

Traditional Chinese Parsing Evaluation at SIGHAN Bake-offs 2012

Yuen-Hsien Tseng¹, Lung-Hao Lee¹ and Liang-Chih Yu²

¹Information Technology Center, National Taiwan Normal University, Taipei, Taiwan.

²Department of Information Management, Yuan Ze University, Chung-Li, Taiwan

{samtseng, lhlee}@ntnu.edu.tw, lcyu@saturn.yzu.edu.tw

Abstract

This paper presents the overview of traditional Chinese parsing task at SIGHAN Bake-offs 2012. On behalf of task organizers, we describe all aspects of the task for traditional Chinese parsing, i.e., task description, data preparation, performance metrics, and evaluation results. We summarize the performance results of all participant teams in this evaluation, in the hope to encourage more future studies on traditional Chinese parsing

1 Introduction

The Association of Computational Linguistics (ACL) is the international scientific and professional society for people working on problems involving natural language and computation. There are about 20 Special Interest Groups (SIG) within ACL. Among these SIGs, SIGHAN provides an umbrella for researchers in industry and academia working in various aspects of Chinese language processing. Bake-offs are important events in SIGHAN, which provides Chinese evaluation platforms for developing and implementing various approaches to solve specific Chinese language issues.

Chinese parsing has been a resurged research area in recent years thanks to the commercial needs in mobile applications, and there is a pressing need for a common evaluation platform where different approaches can be fairly compared. Relevant events include the CoNLL-X (the 10th Conference on Computational Natural Language Learning, 2006) shared task, which evaluates multilingual dependency parsing techniques. This shared task provides the community with a benchmark for evaluating their parsers across different languages. The Chinese data is derived from the Sinica Treebank (Huang et al,

2000; Chen et al., 1999; Chen et al. 2003), which is regarded as the first data set designing for traditional Chinese parsing evaluation. The CoNLL 2007 shared task was the second year event devoted to dependency parsing. The task consists of two separate tasks: a multi-lingual track and a domain adaption track. The designed ideas of the shared task are motivated by the expectation that a parser should be trainable for any language, possibly by adjusting some parameters. The traditional Chinese data set can be used in this multilingual parsing evaluation.

At SIGHAN Bake-offs 2012, we organize the *Traditional Chinese Parsing* task that provides an evaluation platform for developing and implementing traditional Chinese parsers. The hope is that through such evaluation campaigns, more advanced Chinese syntactic parsing techniques will emerge, more effective Chinese language processing resources will be built, and the state-of-the-art techniques will be further advanced as a result.

On behalf of the task co-organizers, we give an overview of *Traditional Chinese Parsing* task at SIGHAN Bake-offs 2012, which is held within the second CIPS-SIGHAN joint conference on Chinese Language Processing (CLP 2012). The rest of this article is organized as follows. Section 2 describes the details of designed tasks, consisting of two sub-tasks, i.e. sentence parsing and semantic role labeling. Section 3 introduces the preparation procedure of data sets. Section 4 proposes the evaluation metrics for both sub-tasks. Section 5 presents the results of participants' approaches for performance comparison. Finally, we conclude this paper with the findings and future research direction in the Section 6.

2 Task Description

For the *Traditional Chinese Parsing* task (Task 4) of Bake-offs 2012, we designed two sub-tasks: 1) Task 4-1: *Sentence parsing* for evaluating the

ability of automatic parsers on complete sentences in real texts. 2) Task 4-2: *Semantic role labeling* for evaluating the ability of automatic parsers on labeling semantic roles.

Each sub-task is separated as closed and open track. In the *Closed Track*, the participants can only use the training data provided by the organizers. In the *Open Track*, the participants can use any data sources in addition to the training data provided by the organizers. Submitted runs in these two tracks will be evaluated separately.

In addition, single systems and combined systems will also be evaluated separately in both tracks. *Single Systems* are parsers that use a single parsing model to accomplish the parsing task. *Combined Systems*, in comparison, are allowed to combine multiple models to hopefully improve performance. For example, collaborative decoding methods will be regarded as a combination method.

We further describe the details and give the examples of both sub-tasks as follows:

2.1 Sentence parsing

The goal of this sub-task is designed to evaluate the ability of automatic parsers on complete sentence parsing in real texts. Complete Chinese sentences with gold standard word segmentation are given as input, in which the word count of each sentence should be greater than 7. The designed parser should assign a POS tag to each word and recognize the syntactic structure of the given sentence as the output result.

The evaluation data sets are derived based on Sinica Treebank. The goal of Sinica Treebank is to provide a syntactic and structure-tagged corpus for improving the performance of automatic parsers by learning the syntactic knowledge. The complete set of part-of-speech tags is defined in the technical report #93-5 (CKIP, 1993). The structural information is defined as the phrase labels for representing syntactic knowledge. The complete set of phrase labels is defined in the construction process (Chen et al, 1999). We give the following example for more information:

- Input: 他 刊登 一則 廣告 在 報紙 上
- Output: S(agent:NP(Nh:他)| Head:VC: 刊登|theme:NP(DM:一則| Na:廣告)| location:PP(P:在|GP(NP(Na:報紙)|Ng:上)))

In this sub-task, we only focus on evaluating the ability of automatic parsers on syntactic structure recognition. That is, the **boundary** and **phrase label** of a **syntactic constituent** should

be completely identical with the gold standard, which is regarded as a correctly recognized case. The semantic roles and part-of-speech tags in the output format will be ignored in this sub-task.

2.2 Semantic role labeling

In addition to syntactic information, the Sinica Treebank also contains semantic roles of each constituent. Hence, we design this sub-task for evaluating the ability of automatic parsers on labeling semantic roles. In this sub-task, the given input sentences are the same as the sentence-parsing sub-task. The parser should assign a semantic role of each top-level constituent. There are 74 abstract semantic roles including thematic roles, e.g. “agent” and “theme”, the second roles of “location”, “time” and “manner”, and roles for nominal modifiers. The complete set of semantic roles is described in the related study (You & Chen, 2004). We also give the example shown as the follows:

- Input: 母親 帶 他們 到 溪 邊 去 釣 魚
- Output: S(agent:NP(Na:母親)|Head:VC: 帶|theme:NP(Nh:他們)|location:PP(P:到|NP(Na:溪|Ncd:邊))|complement:VP(D:去|VA:釣魚))

In this sub-task, we only evaluate the performance of automatic parsers on semantic role labeling. If the **boundary** and **semantic role** of a **syntactic constituent** is completely identical with the gold standard, that is a correct recognition. In the same way, we also ignore the phrase labels and part-of-speech tags in the output format for this sub-task.

3 Data Preparation

The data sets are divided into three distinct ones: 1) Training set: the sentences in this set are prepared for training the designed parsers. 2) Test set: there are 1000 newly developed sentences that are used for formal testing. 3) Validation set: the sentences are adopted for dry run. Table 1 shows the statistics of prepared sets, where #Word and #Sent denote the numbers of words and sentences, respectively. The details are described as follows.

Data Set	#Word	#Sent	Avg. Length
Training	391,505	65,243	6
Test	8,565	1,000	8.57
Validation	341	37	9.2

Table 1: Descriptive statistics of the data sets.

- Training Set

The training set is derived from Sinica Treebank according to sentence lengths and complexities. The original part-of-speech tags in the Treebank are simplified. Only the semantic roles of each top-level constituent are kept, the others are removed. Take the original sentence “S(theme:NP(Head:Nba: 西遊記)|Head:V_11:是|range:NP(property:Ncb: 我國 |property:V · 的(head:VH11: 著名 |Head:DE: 的)|Head:Nac: 小說))” for example, this parsed sentence will be transformed as “S(theme:NP(Nb: 西遊記) |Head:V_11:是|range:NP(Nc:我國 |V · 的(VH: 著名|DE:的)|Na:小說))” for training purpose.

- Test Set

One thousand newly developed sentences were selected from United Daily News Agency news corpus for both sub-tasks to cover different sentence lengths and complexities. Two annotators from the construction team of Sinica Treebank were asked to label the gold standard of the test set. For example, a selected sentence is “聯合國大會今天並未調整會員國出資比例”. Its manually annotated gold standard is “S(agent:NP(Nc:聯合國|Na:大會)|time:Nd:今天|evaluation:D: 並|negation:D: 未|Head:VC: 調整|goal:NP(S(NP(Na: 會員國)|VC: 出資)|Na: 比例))”

- Validation Set

We also prepare additional 37 newly developed sentences as the validation set for dry run. The main purpose of dry run is for output format validation. The participants can submit several runs resulted from different models or parameter settings. During the dry run, each submitted run was evaluated to check whether the output format could be accepted in our developed evaluation tool. The evaluation reports will be returned to the participants to inform the participants whether their output formats are correct and how good are their current performance. With the dry run feedback, the participants can fine-tune their implemented systems to further enhance the performance.

4 Performance Metrics

For the sentence-parsing sub-task, we adopt the Precision (P), Recall (R) and F1 score as metrics for performance evaluation. The computation formulas are listed as follows:

- $P = \# \text{ of correctly recognized constituents} / \# \text{ of all constituents in the parsing output}$
- $R = \# \text{ of correctly recognized constituents} / \# \text{ of all constituents in the gold standard}$
- $F1 = (2 * P * R) / (P + R)$

The criterion for judging correctness is that the *boundary* and *phrase label* of a syntactic constituent should be completely identical with the gold standard. Only six phrase labels (S, VP, NP, GP, PP, and XP) will be evaluated in the test set. The other labels such as “N·的”, “V·地”, and “得·V” will be ignored.

For example, given an input sentence: “他刊登一則廣告在報紙上” and its parsing output of a proposed system: “S(agent:NP(Nh:他)|Head:VC:刊登|theme:NP(DM:一則|Na:廣告)|location:PP(P:在|NP(Na:報紙|Nc:上)))”, the recognized constituents are: S(他刊登一則廣告在報紙上), NP(他), NP(一則廣告), PP(在報紙上), and NP(報紙上). The gold standard of this input sentence is: S(他刊登一則廣告在報紙上), NP(他), NP(一則廣告), PP(在報紙上), GP(報紙上), and NP(報紙). The evaluated tool will yield the following performance metrics:

- $P = 0.8 (=4/5)$ Notes: $\# \{S(\text{他刊登一則廣告在報紙上}), NP(\text{他}), NP(\text{一則廣告}), PP(\text{在報紙上})\} / \# \{S(\text{他刊登一則廣告在報紙上}), NP(\text{他}), NP(\text{一則廣告}), PP(\text{在報紙上}), NP(\text{報紙上})\}$.
- $R = 0.6667 (=4/6)$ Notes: $\# \{S(\text{他刊登一則廣告在報紙上}), NP(\text{他}), NP(\text{一則廣告}), PP(\text{在報紙上})\} / \# \{S(\text{他刊登一則廣告在報紙上}), NP(\text{他}), NP(\text{一則廣告}), PP(\text{在報紙上}), GP(\text{報紙上}), NP(\text{報紙})\}$.
- $F1 = 0.7273 (=2 * 0.8 * 0.6667 / (0.8 + 0.6667))$

For semantic role labeling sub-task, we adopt the same metrics. Similar computations are formulated as follows:

- $P = \# \text{ of correctly recognized roles} / \# \text{ of all roles in the recognized data}$
- $R = \# \text{ of correctly recognized roles} / \# \text{ of all roles in the gold standard data}$
- $F1 = 2 * P * R / (P + R)$

The criterion for judging correctness is that the *boundary* and *semantic role* of a syntactic constituent should be completely identical with the

gold standard. For example, given an input sentence: “母親帶他們到溪邊去釣魚” and its possible parsing output: “S(agent:NP(Na:母親) | Head:VC:帶|agent:NP(Nh:他們)|location:PP(P:到|Na:溪邊)|deontics:D:去|Head:VA:釣魚)”, the recognized semantic roles are: agent(母親), Head(帶), agent(他們), location(到溪邊), deontics(去), and Head(釣魚). The gold standard of this input sentence is: agent(母親), Head(帶), theme(他們), location(到溪邊), and complement(去釣魚). The evaluated tool will yield the following performance metrics:

- $P = 0.5$ ($=3/6$) Notes: $\#\{\text{agent(母親), Head(帶), location(到溪邊)}\} / \#\{\text{agent(母親), Head(帶), agent(他們), location(到溪邊), deontics(去), Head(釣魚)}\}$.
- $R = 0.6$ ($=3/5$) Notes: $\#\{\text{agent(母親), Head(帶), location(到溪邊)}\} / \#\{\text{agent(母親), Head(帶), theme(他們), location(到溪邊), complement(去釣魚)}\}$.
- $F1 = 0.5455$ ($= (2 * 0.5 * 0.6) / (0.5 + 0.6)$)

In addition, we use micro-averaging and macro-averaging to measure the overall performance for both sub-tasks in the test set. Equation (1)~(6) show the formulations for measuring the performance, where P_{micro} , R_{micro} and $F1_{\text{micro}}$ denote micro-averaging precision, recall, and F1 score, respectively; P_{macro} , R_{macro} and $F1_{\text{macro}}$ stand for macro-averaging precision, recall, and F1 score, individually.

$$P_{\text{micro}} = \frac{\sum_{i=1}^{|S|} TP_i}{\sum_{i=1}^{|S|} TP_i + FP_i} \quad (1)$$

$$R_{\text{micro}} = \frac{\sum_{i=1}^{|S|} TP_i}{\sum_{i=1}^{|S|} TP_i + FN_i} \quad (2)$$

$$F1_{\text{micro}} = \frac{2 * P_{\text{micro}} * R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}} \quad (3)$$

$$P_{\text{macro}} = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{TP_i}{TP_i + FP_i} \quad (4)$$

$$R_{\text{macro}} = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$F1_{\text{macro}} = \frac{2 * P_{\text{macro}} * R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}} \quad (6)$$

In the above equations, $|S|$ denotes the number of sentence in the test set; TP is the number of constituents in the gold standard that are correctly recognized in the system output; FN is the number of constituents in the gold standard that are not correctly recognized in the system output; FP is the number of recognized constituents in the system output that are not in the gold standard.

5 Evaluation Results

Table 2 shows the participant teams and their submission statistics. The task 4 of Bakeoffs 2012 attracted 8 research teams. There are 4 teams that come from Taiwan, i.e. CYUT, NCU, NCYU, and NTUT & NCTU. The other 3 teams originate from China, i.e. UM, NEU and PKU. The remaining one is JAPIO from Japan.

Among 8 registered teams, 6 teams submitted their testing results. For formal testing, each participant can submit several runs that use different models or parameter settings. All submitted runs adopt a single parsing model, i.e. *Single System*, to accomplish the evaluated task. In Task 4-1, we received 8 submitted results, including 7 from closed track systems and 1 from an open track system. In Task 4-2, we received 4 submissions, including 3 from closed track systems and 1 from an open track system.

5.1 Analysis of sentence parsing

We evaluated the sentence parsing performance of both tracks separately. Table 3 and Table 4 show the evaluated results in closed track and open track, respectively. For closed track, we implement the baseline system using the Stanford parser (Klein and Manning, 2003; Levy and Manning, 2003) with default parameters for performance comparison. We only adopt the training set to learn the Chinese parsing model. In formal testing phase, there were 75 sentences that cannot be parsed using the re-train Stanford parser. Experimental results indicate that the baseline system achieves micro-averaging and macro-averaging F1 at 0.5822 and 0.5757, respectively.

Parts of the submitted runs perform better than the baseline results. Systems come from NEU-Run1 and NEU-Run2 achieve the best performance, i.e. 0.7078 for micro-averaging F1 and 0.7211 for macro-averaging F1. These two runs have the same syntactic structure, but different semantic role labels. However, only the phrase labels and their boundaries were evaluated in

sub-task 1, so the performance is the same. Note that the NCTU&NTUT-Run1 was submitted a few days after the formal test deadline. However, we also evaluated their results for more information.

Only one team took part in the open track. The performance measures of this submission are micro-averaging F1 score: 0.4355 and macro-averaging F1 score: 0.4287. For performance

comparison, we invited the Chinese Knowledge Information Processing Group (CKIP) in the Institute of Information Science, Academia Sinica, to modify their designed Chinese parser (Yang et al. 2005; 2008; Hsieh et al. 2007) for this evaluation. The CKIP parser achieves the best micro-averaging and macro-averaging F1 scores at 0.7287 and 0.7448, respectively.

ID	Participants	Task 4-1		Task 4-2	
		Open	Closed	Open	Closed
1	Chaoyang University of Technology (CYUT)		1		
2	National Central University (NCU)		1		1
3	National Chiayi University (NCYU)		2		
4	National Chiao Tung University & National Taipei University of Technology (NCTU&NTUT)		1		
5	University of Macau (UM)		1		
6	Northeastern University (NEU)		2		2
7	Peking University (PKU)				
8	Japan Patent Information Organization (JAPIO)	1		1	
Total		1	8	1	3

Table 2: Result submission statistics of all participants in Task 4.

Submitted Runs	Micro-averaging			Macro-averaging		
	Precision	Recall	F1	Precision	Recall	F1
CYUT-Run1	0.6695	0.5781	0.6204	0.6944	0.5999	0.6437
NCU-Run1	0.6215	0.4764	0.5394	0.6317	0.4913	0.5527
NCYU-Run1	0.4116	0.4475	0.4288	0.4354	0.4663	0.4503
NCYU-Run2	0.4167	0.5104	0.4588	0.4352	0.5316	0.4786
*NCTU&NTUT-Run1	0.7215	0.387	0.5038	0.7343	0.4147	0.5301
UM-Run1	0.7165	0.6595	0.6868	0.7229	0.6718	0.6964
NEU-Run1	0.7293	0.6875	0.7078	0.7429	0.7005	0.7211
NEU-Run2	0.7293	0.6875	0.7078	0.7429	0.7005	0.7211
Stanford Parser (Baseline)	0.6208	0.5481	0.5822	0.5885	0.5634	0.5757

Table 3: Sentence parsing evaluation results of Task 4-1 (Closed Track), ordered with participant ID.

Submitted Runs	Micro-averaging			Macro-averaging		
	Precision	Recall	F1	Precision	Recall	F1
JAPIO-Run1	0.4767	0.4008	0.4355	0.5355	0.4195	0.4705
*CKIP Parser (Baseline)	0.7534	0.7057	0.7287	0.7693	0.7218	0.7448

Table 4: Sentence parsing evaluation results of Task 4-1 (Open Track), ordered with participant ID.

5.2 Analysis of semantic role labeling

Table 5 and Table 6 show the evaluation results of semantic role labeling in the closed and open tracks of Task 4-2, respectively. For closed track,

we apply the well-known sequential model Conditional Random Field (CRF) as the baseline system for performance comparison. It scores at 0.4297 for micro-averaging F1 score and 0.4287 for macro-averaging F1 score. NEU’s Run1 and

Run2 perform better slightly than the baseline when micro-averaging F1 is considered, which are 0.4343 and 0.4394, respectively. However, the baseline system achieves the best macro-averaging F1.

For open track, the only one submission achieves 0.2139 and 0.2374 of micro-averaging

and macro-averaging F1 scores, respectively. The CKIP team was also asked to participate in this open track as the baseline system. The modified CKIP parser achieves the best results on labeling semantic roles of each top-level constituent. It accomplishes 0.6034 of micro-averaging F1 score and 0.6249 of macro-averaging F1 score.

Submitted Runs	Micro-averaging			Macro-averaging		
	Precision	Recall	F1	Precision	Recall	F1
NCU-Run1	0.3755	0.3429	0.3585	0.3506	0.3538	0.3522
NEU-Run1	0.4358	0.4328	0.4343	0.4192	0.416	0.4176
NEU-Run2	0.4409	0.4379	0.4394	0.4239	0.4209	0.4224
CRF (Baseline)	0.4382	0.4216	0.4297	0.4347	0.4229	0.4287

Table 5: Semantic role labeling results of Task 4-2 (Closed Track), ordered with participant ID.

Submitted Runs	Micro-averaging			Macro-averaging		
	Precision	Recall	F1	Precision	Recall	F1
JAPIO-Run1	0.2036	0.2255	0.2139	0.2333	0.2417	0.2374
*CKIP Parser (Baseline)	0.6019	0.6049	0.6034	0.6252	0.6247	0.6249

Table 6: Semantic role labeling results of Task 4-2 (Open Track), ordered with participant ID.

6 Conclusions

This paper describes the overview of traditional Chinese parsing evaluation at SIGHAN Bake-offs 2012. We describe the task designing ideas, data preparation details, evaluation metrics, and the results of performance evaluation.

For sentence parsing, the promising parsers achieve about 0.7 of F1 regardless which kind of training data is used to train the parsers. For the sub-task of semantic role labeling, the best system achieves about 0.6 of F1 score.

This Bake-off motivates us to build more Chinese language resources (e.g., modified Treebank and over 1000 new labeled sentences) for reuse in the future to possibly improve the state-of-the-art techniques for Chinese language processing. It also encourages researchers to bravely propose various ideas and implementations for possible break-through. No matter how well their implementations would perform, they contribute to the community by enriching the experience that some ideas or approaches are promising (or impractical), as verified in this bake-off. Their reports in this proceeding will reveal the details of these various approaches and contribute to our knowledge and experience about Chinese language processing.

After this bake-off evaluation, the resources and tools built for this evaluation will be released on the Web for the convenience of future studies.

Acknowledgements

Research fellow Keh-Jiann Chen, the leader of Chinese Knowledge Information Processing Group (CKIP) in Institute of information Science, Academia Sinica, is appreciated for supporting Sinica Treebank. We would like to thank Su-Chu Lin and Shih-Min Li for their hard work to prepare the test set for the evaluation. We would like to thank Kuei-Ching Lee for implementing the baseline systems for performance comparison. We thank Wei-Cheng He for developing the evaluation tools. Finally, we thank all the participants for taking part in the evaluation.

This research was partially supported by the ‘‘Aim for the Top University Project’’ of National Taiwan Normal University, sponsored by the Ministry of Education, Taiwan.

References

- Chinese Knowledge Information Processing Group 1993. Categorical Analysis of Chinese. ACLCLP Technical Report # 93-05, Academia Sinica.
- Chu-Ren Huang, Keh-Jiann Chen, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao, and Kuang-Yu

- Chen. 2000. Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface. In *Proceedings of the 2nd Chinese Language Processing Workshop*, 29-37.
- Dan Klein, and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. 423-430.
- Duen-Chi Yang, Yu-Ming Hsieh, and Keh-Jiann Chen. 2005. Linguistically-Motivated Grammar Extraction, Generalization and Adaptation. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, 177-187.
- Duen-Chi Yang, Yu-Ming Hsieh, and Keh-Jiann Chen. 2008. Resolving Ambiguities of Chinese Conjunctive Structures by Divide-and-Conquer Approaches. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, 715-720.
- Feng-Yi Chen, Pi-Fang Tsai, Keh-Jiann Chen, and Chu-Ren Huang. 1999. The Construction of Sinica Treebank. *International Journal of Computational Linguistics and Chinese Language Processing*, 4(2): 87-104. (in Chinese)
- Jia-Ming You and Keh-Jiann Chen. 2004. Automatic Semantic Role Assignment for a Tree Structure. In *Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing*, 1-8.
- Keh-Jiann Chen, Chu-Ren Huang, Feng-Yi Chen, Chi-Ching Luo, Ming-Chung Chang, Chao-Jan Chen, and Zhao-Ming Gao. 2003. Sinica Treebank: Design Criteria, Representational Issues and Implementation. In *Anne Abeille (Ed.) Treebanks Building and Using Parsed Corpora*, Dordrecht:Kluwer, 231-248.
- Roger Levy and Christopher D. Manning. 2003. Is it Harder to Parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. 439-446.
- Yu-Ming Hsieh, Duen-Chi Yang, and Keh-Jiann Chen. 2007. Improve Parsing Performance by Self-Learning. *International Journal of Computational Linguistics and Chinese Language Processing*, 12(2):195-216.