Tibetan Syntactic Parsing based on Syllables

Fucheng Wan [1,a*], Xiangzhen He[1,b]

1（Key lab of China’s National Languages Information Technology, Northwest University for Nationalities,730030, Lanzhou Gansu, ）

awanfucheng@126.com.b306261663@qq.com,

Keywords: Tibetan syntactic parsing; Tibetan syllables; syntactic Treebank

Abstract: We propose a model of Tibetan syntactic parsing which is based on Tibetan syllables instead of Tibetan words, change Tibetan syntactic Treebank use algorithm of labeling.

Introduction

Tibetan phrase syntactic Treebank is on the basis of Tibetan words, Tibetan syntactic parsing is similar to Chinese syntactic parsing which is based on words, but Tibetan syllables are more useful in NLP applications. In recent years, a lot of paper are about Chinese word segmentation, Chinese syntactic parsing which is based on Chinese character. Luo propose a joint parsing method of Chinese word segmentation, part of speech and syntactic parsing. From then, more and more researcher pay more attention on the model of joint method. Li focuses on the structure of words and put forward a new method of joint parsing on syntax and word structure. Qian and Liu train three model during the training process, and join the three model together in the distinguishing framework. Zhang et al propose a method of joint morphological and syntactic parsing through establishing joint model of the feature of Chinese character on the basis of syntactic parsing which is based on shift-reduction algorithm, nowadays, these methods have become main directions.

The structure of this paper is like this: first, the algorithm of joint parsing labeling; then, the changing program of Tibetan Treebank structure; At last, a system of validation.

Algorithm of joint parsing labeling

The leaf nodes of Tibetan phrase Treebank are all kinds of Tibetan words, the labeling program is on the basis of Tibetan words, the level of words, Tibetan phrase syntactic parsing based on Tibetan syllables is on the basis of Tibetan syllables, the level of characters, so we need change the structure of Tibetan phrase syntactic Treebank based on words to that based on syllables, the detail of labeling and changing method is like Fig 1,

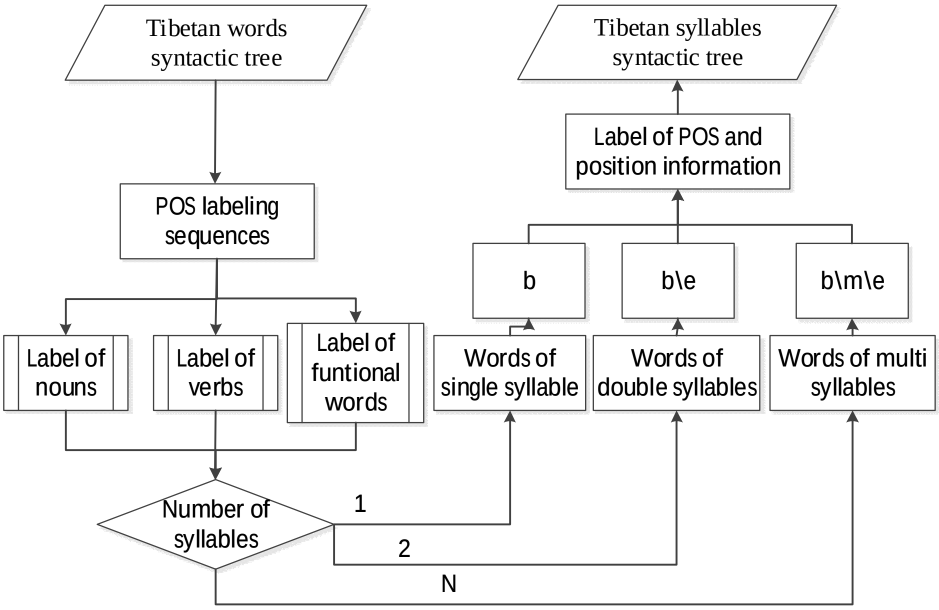

Fig 1 Algorithm of joint parsing labeling

(1) Output the sequences of Tibetan phrase syntactic Treebank part of speech labeling based on Tibetan words;
(2) Word-Level part of speech become the label of syllables sequences;
(3) Level of Tibetan syllables use the label of primary part of speech and add the extra position information;
(4) Judge the number of Tibetan syllables, single syllables, double syllables or multi-syllables;
(5) Position information: for words of single syllables, the label of position information is “s’; for words of double and multi syllables, the label of position is like this, the first syllable label of position information is “b”, the last syllable label of positon information is “e”, and the middle syllables label of position information is “m”;
(6) Output the part of speech and syllables sequences label;
(7) Output the Tibetan phrase syntactic Treebank based on Tibetan syllables.

During the process, for a Tibetan sentence, integrate the process of word segmentation, part of speech labeling, Named Entity Recognition and syntactic parsing. So, out method is joint parsing method.

Program of Tibetan phrase syntactic Treebank

The leaf nodes of Tibetan phrase syntactic Treebank are Tibetan words sequences, to establish the model of Tibetan phrase syntactic parsing based on Tibetan syllables, we need change Tibetan phrase syntactic tree based on Tibetan words to that based on Tibetan syllables, Tibetan syllables become the leaf nodes of Tibetan phrase syntactic tree, the program is as below:

(1) Label of POS in Tibetan phrase syntactic tree based on Tibetan words become phrase label in Tibetan phrase syntactic tree based on Tibetan syllables;

(2) Label of POS in Tibetan phrase syntactic tree based on Tibetan syllables use the label of Tibetan phrase syntactic tree based on Tibetan words, and add extra position information;

(3) For Tibetan words of single syllable, label of position information is “s”, for Tibetan words of double syllables, label of position information is “b”, ”s”, for Tibetan words of multi syllables, label of the first syllable is “b”, label of the last syllable is “e”, and label of the other syllables is “m”, that is the program of “bmes”.

Take Tibetan phrase syntactic tree for example, change Tibetan phrase syntactic tree based on Tibetan words to that based on Tibetan syllables, (IP(NP(NS(NSb པེ་)(NSe ཅིན་)))(VP(UN(UNs ནི་))(NP(DNP(NS(NSb ཀྲུང་)(NSe གོ))(DEG(DEGs འི་)))(NP(NN(NNb རྒྱལ་)(NNe ས་))))(VV(VVs ཡིན)))(PU །))， and the Chinese translation is "Beijing shi zhongguo de shoudu". Structure of Tibetan phrase syntactic tree based on Tibetan syllables is as shown in Fig 2

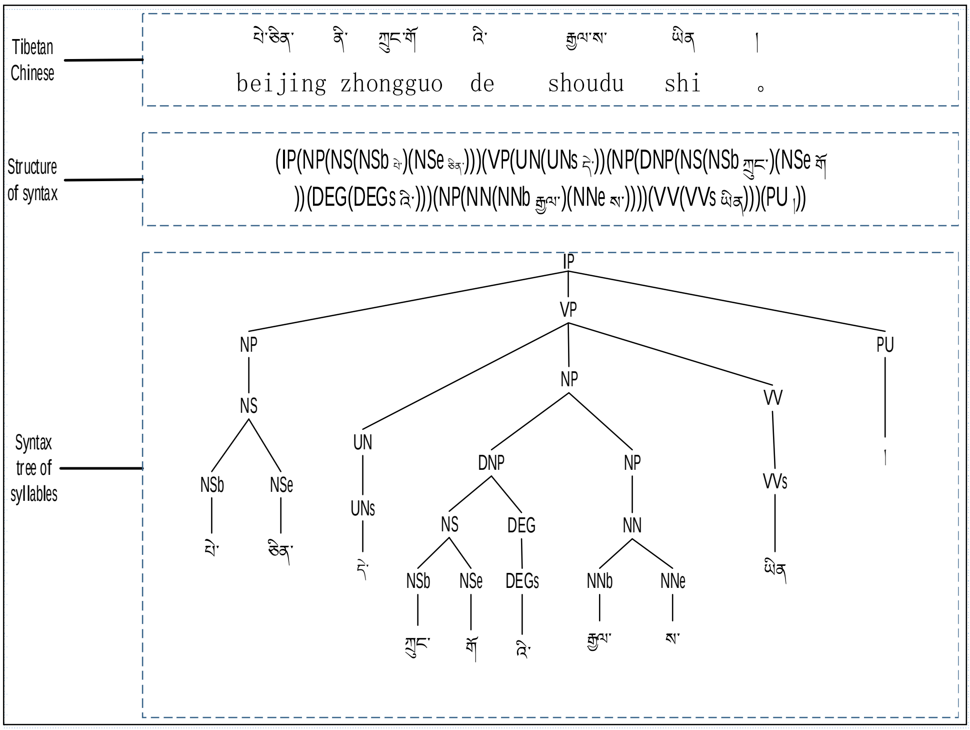
Tibetan
Chinese
པེ་ཅིན་ ནི་ ཀྲུང་གོ འི་ རྒྱལ་ས་ ཡིན །
beijing zhongguo de shoudu shi 。
Structure of syntax
(IP(NP(NS(NSb པེ་)(NSe ཅིན་)))(VP(UN(UNs ནི་))(NP(DNP(NS(NSb ཀྲུང་)(NSe གོ))(DEG(DEGs འི་)))(NP(NN(NNb རྒྱལ་)(NNe ས་))))(VV(VVs ཡིན)))(PU །))
Syntax tree of syllables
IP
VP
NP
PU
NS
NP
VV
UN
NSb
NSe
DNP
NP
UNs
VVs
NS
DEG
NN
པེ་
ཅིན་
ནི་
NSb
NSe
DEGs
NNb
NNe
ཡིན
ཀྲུང་
གོ
འི་
རྒྱལ་
ས་
།

Fig 2 Syntax tree of syllables

In Fig 2, Tibetan phrase syntactic tree based on Tibetan words, the label of POS nodes of "NR" has become the label of phrase "པེ་"、"ཅིན་", these are label of phrase structure, below the nodes of "NR", "པེ་ཅིན་"（beijing）has been divided into two syllables according to the sign of syllable, every syllable has its own label, the node above Tibetan syllables has become the label of "NR", the label of POS, add the extra position information, "b", "e", like this "NRb", "NRe", so the label of "པེ་"is "NRb", the label of "ཅིན་"is "NRe", label of single syllable is "s', like this, the label of "འི་" is "DEGs".

Acknowledgements

This research is supported by The Northwest University for Nationalities fundamental research funds for the Central Universities project supported by the special funds and the project of introduction of talents in Northwest University.

References：

[1] X. Luo. “A maximum entropy Chinese character-based parser”. In: Proceedings of the 2003 conference on empirical methods in natural language processing， 2003: 192–199.

[2] X. Qian and Y. Liu. “Joint Chinese word segmentation， POS tagging and parsing”. In:Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning， 2012: 501–511.

[3] Z. Li. “Parsing the enternal structure of words: A new paradigm for Chinese word segmentation.” In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies， 2011: 1405–1414.

[4] P. Fung et al. “A maximum-entropy Chinese parser augmented by transformation-based learning”. ACM transactions on Asian language information processing (TALIP)，2004， 3(2): 159–168.

[5] M. Zhang et al. “Chinese parsing exploiting characters”. 51st annual meeting of the Association for Computational Linguistics， 2013.

[6] Y. Zhang and S. Clark. “Syntactic processing using the generalized perceptron and beam search”. Computational linguistics， 2011， 37(1): 105–151.

[7] M. Collins. “Ranking algorithms for named-entity extraction: Boosting and the voted perceptron”. In: Proceedings of the 40th annual meeting on Association for Computational Linguistics， 2002: 489–496.

[8] Li Z. Parsing the Internal Structure of Words: A New Paradigm for Chinese Word Segmentation[C]，ACL. 2011: 1405-1414.