsing the test sentences from both simplified Chinese parsing task and traditional Chinese parsing task, we transfer the simplified Chinese characters of the dictionary of CSSG to traditional Chinese characters to obtain a traditional Chinese morphological dictionary.

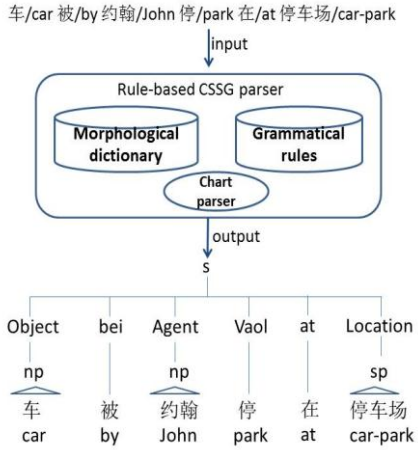

Figure 10: the input and output of (a) of the CSSG parser

We parse simplified and traditional Chinese test sentences with the same grammatical rules and the different morphological dictionaries. Since the scale of the dictionaries is not large enough, there are some

unknown words for CSSG in both test data of simplified and traditional Chinese. We add the unknown words to CSSG dictionaries before parsing.

As figure 10 shown: 1) the CSSG parser consists of three parts: the grammatical rules, a morphological dictionary and a chart parsing engine; 2) the input is a word-segmented sentence and the output is a CSSG parsing tree; 3) since there is not yet a postager based on CSSG, we have to parse all possible POS tag lists of a sentence with the CSSG parser.

After parsing the test data, we convert the CSSG parsing trees and make them are as similar as possible to TCT trees and Sinica-Treebank trees.

4.2 Evaluation Results

Table 3 and 4 summarize the evaluation results of the simplified Chinese parsing. Table 3 shows the results of the constituent boundary recognition. Table 4 represents the evaluation results of the parsing (both phrase boundaries and phrase labels recognition).

correct	gold	system	P	R	F1
85	92	158	53.8%	92.4%	68.0%

Table 3: the result for phrase boundary recognition

correct	gold	system	P	R	F1
85	92	158	42.4%	72.8%	53.6%

Table 4: the result of the simplified Chinese parsing task

Table 5 summarizes the evaluation results of the traditional Chinese parsing.

Micro-averaging			Macro-averaging		
P	R	F1	P	R	F1
47.7%	40.1%	43.6%	53.6%	42.0%	47.1%

Table 5: the results of the traditional parsing parsing task

4.3 Discussion

As we anticipated, the evaluation results are lower than the real performance of the CSSG parser.

There are two reasons should be considered: 1) because of the large differences between the design of CSSG and the two gold data: Sinica Treebank and CSSG, it is impossible to convert some CSSG trees to TCT trees or Sinica-Treebank trees. For instance, f) is treated as a temporal phrase in CSSG, so it does not correspond to any phrase in TCT or Sinica Treebank; 2) there is much inaccuracy in tree-conversion works. As shown in table 3 and 4, the system phrase counts is 158, that is much more than the gold phrase counts 92 so that the recall scores (92.4% and 72.8%) are much higher than the precision scores (68.0% and 53.6%). We checked the evaluation data and found that we converted noun phrases of CSSG like h) to TCT format like i), which might be counted as two noun phrases.

h. (np (nnp (n 葡萄牙) (n 政府)))
i. (np (np (n 葡萄牙) (n 政府)))

As discussed above, the evaluation results do not reflect the real performance of the CSSG parser because of the large differences between CSSG and the two gold data. We expect that more neutral evaluation metrics would be drawn up for the open parsing task.

The experiments show that the evaluation results of the traditional Chinese parsing task are lower than the simplified Chinese parsing task. One of the possible reasons is that there are some differences between the constructions of simplified Chinese and traditional Chinese. We noticed that a few traditional Chinese constructions differ from simplified Chinese.

5 Conclusion and Future Work

In this paper, we introduced a broad-coverage rule-based Chinese grammatical resource CSSG, which was developed based on a new grammar formalism idea: Sentence Structure Grammar; we compared briefly CSSG with a simplified Chinese Treebank TCT and a traditional Chinese resource Sinica Treebank; we also introduced our participation of CIPS-SIGHAN-2012 parsing task. We use a same rule-based chart parser implemented CSSG to participate in both simplified Chinese parsing task and traditional Chinese parsing task. The experiment shows that the rule-based grammatical resource CSSG that was developed for covering simplified Chinese constructions can also parse traditional Chinese sentences with a lower performance.

References

Chen Feng-Yi, Pi-Fang Tsai, Keh-Jiann Chen, Chu-Ren Hunag. 1999. *The Construction of Sinica Treebank*. Computational Linguistics and Chinese Language Processing, vol. 4, No. 2. pp.87-104.

Chen Keh-jiann and Chu-Ren Huang. 1990. *Information-based Case Grammar*. Proceedings of the 13th Conference on Computational Linguistics, Volume 2, pages 54-59.

Fan Xiao. 1998. *The types of Chinese Sentences (In Chinese)*. Shanxi Shuhai Press.

Fillmore, Charles J. (1968). *The Case for Case*. In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88.

Liu Yuehua, Wenyu Pan and Wei Gu. 2001. *Practical Modern Chinese Grammar (In Chinese)*. Beijing: Commercial Press.

Wang Xiangli, Yusuke Miyao and Yuan Li. 2012. *Chinese Grammatical resources based on Sentence Structure Grammar and its application on patent field (In Japanese)*. Proceeding of Japan Natural Language Processing. 2012.

Xue Nianwen and Fei Xia. 2000. *The bracketing Guidelines for the Penn Chinese Treebank Project*. Technical Report IRCS 00-08, University of Pennsylvania.

Yu Kun, Yusuke Miyao, Takuya Matsuzaki, Xiangli Wang, Yaozhong Zhang, Kiyotaka Uchimoto, Junichi Tsujii. *Comparison of Chinese Treebanks for Corpus-oriented HPSG Grammar Development.* Journal of Natural Language Processing (Special Issue on Empirical Methods for Asian Language Processing). April 2010.

Zhu Dexi. 1982. *Lecture Notes on Grammar (In Chinese)*. Beijing: Commercial Press.

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