The Deep Re-annotation in a Chinese Scientific Treebank

Kun Yu[1] **Xiangli Wang**[1] **Yusuke Miyao**[2] **Takuya Matsuzaki**[1] **Junichi Tsujii**[1,3]
1. The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033, Japan
{kunyu, xiangli, matuzaki, tsujii}@is.s.u-tokyo.ac.jp
2. National Institute of Informatics, Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo, 101-8430, Japan
yusuke@nii.ac.jp
3. The University of Manchester, Oxford Road, Manchester, M13 9PL, UK

Abstract

In this paper, we introduce our recent work on re-annotating the deep information, which includes both the grammatical functional tags and the traces, in a Chinese scientific treebank. The issues with regard to re-annotation and its corresponding solutions are discussed. Furthermore, the process of the re-annotation work is described.

1 Introduction

A Chinese scientific Treebank (called the *NICT Chinese Treebank*) has been developed by the National Institute of Information and Communications Technology of Japan (NICT). This treebank annotates the word segmentation, pos-tags, and bracketing structures according to the annotation guideline of the Penn Chinese Treebank (Xia, 2000(a); Xia, 2000(b); Xue and Xia, 2000). Contrary to the Penn Chinese Treebank in news domain, the NICT Chinese Treebank includes sentences that are manually translated from Japanese scientific papers. Currently, the NICT Chinese Treebank includes around 8,000 Chinese sentences. The annotation of more sentences in the science domain is ongoing.

The current annotation of the NICT Chinese Treebank is informative for some language analysis tasks, such as syntactic parsing and word segmentation. However, the deep information, which includes both the grammatical functional tags and the traces, are omitted in the annotation. Without grammatical functions, the simple bracketing structure is not informative enough to represent the semantics for Chinese. Furthermore, the traces are critical elements in detecting long-distance dependencies.

Gabbard et al. (2006) and Blaheta and Charniak (2000) applied machine learning models to automatically assign the empty categories and functional tags to an English treebank. However, considering about the different domains that the Penn Chinese Treebank and the NICT Chinese Treebank belong to, the machine learning model trained on the Penn Chinese Treebank may not work successfully on the NICT Chinese Treebank. In order to guarantee the high annotation quality, in our work, we manually re-annotate both the grammatical functional tags and the traces to the NICT Chinese Treebank. With the deep re-annotation, the NICT Chinese Treebank could be used not only for the shallow natural language processing tasks, but also as a resource for deep applications, such as the lexicalized grammar development from treebanks (Miyao 2006; Guo 2009; Xia 1999; Hockenmaier and Steedman 2002).

Considering that the translation quality of the sentences in the NICT Chinese Treebank may affect the quality of re-annotation, in the current phase, we only selected 2,363 sentences that are of good translation quality, for re-annotation. In the future, with the expansion of the NICT Chinese Treebank, we will continue this re-annotation work on large-scale sentences.

2 Content of Re-annotation

Because the NICT Chinese Treebank follows the annotation guideline of the Penn Chinese Treebank, our re-annotation uses similar annotation criteria in the Penn Chinese Treebank.

Figure 1 exemplifies our re-annotation to a sentence in the NICT Chinese Treebank. In this example, we first re-annotate the trace (as indicated by the italicized part in Figure 1(b)) for the extracted head noun '词/word'. Furthermore, we re-annotate the functional tag of the trace (as indicated by the dashed-box in Figure 1(b)), to indicate that the extracted head noun should be restored into the relative clause as a topic.

There are 26 functional tags in the Penn Chinese Treebank (Xue and Xia, 2000), in which seven functional tags describe the grammatical

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 123–126,
Uppsala, Sweden, 15-16 July 2010.

roles and one functional tag (i.e. LGS) indicates a logical subject. Since the eight functional tags are crucial for obtaining the grammatical function of constituents, we re-annotate the eight functional tags (refer to Table 1) to the NICT Chinese Treebank.

(NP (CP (IP (NP (NN 单词)
(NN 亲密度))
(VP (VA 高)))
(DEC 的))
(NP (NN 词)))
(*the word of which the word cohesion is high*)
(a) A relative clause in the NICT Chinese Treebank

(NP *(CP (WHNP-1 (-NONE- *OP*)*
(CP (IP *(NP-TPC (-NONE- *T*-1))*
(NP (NN 单词)
(NN 亲密度))
(VP (VA 高)))
(DEC 的)))
(NP (NN 词)))
(b) The relative clause after re-annotation

Figure 1. Our re-annotation to a relative clause.

Functional Tag	Description
IO	indirect object
OBJ	direct object
EXT	post-verbal complement that describes the extent, frequency, or quantity
FOC	object fronted to a pre-verbal but post-subject position
PRD	non-verbal predicate
SBJ	surface subject
TPC	topic
LGS	logical subject

Table 1. Functional tags that we re-annotate.

(IP (NP-*TPC-1* (NN 信息))
(VP (ADVP (AD 比较))
(VP (ADVP (AD 容易))
(VP (VV 获得)
*(NP-OBJ (-NONE- *T*-1))*))))
(*It is easier to obtain information.*)
(a) A topic construction with long-distance dependency after re-annotation of functional tag and trace

(IP (NP-*TPC* (DP (DT 该))
(NP (NN 算法)))
(NP-SBJ (NP (PN 其))
(NP (NN 合理性)))
(VP (ADVP (AD 已))
(VP (VV 得到)
(VV 证实))))
(*The rationality of this algorithm has been verified.*)
(b) A topic construction without long-distance dependency after re-annotation of functional tag

Figure 2. Our re-annotation to topic constructions.

In addition, in the annotation guideline of the Penn Chinese Treebank, four constructions are annotated with traces: *BA-construction*, *BEI-construction*, *topic construction* and *relative clause*. The *BEI-construction* and *relative clause* introduce long-distance dependency. Therefore, we re-annotate the traces for the two constructions. The *topic construction* introduces the topic phrase. For the topic constructions that contain long-distance dependency, we re-annotate both the traces and the functional tags (refer to the italicized part in Figure 2(a)). Some topic constructions, however, do not include long-distance dependency. In such cases, we only re-annotate the functional tag to indicate that it is a topic (refer to the italicized part in Figure 2(b)). In addition, the *BA-construction* moves the object to a pre-verbal position. Although the *BA-construction* does not contain long-distance dependency, we still re-annotate the trace to acquire the original position of the moved object in the sentence.

3 Issues and Solutions

3.1 Trace re-annotation in the BA/BEI construction

The NICT Chinese Treebank follows the word segmentation and pos-tag annotation guideline of the Penn Chinese Treebank. Therefore, there are some BA-constructions and BEI-constructions that cannot be re-annotated with traces. The principle reason for this is that the moved object has semantic relations with only part of the verb. For example, in the sentence shown in Figure 3(a), the moved head noun '家乡/hometown' is the object of '建/construct', but not for '建成/construct to be'.

(VP (BA 把)
(IP (NP (NN 家乡))
(VP (VV 建成)
(NP (NN 花园)))))
(*construct the hometown to be a garden*)
(a) The annotation in the NICT Chinese Treebank

(VP (BA 把)
(IP (NP-*SBJ-1* (NN 家乡))
(VP (*VV* 建)
*(NP-OBJ (-NONE- *-1))*
(*AM* 成)
(NP (NN 花园)))))
(b) Our proposed re-annotation of functional tag and trace

Figure 3. Our re-annotation to a BA construction with split verb.

Our analysis of the Penn Chinese Treebank shows that only a closed list of characters (such as '成/to be') can be attached to verbs in such a case. Therefore, we solve the problem by following four steps (for an example, refer to Figure 3(b)):

(1) A linguist manually collects the characters that can be attached to verbs in such a case from the Penn Chinese Treebank and assigns them a new pos-tag 'AM (argument marker)'.

(2) The annotators use the character list as a reference during the re-annotation. When the verb in a BA/BEI construction ends with a character in the list, and the annotators think the verb should be split, the annotators record the sentence ID without performing any re-annotation.

(3) The linguist collects all of the recorded sentences, and defines pattern rules to automatically split the verbs in the BA/BEI constructions.

(4) The annotators annotate trace for the sentences with the split verbs. This step will be finished in our future work.

3.2 Topic detection

In the annotation guideline of the Penn Chinese Treebank, a topic is defined as '*the element that appears before the subject in a declarative sentence*'. However, the NICT Chinese Treebank does not annotate the omitted subject. Therefore, we could not use the position of the subject as a criterion for topic detection.

In order to resolve this issue, we define some heuristic rules based on both the meaning and the bracketing structure of phrases, to help detect the topic phrase. Only the phrase that satisfies all the rules will be re-annotated as a topic. The following exemplifies some rules:

(1) If there is a phrase before a subject, the phrase is probably a topic.

(2) A topic phrase must be parallel to the following verb phrase.

(3) The preposition phrase and localization phrase describing the location or time are not topics.

3.3 Inconsistent annotation in the NICT Chinese Treebank

There are some inconsistent annotations in the NICT Chinese Treebank, which makes our re-annotation work difficult.

These inconsistencies include:

(1) Inconsistent word segmentation, such as segmenting the word '相对应/corresponding' into two words '相对/opposite' and '应/ought'.

(2) Inconsistent pos-tag annotation. For example, when the word '的' exists between two noun phrases, it should be tagged as an associative marker (i.e. DEG), according to the guideline of the Penn Chinese Treebank. However, in the NICT Chinese Treebank, sometimes it is tagged as a nominalizer (i.e. DEC).

(3) Inconsistent bracketing annotation. Figure 4(a) shows the annotation of a relative clause in the NICT Chinese Treebank. In this annotation, the noun phrase '大阪/Osaka 地铁/subway' is incorrectly treated as the extracted head; furthermore, the adverb '人工/by hand' that modifies the verb '制作/make' is incorrectly annotated as an adjective that modifies the noun '变形图/deformation graph'. After correcting these inconsistencies, the relative clause should be annotated as shown in Figure 4(b).

```
(NP (QP (CD 很多))
    (ADJP (JJ 人工))
    (DNP (NP (CP (IP (VP (VV 制作)))
                 (DEC 的))
             (NP (NR 大阪)
                 (NN 地铁)))
         (DEG 的))
    (NP (NN 变形图)))
```

(*many deformation graphs of Osaka subway that are made by hand*)

(a) The inconsistent annotation of a relative clause

```
(NP (QP (CD 很多))
    (NP (CP (IP (VP (ADVP (AD 人工))
                    (VP (VV 制作))))
            (DEC 的))
        (NP (DNP (NP (NR 大阪)
                     (NN 地铁))
                 (DEG 的))
            (NP (NN 变形图)))))
```

(b) The annotation after correcting the inconsistencies

Figure 4. An inconsistent annotation in the NICT Chinese Treebank and its correction.

In our re-annotation, these inconsistently annotated sentences in the NICT Chinese Treebank were recorded by the annotators. We then sent them back to NICT for further verification.

4 Process of Re-annotation

4.1 Annotation Guideline

During the re-annotation, we basically follow the annotation guideline of the Penn Chinese Treebank (Xue and Xia, 2000). However, in order to fit with the characteristics of scientific sentences in the NICT Chinese Treebank, some constraints are added to the guideline.

For example, in the science domain, the relative clause is often used to describe a phenomenon, in which the extracted head noun is usually an abstract noun, and the relative clause is an appositive of the extracted head noun. Figure 5 shows an example in which the relative clause '系统/system 停止/stop 工作/working' is a de-

scription of the extracted head noun '现象/phenomenon'. In such a case, the head noun cannot be restored into the clause. Therefore, we add the following restriction in our re-annotation guideline: *Do not re-annotate the trace when the head noun of a relative clause is an abstract noun and it is an appositive of the relative clause*.

(NP (CP (IP (NP (NN 系统))
(VP (VV 停止)
(NP (NN 工作))))
(DEC 的))
(NP (NN 现象)))
(*the phenomenon that the system stops working*)

Figure 5. A relative clause in the NICT Chinese Treebank.

4.2 Quality Control

Several processes were undertaken to guarantee the quality of our re-annotation:

(1) We chose graduate students who major in Chinese for all of the annotators.

(2) A visualization tool - XConc Suite (Kim et al., 2008) was used as assistance during the re-annotation.

(3) Only 2,363 sentences with good translation quality in the NICT Chinese Treebank were chosen for re-annotation in the current phase.

(4) Before starting the re-annotation, a linguist selected 200 representative sentences, which contain all the linguistic phenomena that we want to re-annotate, from among the 2,363 sentences in the NICT Chinese Treebank. The selected 200 sentences were manually re-annotated by the linguist, and were split into two sets for training the annotators sequentially. We evaluated the annotation quality of the annotators during training. The average annotation quality of all the annotators after training is shown in Table 2.

Annotation Quality		Inter-annotator Consistency	
Precision	Recall	Precision	Recall
70.71%	70.75%	61.59%	61.59%

Table 2. The average annotation quality of the annotators after training.

(5) After training, the remaining sentences were split into several parts and assigned to the annotators for re-annotation. In each part, there were around 20% sentences that were shared by all of the annotators. These shared sentences were used to check and guarantee inter-annotator consistency during the re-annotation.

5 Conclusion and Future Work

We re-annotated the deep information, which includes eight types of grammatical functional tags and the traces in four constructions, to a Chinese scientific treebank, i.e. the NICT Chinese Treebank. Since the NICT Chinese Treebank is based on manually translated sentences, only 2,363 sentences with good translation quality were re-annotated in the current phase to guarantee the re-annotation quality.

In the future, we will finish the trace annotation for the BA and BEI constructions with split verbs. Furthermore, we will continue our re-annotation on more sentences in the NICT Chinese Treebank.

Acknowledgments

We would like to thank Dr. Kiyotaka Uchimoto and Dr. Junichi Kazama for providing the NICT Chinese Treebank.

References

Don Blaheta and Eugene Charniak. 2000. Assigning Function Tags to Parsed Text. *Proceedings of NAACL 2000*.

Ryan Gabbard, Seth Kulick and Mitchell Marcus. 2006. Fully Parsing the Penn Treebank. *Proceedings of HLT-NAACL 2006*.

Yuqing Guo. 2009. *Treebank-based acquisition of Chinese LFG Resources for Parsing and Generation*. Ph.D. Thesis. Dublin City University.

Julia Hockenmaier and Mark Steedman. 2002. Acquiring Compact Lexicalized Grammars from a Cleaner Treebank. *Proceedings of the 3rd LREC*.

Jindong Kim, Tomoko Ohta, and Junichi Tsujii. 2008. Corpus Annotation for Mining Biomedical Events from Literature. *BMC Bioinformatics*, 9(10).

Yusuke Miyao. 2006. From Linguistic Theory to Syntactic Analysis: Corpus-oriented Grammar Development and Feature Forest Model. Ph.D Thesis. The University of Tokyo.

Fei Xia. 1999. Extracting Tree Adjoining Grammars from Bracketed Corpora. *Proceedings of the 5th NLPRS*.

Fei Xia. 2000 (a). The Segmentation Guidelines for the Penn Chinese Treebank (3.0).

Fei Xia. 2000 (b). The Part-of-speech Tagging Guidelines for the Penn Chinese Treebank (3.0).

Nianwen Xue, Fudong Chiou, and Martha Palmer. 2002. Building a Large-Scale Annotated Chinese Corpus. *Proceedings of COLING 2002*.

Nianwen Xue and Fei Xia. 2000. The Bracketing Guidelines for the Penn Chinese Treebank.

Shiwen Yu et al. 2002. The Basic Processing of Contemporary Chinese Corpus at Peking University Specification. *Journal of Chinese Information Processing*, 16 (5).

Qiang Zhou. 2004. Annotation Scheme for Chinese Treebank. *Journal of Chinese Information Processing*, 18 (4).