Build a Large-Scale Syntactically Annotated Chinese Corpus

Qiang Zhou

State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science and Technology
Tsinghua University, Beijing 100084, P. R. China
zhouq@s1000e.cs.tsinghua.edu.cn

Abstract. This paper reports on our research to build a large-scale Tsinghua Chinese Treebank (TCT). We propose a two-stage approach to reduce manual proofreading labors as much as possible. The insertion of an intermediate functional chunk level creates a good information bridge to link simple chunk annotation with detailed syntactic tree annotation. We describe our chunk and tree annotation schemes, focus on two grammatical relation tag sets designed to give more detailed description for most of the special language phenomena in the Chinese language. We also briefly introduce our current progress in building a Chinese chunk bank with 2,000,000 Chinese characters, developing an efficient Chinese chunk-based parser and building a 1,000,000 words Chinese treebank. All this work lays good foundations for further research project to build a good Chinese parser.

1 Introduction

Corpus-based methods play an important role in empirical linguistics as well as in machine learning methods in natural language processing. The key issue in these two areas of research is to build large natural language corpora enriched with syntactic information. Therefore, in recent years, many researchers dedicated themselves to the construction of these syntactically annotated corpora, commonly called 'treebanks'.

For the English language, one of the best-known treebanks is the Penn Treebank (PTB1), which consists of about 1 million words of newspaper text annotated with rough syntactic and semantic tags in a bracketing format [8]. PTB2 further added some predicate-argument relation tags and trace-filler mechanisms to represent discontinuous phenomena [7]. The newest research work involves adding a layer of semantic annotation to the PTB2 and creating a Proposition Bank [6].

For languages other than English, a fairly well known treebank is the Prague Dependency Treebank (PDT) for Czech [3]. It contains about 450,000 tokens and is annotated on morphological, syntactic and tectogrammatical levels. Another large treebank is the TIGER treebank for German [2]. It extends the annotation scheme used in the NEGRA treebank [15] and currently contains 35,000 German newspaper

sentences annotated with part-of-speech information, phrase categories, syntactic functions, lemmata and morphology information.

For the Chinese language, there are two announced treebanks now. One is the Penn Chinese Treebank. Its first release version (CTB-1) contains about 100,000 Chinese words of Xinhua newswire texts and adopts the annotation scheme similar to English PTB2 [10]. Now, the 400,000-word CTB-2 is being developed and to be ready early in the year 2003 [11]. The other is the Taiwan Sinica Treebank [4]. It annotated syntactic categories and some thematic role information in Chinese sentences Its release version 1.0 contains about 240,000 Chinese words.

This paper reports on the Tsinghua Chinese Treebank (TCT) project, which aims at building the largest and most exhaustively annotated treebank for the Chinese language. In this project, we extended the single constituent tag set used in a small-size (about 200,000 Chinese words) test suite for Chinese treebank construction [13], and added grammatical relation tags to give more detailed syntactic description in treebank annotation. We also inserted an intermediate functional chunk annotation level to link grammatical relation tags with syntactic constituent tags and proposed a two-stage approach to improve the overall annotation efficiency.

The rest of the paper is organized as follows: Section 2 introduces the overview of TCT project, including its basis corpora, research goal and annotation method. Section 3 and 4 give more detailed description of our chunk and tree annotation schemes. Section 5 briefly introduces our current progress. Finally, section 6 summarizes the paper and sketches some ideas for future work.

2 Overview of TCT project

The basis of the TCT project is HYCorp, a corpus containing two million Chinese characters of text drawn from a balanced collection of journalistic, literary, academic, and other documents published in 1990s [9]. Only complete articles were used. All of the material has been hand corrected after sentence splitting process, word segmentation and part-of-speech tagging by automatic tools. Table 1 lists some basic statistics of HYCorp, where the Word Sum includes Chinese words and punctuations, the Char. Sum includes Chinese characters and punctuations.

Table 1. Basic statistics of HYCorp

Text Type	Article Sum	Sent. Sum	Word Sum	Char. Sum	Average Length (Word/Sent.)
Academic	29	9846	273017	447288	27.73
Journal	376	16921	427649	674566	25.27
Literary	295	38258	740445	1018839	20.56
Others	258	4302	88452	144027	19.35
Total	958	69327	1529563	2284720	22.06

2.1 Project Goal

The goal of the TCT project is to extract about 50,000 sentences (with about 1 million Chinese words) from HYCorp and to annotate them with correct syntactic trees. It is a five years project from 1998 to 2003. Figure 1 shows the complete parse tree of an annotation example, where word segmentation and part-of-speech information is encoded in terminal nodes, separated with '/'. All non-terminal nodes are labeled with syntactic constituent and grammatical relation tags, separated with '-'.

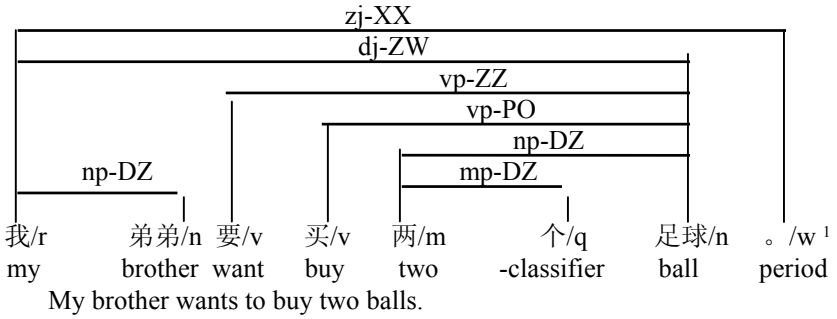

Figure. 1. Different levels of annotations in TCT

2.2 Annotation Method

In our opinion, the manual efforts are inevitable in the construction of a good syntactically annotated corpus. The key issue is how to reduce them as far as possible through suitable human-machine collaboration. As we know, the biggest problem of many current automatic parsers lies in their poor disambiguation ability. In these respects, humans have their advantages. If we can separate the complex sentence into several chunks with special syntactic functions through suitable manual preprocessing, then provide them to the automatic parser for syntactic parsing, many ambiguous structures in the sentence will be eliminated or restricted in the smaller context. Therefore, the accuracy of the parsed results will be greatly improved and the man workload for post-proofreading will be greatly reduced.

Starting from the above ideas, we inserted an internal annotating level: functional chunk and proposed a two-stage approach for Chinese treebank construction. Firstly, we manually partitioned sentences into several functional chunks to reduce the difficulty of automatic parsing. Then, we developed a chunk-based parser to provide high-accuracy parse outputs for manual proofreading. Because much useful disambiguation knowledge has been introduced by manual chunk annotation, the accuracy of the chunk-based parser can be improved to about 85%. Therefore, only less than 15% of

[1] The part-of-speech and syntactic tags used in this sentence are as follows: r—pronoun, n--noun, v--verb, m--numeral, q--classifier, w--punctuation; np--noun phrase, mp--numeral phrase , vp--verb phrase, dj--simple sentence, zj--complete sentence; DZ--attribute-head relation, PO--predicate-object relation, ZZ--adverbial-head relation, ZW--subject-predicate relation, XX--default relation.

the constituent structures need manual proofreading and modification. So the manual proofreading labor will be greatly reduced and the overall annotating efficiency will be greatly improved.

3 Functional chunk scheme

Our functional chunk scheme represents information about grammatical relations between sentence-level predicates and their arguments. Under this framework, each simple sentence (or clause) is exhaustively partitioned into a series of non- nested, non-overlapped labeled units, or functional chunks, while any structural relations within or between these chunks are left implicit.

In the typical case, each clause contains one predicate (P) chunk. Preceding the predicate there may be some number of adverbial (D) chunks, possibly with one subject (S) chunk among them. Following the predicate, there may be one direct object (O) chunk, one indirect object (O) chunk or one complement (C) chunk, possibly followed by a modal particle chunk (Y). For instance, for the Chinese sentence cited in Figure 1, the correct functional chunk annotated result after manual proofreading is as follows:

[S 我/r 弟弟/n] [D 要/v] [P 买/v] [O 两/m 个/q 足球/n] 。/w

The functional chunk annotation builds basic links between functional structure [5] and argument structure [1]. Although there is no explicit annotation of connections between specific predicates and specific arguments, that information should be largely recoverable from the sequence of chunk categories.

In fact, there are several other annotation schemes to describe argument structure. The ongoing PropBank project tries to provide a layer of consistent argument labels to the Penn treebank [5]. It is the research work to describe the argument structure through predicate-argument annotation itself.

Another interested project is FrameNet developed in UC, Berkeley [16]. Based on frame semantics proposed by Fillmore [17], they selected and labeled those constituents in the sentences which instantiated the concepts of frame elements (FEs). The constituents identified as FEs were then to be classified (automatically if possible) as to their phrase type (PP, etc.) and in respect to their grammatical function (Object, etc.). Therefore, a lexicon with full descriptions of the frame-semantic and syntactic combinational properties of the words can be constructed automatically from the set of annotations [18]. It is the research work to describe the argument structure through semantic role annotation.

4 Tree annotation scheme

Unlike the flattened structure trees used in PTB and TIGER projects, we adopted a deeper and more complete parse trees to annotate Chinese sentences. In our annotation scheme, each non-terminal constituent in a sentence will be assigned with two tags. One is the syntactic constituent tag, which describes its external functional rela-

tion with other constituents in a parse tree. The other is the grammatical relation tag, which describes the internal structural relation of its sub-components. These two tag sets form an integrated annotation for the syntactic constituent in a parse tree through top-down and bottom-up descriptions.

Our syntactic constituent tag set consists of 15 tags, focusing on the description of different levels of syntactic constituents in a parse tree, including:

(1) phrases, such as noun phrases, verb phrases, preposition phrases, etc.;
(2) sentences, such as simple sentences and complex sentences;
(3) other special language phenomena, such as quotation, independent constituent (e.g. parenthesis, exclamation, etc.).

Our grammatical relation tag set consists of 26 tags, focusing on the description of following information:

(1) The structural relation of sub-components in phrases and simple sentences
- Governor relations, such as subject-predicate, predicate-object, predicate-complement;
- Modification relations, such as attribute-head, adverbial-head;
- Coordination relations, such as conjunction structure, co-predicate structure;
- Syntactic relations, such as addition, overlapping, location, etc.

(2) The internal logical relation of clauses in complex sentences
- Coordination relations, such as coordinate clauses, coherent clauses, promotional clauses, selective clauses;
- Causality relations, such as cause-result, purpose-action;
- Conditional relations, such as conditional clauses, presumptive clauses, transitional clauses;
- Explicatory relations;
- Topic-comment relations.

These two relational tag sets form a good bridge linking our tree annotation scheme with functional chunk scheme. Compared with the single functional tag attachment method used in PTB and TIGER projects, our binary or multiple (only for conjunction structures) relation tag shows more flexible description capability. Not only the grammatical relations of different functional chunks at sentence level can be easily represented by them; some detailed syntactic relations, such as addition, overlapping, and so on, can be also described easily. This characteristic is very suitable for the Chinese language. There are not very obvious boundaries between Chinese words and phrases. Therefore, some special syntactic relational tags may need to describe those intersected constituents.

According to our tentative statistics in HYCorp, more than 50% of the sentences[2] are complex sentences with two or more clauses. They comprise abundant information content, especially in the case integrating with some special language phenomena, such as quotations and independent constituents. Therefore, some detailed tags should be specially designed to describe the complex logical relation between different clauses inside them. Our current relation tag set contains 11 tags, covers most commonly-used event logical relations, including coordination relations, causality

[2] Here the sentence is defined as the word sequences ended with free period, interrogation or exclamatory mark.

relations and conditional relations, and some special event relations used in the Chinese language, such as explicatory relations and topic-comment relations. So far as we know, it is the most detailed event relation tag set used in current treebank projects.

5 From chunk bank to treebank

Our current chunk bank was built through manual annotation and proofreading. Our tentative count shows that original annotating speed for an annotator is about 1200 words per work hour. As they are familiar with the annotation scheme and processing procedure deeper and deeper, the annotation speed will gradually increase and reach about 2400 words per work hour after 1 or 2 months.

We designed a two-level checking system to guarantee the quality of the final annotating results. Firstly, we developed an automatic checking program based on the basic principles and rules listed in chunk scheme. Most wrong chunks can be checked out and provided to annotator for further confirmation or modification. Secondly, we checked the final annotating results through random sampling, found and modified the chunk errors left, until the error ratio is below 1%.

The chunk bank project began in March 2000 and ended in June 2001. All the YHCorp were processed and a Chinese chunk bank with two million Chinese characters has been build. Detailed information can be found in [14].

Based on the functional chunk information, our chunk-based parser, a revised version of the former statistics-based Chinese parser [12], can only focus on the following parsing tasks: (1) the intra-chunk phrase parsing; (2) the inter-chunk clause parsing; and (3) The clause relation analyzing.

A comparison experiment shows that functional chunk information brings in great improvement in parsing performance: the labeled precision an recall of the syntactic parser increase about 13%, from 76% to 89%, and the average number of crossing brackets in a sentence is reduced from 3.04 to 1.17. This is a close test on 7595 very long Chinese sentences (about 200,000 Chinese words).

Therefore, our current treebank can be built through manual proofreading on the syntactically annotated sentences output by the chunk-based parser. The detailed proofreading procedure is as follows: Firstly, each annotator is assigned several annotated sentences for first-level proofreading. After that, the error-checking tool is used to find and modify most obvious errors so as to obtain a better annotation version. Then, about 30% of the complex sentences are randomly selected from the annotated sentences after first-level proofreading and sent to another annotator for second-level proofreading. Some remained or overlooked errors can be found and corrected. Thus, the best annotation version can be obtained through this two-level proofreading approach.

The treebank project began in July 2001. About 50,000 sentences annotated with correct functional chunks, consisting of 38% literary texts, 34% news texts, 20% academic texts and 8% other texts, were extracted from the chunk bank and analyzed through chunk-based parser. By the end of March 2003, all the first-level proofreading work and about 35% second-level proofreading work have been finished. This

work still goes on and we plan to complete the first release version in July 2003.

6 Summary and future work

This paper reports on our research to build the largest syntactically annotated Chinese corpus: Tsinghua Chinese Treebank (TCT). We mainly focused on the following two issues: (1) How to design a suitable annotation method to reduce manual proofreading labor as much as possible; (2) How to design a suitable annotation scheme to describe syntactic phenomena in as detailed manner as possible in the Chinese language.

For the first issue, we proposed a two-stage approach. The insertion of an intermediate functional chunk level provided us with good opportunity to simplify the automatic parsing and created an information bridge to link simple chunk annotation with detailed syntactic tree annotation. For the second issue, we extended the single constituent tag set used in our small-size treebank test suite, and specially designed two grammatical relation tag sets to describe more detailed phrase structural relations and clause logical relations in the Chinese sentences.

Furthermore, we explained our chunk and tree annotation schemes and also briefly introduced the current progress in building a Chinese chunk bank with 2,000,000 Chinese characters, developing an efficient Chinese chunk-based parser and building a 1,000,000 words Chinese treebank.

Future work will be concerned with the application of current treebank in Chinese automatic parsing and understanding systems. For instance, we can extract useful verb sub-categorization templates from TCT and apply them in current parser to improve parsing performance. Based on the lexical collocation information extracted from TCT, we can explore efficient computational model for semantic similarity and cohesion and find useful heuristic rules for automatic mapping from grammatical relations to semantic theta role relations. All these research work will give impetus to the development of a better Chinese parser and understanding system.

Acknowledgements

This work was supported by the Chinese National Science Foundation (Grant No. 69903007, 60173008), National 973 Foundation (Grant No. 1998030507) and National 863 plan (Grant No. 2001AA114040). The author would like to thank all the corpus annotators for their hard work, and thank three anonymous reviewers for their helpful comments and advices.

References

1. Alsina. A. (1996). *The Role of Argument Structure in Grammar: Evidence from Romance*. CSLI Lecture Notes No. 62, CSLI Publications: Stanford, California, USA.

2. Brants, S., & Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus. In *Proc. of the Third Conference on Language Resources and Evaluation LREC-02,* Las Palmas, Spain.
3. Hajic, J. (1999). Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajicova (Ed.), *Issues of valency and meaning. Studies in honour of Jarmila Panevova.* Prague, Czech Republic: Charles University Press.
4. Huang Chu-Ren, Chen Feng-Yi, & al.(2000). Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface, In *Proc. of the Second Chinese Language Processing Workshop, HongKong*. 29-37.
5. Kaplan, R. and Bresnan, J. (1982). Lexical-Functional Grammar: A Formal System of Representation, In Joan Bresnan (ed.) *The Mental Representation of Grammatical Relations.* MIT Press, Cambridge, Mass. 173-281.
6. Kingsbury, P.; Martha Palmer, and Marcus, M. (2002). Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference, San Diego, California*.
7. Marcus, M., Kim, G., Marcinkiewicz, M.,& al. (1994). The Penn Treebank: Annotating predicate argument structure. In *Proc. of the ARPA Human Language Technology Workshop.* San Francisco, CA
8. Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics,* 19(2), 313-330
9. Sun M.S., Zhou Q., & al. (2000). Constructing a Word-segmented & POS-tagged Chinese Corpus and a Chinese Treebank, In *Proc. of International conference on Chinese language computing (ICCLC'00)*, p239-243
10. Xia, F., Martha Palmer, & al. (2000) Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In *Proc. of the second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece.*
11. Xue N.W., Chiou F. and Martha P. (2002). "Building a Large-Scale Annotated Chinese Corpus". In *Proc. of 19th International Conference on Computational Linguistics (COLING-02*), Taiwan.
12. Zhou, Q. (1997). A Statistics-Based Chinese Parser. In *Proc. of the Fifth Workshop on Very Large Corpora,* 4-15. Bejing, China.
13. Zhou, Q. and Sun, M.S. (1999). Build a Chinese Treebank as the test suite for Chinese parser. In *Proc. of the Workshop MAL'99 (Multi-lingual information Processing and Asian Language Processing)*, Beijing, China.
14. Zhou Q., Zhang W. D., Ren H. B. (2001). Build a large scale Chinese functional chunk bank. In *Changning Huang, Po Zhang(eds.) Natural language understanding and machine translation. Tsinghua University Press.* 102-107. (In Chinese)
15. Skut,W., Brants, T., Krenn, B., & Uszkoreit, H. (1998). A linguistically interpreted corpus of German newspaper text. In *Proceedings of the Conference on Language Resources and Evaluation LREC-98* (pp. 705–711). Granade, Spain.
16. Baker, C.F., Fillmore, C.J., and Lowe, J.B. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL'98, Montreal, Canada*, 86-90.
17. Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm*, Hanshin Publishing Co., Seoul, South Korea. 111-137.
18. Fillmore, C.J., Wooters, C., and Baker, C.F. (2001). Building a Large Lexical Databank Which Provides Deep Semantics. In *Proc. of the Pacific Asian Conference on Language, Information and Computation. Hong Kong.*