International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.11 (2015), pp.1-6
http://dx.doi.org/10.14257/ijsip.2015.8.11.01

Syntactic Parsing Tree in Tibetan Language based on Context Free Grammars

Lirong Qiu

School of Information Engineering, Minzu University of China
Beijing, China
qiu_lirong@126.com

Abstract

Serving as a basic and key research of Tibetan information processing, Tibetan syntax analysis plays an important role in promoting the research on Tibetan natural processing technology such as machine translation, Tibetan information retrieval and semantic analysis. The existing Tibetan syntactic analysis system shows poor performance in general fields. Based on unique characteristics of Tibetan, through analysis and study on Tibetan syntax, the efficiency of sentence processing system can be improved, thus speeding up the progress of Tibetan sentence processing. In this paper, we proposed a syntactic parsing method based on context free grammars using up-bottom parsing technology.

Keywords: *syntactic parsing tree, context free grammar, up-bottom parsing, CKF algorithm*

1. Introduction

Information processing technology for English and Chinese, which have rich language resources, are carried out on the basis of large scale language resources statistics [5]. Although Tibetan is a widely spoken language among the minority languages, compared to Chinese and English, available Tibetan electronic resources are relatively scarce, especially annotated text and corpus. This brings adverse effect on Tibetan information processing.

Syntactic analysis is a basic technology of natural language processing, and also an important stage for natural language processing. It plays an important role in natural language processing fields, which includes machine translation, information extraction and many other applications. Therefore syntactic analysis is an indispensable part in natural language processing research of any languages.

Syntactic analysis is grammatical function analysis for the vocabulary in sentence, which aims at deriving syntactic structure law of Tibetan sentence automatically according to the given grammar rules, and identifying the syntactic unit and their relationship according to the syntactic structure law obtained above. The main contents include recognizing the words in sentence; stipulating syntactic role for each word; figuring out the containing word group or phrases (adjective phrases, noun phrases, adverb phrases, verb phrases, *etc*); and the way the syntactic constituent above combine into the whole sentence structure.

The result of parsing is a syntactic tree that clearly measures the relationship type and primary and secondary relations, and the syntax tree expresses the sentence structure in the form of dendritical structure.

Serving as a basic and key research of Tibetan information processing, Tibetan syntax analysis plays a tremendous role in promoting the research on Tibetan natural processing

ISSN: 2005-4254 IJSIP

technology such as machine translation, Tibetan information retrieval and semantic analysis.

The unique linguistic characteristics of Tibetan language makes the research achievement, which applied in English and Chinese role labeling, cannot directly copy to Tibetan. For Tibetan, on the one hand, the grammatical markers in Tibetan language can be directly used to mark the semantic role information of subject, object, time, location and pattern; on the other hand, according to the result of semantic role marking on predicate, through the retroaction on parsing process, the only uncertain effect of syntactic markers can be reduced, thus improve the performance of sentence processing system.

The existing Tibetan syntactic analysis system shows poor performance in general fields. Based on unique characteristics of Tibetan, through analysis and study on Tibetan syntax, the efficiency of sentence processing system can be improved, thus speeding up the progress of Tibetan sentence processing. In this paper, we proposed a syntactic parsing method based on context free grammars using up-bottom parsing technology.

The remainder of the paper is organized as follows. In Section 2, we provide an overview of preliminaries of syntactic parsing. Section 3 presents our work on syntactic parsing tree analysis with context free grammar approach. The related work are introduced in Section 4, followed by the conclusions, discussions, and future work in Section 5.

2. Preliminaries

The main methods of syntactic analysis research are rule-based method and statistic-based method.

Rule-based method is based on the conclusion and summary of language phenomenon from linguists, then compile the descriptions of internal rule for language phenomenon. Rule based method has plenty of advantages: maximize the approach to the syntax habit of natural language, can be quickly grasped by linguists; have flexible and diverse forms of expression, maximize the expression of thoughts from researchers, effectively express natural language, and have good computability.

Linguists believe that natural language has character of constitutive property and hierarchy, and the layered structure of language can be described formalized rules, these rule sets are syntax of natural language [6]. For a sentence (essentially an arbitrary string) the corresponding syntactic structure can be derived according to the grammar rules, and whether the string is a reasonable sentence, that is to say, whether the string conforms to grammar rules, can be determined from the syntactic structure.

Statistic-based method indicates grammatical rules of language phenomena with the method of probability, this method could not determine whether a sentence conforms to grammar rules, instead, it determines the probability of a certain sentence constituent through the statistical result of a language phenomenon in large corpora.

Both rule-based and statistic-based method have their advantages, but the same as other natural language processing mission, there are defects in application no matter which method was used. For example: the rule-based method could not perfectly describe all the language phenomena rules; in addition, formulation of rules is also inevitably subjective; statistic-based method usually could not 100% correctly represent the syntactic structure, there is a certain degree of error, and error accumulation will influence on the subsequent language processing.

According to the requirements of formalizing Tibetan natural language, Zasiga and Dora [15] put forward exploratory research idea, demonstrates the necessity of describe Tibetan sentences with complex characteristics. Starting with formalizing Tibetan natural language, they put forward the research idea of uniting the calculation of vocabulary, syntax, semantic rules and sentences.

3. Up-Bottom Parsing Method based on Context Free Grammar

Most of the center words in Tibetan in the position of sentence tail, while subject, object and complement are before the center word, which makes the interior relationship of sentence elements very complex and difficult to distinguish, thus becoming the problems that must be solved in Tibetan dependency grammar analysis.

The unique language characteristics of Tibetan must be considered while carrying out sentence processing research on Tibetan. Gesangjumian's Practical Tibetan Grammar, Zhou Jiwen's Grammar of Tibetan Lhasa Dialect, and paper of Jiang Di [11] have pointed out repeatedly that there are three main characteristics as follows that occupied the core status of Tibetan sentences:

(1) Word order: the word order of Tibetan sentences is "subject + object + verb" (SOV), this is a kind of predicate postposition language.

(2) Grammatical markers: words or phrases serves as subject, object, purpose and locations in Tibetan sentences, usually include grammatical markers.

(3) Predicate verb: when Tibetan verb serves as predicate of the sentence, except the rich syntactic category information (person, tense, *etc.*), the semantic categories of verbs (verb attributes, possession, intention, *etc.*) are also included.

The syntactic structure of Tibetan varies a lot with Chinese and English, such as: the Tibetan structure is subject-object-predicate, while the Chinese, English and other language are subject-predicate-object structure; Chinese is a language with no morphologic change, while Tibetan has rich vocabulary morphology, that is to say, Tibetan has obvious morphological cases, usually appear in the middle of sentence.

So the construction of Tibetan dictionary and rule base cannot consult the dictionary and rule base of Chinese or English, it needs to establish the rule base and dictionary according to the characteristics of Tibetan itself.

Start search space of reverse derivations from the terminal symbols in the string: given a set of example trees, the underlying CFG can simply be all rules seen in the corpus. According to the literature [14], the following grammar form is adopted:

$G = <Vn, Vt, S, P>$

$Vn = <S, NP, VP, Det, N, V, Prep>$

$S = S$

<1> $S \rightarrow NP+VP$

<2> $NP \rightarrow N+Det$

<3> $VP \rightarrow DP+VP$

<4> $DP \rightarrow D+Det$

<5> $VP \rightarrow N+V$

<6> $N \rightarrow$ [རྒྱ་བཟང་]

<7> $Det \rightarrow$ [གིས་]

<8> $D \rightarrow$ [འདང་ངེ]

<9> $Det \rightarrow$ [ས]

<10> $N \rightarrow$ [རྒྱུ་]

<11> $V \rightarrow$ [བཟོས]

（1）Rule VP→· V NP: represents this rule has not been matched yet;

（2）Rule VP→V · NP: represents the V in the left side of this rule has been successfully matched, yet the NP have not;

（3）Rule VP→V NP: represents this rule has been matched and formulates a phrase

VP.

Similar to CYK algorithm, Earley algorithm is a kind of parallel algorithm, which needs no backtrack [8]. Earley algorithm makes use of a two-dimensional matrix to store the analyzed results; another important contribution of Earley algorithm is the introduction of dot rule, which further reduce the redundant operation in rule matching.

Data structure: for a two dimensional matrix {E (i,j)}, each element is a collection of dot rules, and stores all the dot rules obtained from the span of word i to the word j in sentences after analysis.

Initialization:

（1）For the centralized rule S→ α ,whose left side is initial character S, add S→· α into E(0,0).

（2）If E(0,0) contains B→· A β, add S→· α into E(0,0) for all A→α rule whose left side is initial character A.

Circulate the following steps until success or failure appears:

（1）If E(i,j-1) contains A→α·xjβ, add A→αxj·β into E(i,j).

（2）If E(i,j) contains A→α·Bβ, add B→·γ into E(j,j) for all the rule B→γ whose left side is character B.

（3）If E(i,j) contains B→γ, meanwhile E(k,i-1) contains A→α·Bβ, add A→αB·β into E(k,j).

“ས་སྒེ་རེ་ཡིས་ཤིང་བཅད།” (Lausanne felled the tree with an axe) this sentence starts with the initial symbol S, generates the syntax analysis tree through top-down scanning until the end of analysis on terminator. S is in the first search target to scan. Start from the initial character S, select the corresponding rules in grammar to replace the search target, and match the words in sentence with the right part of the grammar rule. If the matching is successful, then erase the word, and record the corresponding rules of this word in search target, then search the remaining part of the input sentence; the analysis completes when analyzing to the terminator and the search target is 0.

Next to The arrow pointing to right “→”, the rule number is labeled, the search target with drawing line represents using the left part of the rule. According to this search process, sentence “ས་སྒེ་རེ་ཡིས་ཤིང་བཅད།” (Lausanne felled the tree with an axe) is processed as follows:

(1) S→rule <1>($S \rightarrow NP+VP$): སྐོ་བཟང་གིས་སྒེ་རེ་ཡིས་ཤིང་བཅད། (Status)

(2) NP+VP→rule <2>($NP \rightarrow N+Det$): སྐོ་བཟང་གིས་སྒེ་རེ་ཡིས་ཤིང་བཅད། (Status)

(3) N+Det+VP→rule <6>($N \rightarrow$སྐོ་བཟང་): སྐོ་བཟང་གིས་སྒེ་རེ་ཡིས་ཤིང་བཅད། (Status)

(4) Det+VP→rule <7>($Det \rightarrow$གིས་): གིས་སྒེ་རེ་ཡིས་ཤིང་བཅད། (Status)

(5) VP→rule <3>($VP \rightarrow DP+VP$): སྒེ་རེ་ཡིས་ཤིང་བཅད། (Status)

(6) DP+VP→rule <4>($DP \rightarrow D+Det$): སྒེ་རེ་ཡིས་ཤིང་བཅད (Status)

(7) D+Det+VP→rule <8>($D \rightarrow$སྒེ་རེ་): སྒེ་རེ་ཡིས་ཤིང་བཅད། (Status)

(8) Det+VP→rule <9>($Det \rightarrow$ཡིས་): ཡིས་ཤིང་བཅད། (Status)

(9) VP→rule <5>($VP \rightarrow N+V$): ཤིང་བཅད། (Status)

(10) N+V→rule <10>($N \rightarrow$ཤིང་): ཤིང་བཅད། (Status)

(11) V→rule <11>(V →བཅད): བཅད། (Status)

OverThe sentence “ས་སྒེ་རེ་ཡིས་ཤིང་བཅད།” (Lausanne felled the tree with an axe) can be translated into a syntactic parsing tree, as shown in Figure 1.

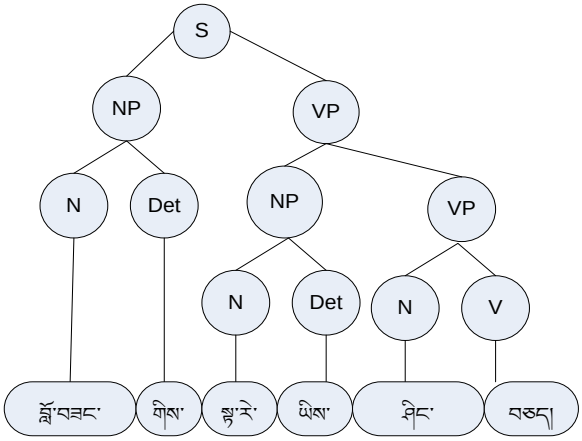


Figure 1. Syntactic Parsing Tree

4. Related Work

A number of extracting studies investigate the syntactic parsing using two sources of information: lexicographic resources and distributional similarity [1-4]. However there are not much study about the syntactic parsing methods.

According to the requirements of formalizing Tibetan natural language, Zha Xijia [14] describes a syntactic parsing method with a context free method and Zasiga and Dora [15] put forward exploratory research idea, demonstrates the necessity of describe Tibetan sentences with complex characteristics. Starting with formalizing Tibetan natural language, they put forward the research idea of uniting the calculation of vocabulary, syntax, semantic rules and sentences.

Based on unique characteristics of Tibetan, through analysis and study on Tibetan syntax, the efficiency of sentence processing system can be improved, thus speeding up the progress of Tibetan sentence processing. In this paper, we proposed a syntactic parsing method based on context free grammars using up-bottom parsing technology.

5. Conclusion and Future Work

The study on Tibetan syntactic analysis has outstanding significance. But at present, it is difficult to obtain deep syntactic analysis results of Tibetan. The existing Tibetan syntactic analysis system shows poor performance in general fields [7]. Based on unique characteristics of Tibetan, through analysis and study on Tibetan syntax, the efficiency of sentence processing system can be improved, thus speeding up the progress of Tibetan sentence processing.

This paper addressed the problem of syntactic parsing tree in Tibetan language, and proposed an analysis approach.

The work of this paper is a part of our ongoing research work, which aims to provide a Tibetan sentence parsing tree corpus for further bilingual named entity recognition, translation and other applications of Tibetan language [9].

Various experiments and applications have been conducting in our current research. Future work includes how to acquire and verify bilingual named entities from Tibetan and Chinese free text, how to obtain entity patterns automatically and how to acquire language features for entity recognition.

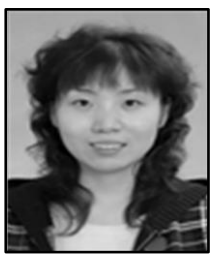
Acknowledgments

Our work is supported by the National nature science foundation of China (No. 61103161) and the Program for New Century Excellent Talents in University (NCET-12-0579).

References

[1] Z. Lin, M.-Y. Kan and H. Tou Ng, "Recognizing Implicit Discourse Relations in the Penn Discourse Treebank", Proceedings of the 2009 Conference on Empirical Method in Natural Language Processing, vol. 1, **(2009)** EMNLP, pages 343-351.
[2] R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, L. Robaldo and B. Webber, "The Penn Discourse Treebank 2.0 Annotation Manual", The PDTB Research Group, **(2007)** December.
[3] K. Moritz Hermann and P. Blunsom, "Multilingual Models for Compositional Distributed Semantics", ACL, **(2014)**, pp. 58-68.
[4] R. Subba and B. Di Eugenio, "An effective discourse parser that uses rich linguistic information", In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, **(2009)**, pp. 566-574.
[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural language processing (almost) from scratch", Journal of Machine Learning Research, **(2011)**, pp. 2493-2537.
[6] P. Hackett, "A Tibetan Verb Lexicon: Verbs, Classes, and Syntactic Frames", Ithaca: Snow Lion Publications, **(2003)**.
[7] L. Qiu, "Verb classification using bilingual lexicon and translation information in Tibetan language", International Journal of Multimedia and Ubiquitous Engineering, vol. 9, no. 2, **(2014)**, pp. 45-62.
[8] S. Petrov, L. Barrett, R. Thibaux and D. Klein, "Learning accurate, compact, and interpretable tree annotation", In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,Sydney, Australia, **(2006)**, pp. 433-440.
[9] L. Qiu, "Finding and typing new named entities in Tibetan from Chinese-Tibetan parallel corpora", International Journal of Multimedia and Ubiquitous Engineering, vol. 9, no. 9, **(2014)**, pp. 143-150.
[10] D. Albright, A. Lanfranchi, A. Fredriksen, W. F. Styler, C. Warner, J. D. Hwang, J. D. Choi, D. Dligach, R. D. Nielsen, J. Martin, W. Ward, M. Palmer and G. K. Savova, "Towards comprehensive syntactic and semantic annotations of the clinical narrative", Journal of the American Medical Informatics Association, vol. 5, no. 20, **(2013)**, pp. 922-930.
[11] D. Jiang, "The Classification of Tibetan Verbs and Relative Patterns Based on Semantics and Syntax", Journal of Chinese Information Processing, vol. 1, no. 25, **(2006)**, pp. 37-43.
[12] G. Durrett, A. Pauls and D. Klein, "Syntactic Transfer Using a Bilingual Lexion", In proceedings of the EMNLP-CoNLL, Jeju Island, Korea, **(2012)** July 12-14.
[13] R. Barziley and M. Lapata "Modeling local coherence: An entity-based approach", Computational Linguistics, vol. 1, no. 34, **(2008)**, pp. 1-34.
[14] Z. Jia, "Syntactic parsing based on context free grammar", Xizang University Natural science Journal, vol. 28, no. 2, **(2013)**.
[15] Z. and D., "Tibetan Syntax Formalization based on FUG", Journal of Chinese information processing, vol. 28, no. 3, **(2014)**, pp. 99-103.

Author



Lirong Qiu, she received her Ph.D. in Computer Sciences (2007) from Chinese Academy of Science. Now she is an associate professor of computer sciences at Information Engineering Department, Minzu University of China. Her current research interests include different aspects of natural language processing, artificial intelligence and distributed systems.

