International Journal of Database Theory and Application
Vol.8, No.2 (2015), pp. 25-32
http://dx.doi.org/10.14257/ijdta.2015.8.2.03

A Method for Building Naxi Language Dependency Treebank Based on Chinese-Naxi Language Relationship Alignment

Gao Sheng-Xiang[1], An Ming-Jia[1], Mao Cun-Li[1], Xian Yan-Tuan[1] and Yu Zheng-Tao[1, 2]

[1]School of Information Engineering and Automa-tion, Kunming University of Science and Technology, Chenggong District, 650500, Kunming, China
[2]Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Chenggong District, 650500, Kunming, China
ztyu@hotmail.com

Abstract

Relative to Chinese, as to Naxi Language, its corpus is very rare, its annotation is also difficult, and these factors make its Syntactic Analysis much too difficult. Aiming to the problem, in the paper, it is proposed a method for building Naxi Language Dependency Treebank based on Chinese-Naxi Language relationship alignment. Firstly, the corresponding words of Chinese-Naxi sentence pairs are aligned; then, the dependency grammar on Chinese sentences; Finally, some characteristics and rules of Naxi Language in itself being considered, the generated Chinese Dependency Tree is mapped to Naxi Sentence by using Chinese-Naxi Languages relationship alignment, as a result, Naxi Dependency Parsing Tree is generated. Experimental results show that: This approach can simplify the process of manual collection and annotation of Naxi Treebank, and save manpower and time to build the dependency treebank of Naxi Language.

Keywords: *Naxi Language; Dependency Treebank; Dependency Parsing; Word Alignment*

1. Introduction

It is of great academic and practical values that can promote the informatization that Naxi Language is translated into Chinese. In the process of the translation, Naxi Language Syntactic Analysis is a very important basic work. As Naxi Language structure is simple, affiliation between the words is very clear, therefore, in comparison to Phrase Tree Parsing, Dependency Tree Parsing is more applicable to Naxi Language. Naxi Language is difficult to be tagged, so the construction of Naxi Language Dependency Parsing annotation system and Dependency Treebank has become the core work of Naxi Language Dependency Parsing. If the problem can be effectively solved, it will provide strong support to the further application such as Naxi Language Parsing, Machine Translation, Information extraction etc.

The research work about Dependency Treebank has been carried out in some foreign language. Some of well-known Treebank are as follows, Czech Prague Treebank[1], English PARC Treebank [2], Italian and other languages Treebank [3-4]. In the field of Chinese Dependency Treebank construction, Lai, who carried out the works of analysis and annotation of dependency grammar, but the tagging corpus, is too small and there are just about 5000 words [5]. In Chinese area, the following treebanks are widely acceptted: （1） TCT,Tsinghua phrase structure Treebank contains 1,000,000 words and about 50,000 sentences [6]；（2）Penn Chinese Treebank, University of Pennsylvania Chinese Treebank, phrase structure type, contains 780,000 words [7], 28000 sentences;（3） Sinica Treebank 3.0, Academia Sinica, Taiwan, similar to the phrase structure type,

ISSN: 2005-4270 IJDTA

360,000 words, 60000 sentences [8]; （4） HIT-CIR-CDT, Chinese Treebank of HIT Social Computing and Information Retrieval Research Center, dependency structure type, 1.2 million words, 60000 sentences [9].

It can be seen that the construction of treebank has made great achievements, and there have been also some studies as to Naxi-English and Chinese-Naxi Bilingual Word Alignment in domestic, the details can be seen from the following references [10-,15]. However, there are difficulties in corpus annotation of Naxi language, and fewer people who are versed in Naxi language. So far, it is almost nobody that casts attentions to the construction of Naxi language Dependency Treebank. Therefore, this brings some difficulties to the work of Dependency Analysis of Naxi language, and the difficulties are just the main problems to be solved in this paper.

2. The Process of Dependency Treebank Building Based on Chinese-Naxi Languages Relationship alignment

2.1. Chinese-Naxi Language Word Alignment

Word Alignment is a very important concept in the area of Statistical Machine Translation. Figure1 shows an example about Chinese - Naxi Language Word Alignment. In this example, there are five words need to be aligned: (老爷爷(old man),), (坐在(is sitting on),), (桥头(the bridge),), (喝茶(drink tea),). In this article, we regard the representation of P. F. Brown [16] as an example and then Chinese-Naxi Language Word Alignment can be expressed as follows: (|老爷爷(old man)(1) 坐(is sitting)(3) 在(on)(3) 桥头 (the bridge) (2) 喝(drink)(4) 茶(tea)(4)). Wherein, The number behind the Chinese word indicate the position of the Naxi word which is aligned to this Chinese word. For example,”桥头 (the bridge) (2)” indicate” 桥头 (the bridge)” aligned with " " which is the second word in Naxi sentence.

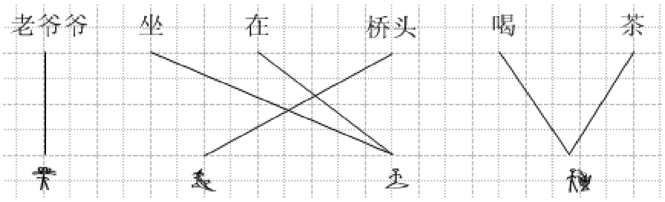


Figure 1. The Example of Word Alignment

We use open-source tools GIZA++ to align words of Chinese-Naxi Language pairs. GIZA package was firstly proposed by the Johns Hopkins University. Later, Och *et al* make GIZA package more optimizing and it was called GIZA++. GIZA++ implements the five Machine Translation models which were proposed by IBM. The main idea of GIZA++ is to use the Bilingual Parallel Corpus to make words alignment between different languages. Today, GIZA++ is still the core component of the most Statistical Machine Translation Systems, and has been widely applied in the field of Word Alignment [17].

2.2. Chinese Dependency Parsing

Syntactic Analysis should follow one grammatical system, and the syntax tree representation is determined according to the syntax of the system. Currently, in Syntactic Analysis, Phrase Structure Grammar and Dependency Grammar are widely used. However, phrase structure grammar is proposed and improved based on English, and most of its current study is focused on English. At the same time

Dependency Grammar studies have been carried out in many languages. Thus, we used Dependency Grammar as the grammar of parsing in the experiment. Figure2 is a Chinese dependency tree and the figure shows that： The representation Dependency Grammar is simple and easy to understand. Dependency grammar directly shows the relationship between words without additional grammar symbols. Even though, people without linguistic backgrounds can easily know the dependency grammar well. These advantages are very conducive to tree bank building. And Dependency Syntax Parsing can be applied into a wider scope of language.

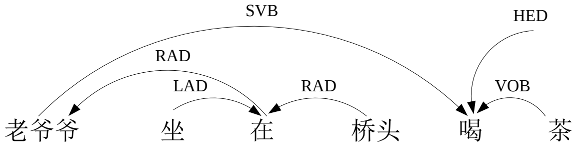


Figure 2. The Structure of Chinese Dependency Tree

The structure of Chinese is complicated. In this experiment, Chinese Dependency Parsing serves for the construction of Naxi Dependency Treebank and we define dependence relationship mainly based on the structural characteristics and semantic relations of Naxi Language. The structure of Naxi sentence is simple. And in order to avoid the problem of data sparseness, we do not define too many dependencies collections. In this experiment, we only defined 10 kinds of dependencies which are shown in Table1.

Table 1. Naxi Dependencies

Relation	Symbol	Relation	Symbol
attribute	ATB	right adjunc	RAD
quantity	QUN	verb-object	VOB
coordinate	COO	subject-verb	SBV
appositive	APP	verb-verb	VV
left adjunc	LAD	head	HED

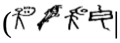
2.3. The Method of Chinese Syntax Tree mapping Naxi Syntax Tree

Now that we have had the base of Chinese-Naxi word alignment and Chinese Syntactic Analysis, The next work that we need to do is to find a mapping between Chinese Dependency Tree and Naxi Language Dependency Tree. In other word, we need to generate Naxi Language Syntactic Dependency Tree based on Chinese Dependency Tree and Chinese-Naxi word aligned Relation.

Obviously, the syntax structure is difference between these two languages. And we have found that Naxi language and Chinese mainly have two different points. During the course of the experiment, we must fully consider the differences between the two languages. Firstly, Some grammatical structures are different between these two languages, for example, the sentence structure ” Subject predicate object ” of Chinese corresponds to the sentence structure ”Subject object predicate” of Naxi Language. In addition to the sentence structure”Subject object predicate” of Naxi Language, Naxi Language usually use attributive rear. This is the major differences of the grammatical structure between Naxi Language and Chinese. Secondly, a lot of Naxi words which equivalent to the Chinese Phrases have multiple meaning due to the small vocabulary of Naxi language. After study Naxi dictionary, we found that One Naxi word can corresponds to one or more Chinese words (One Naxi word corresponds to a Chinese phrase), and basically multiple Naxi words usually do not correspond to one Chinese word, For example, The meaning of “ ” is “黄鼠狼吹火 (weasel is blowing the fire)”, the meaning of “ ” is “听到好消息 (hear the good news)”. In this paper, we call this type of Naxi words for Special Naxi Word. This is also the main research problem in this experiment。

For the first point of the differences, even though, we found that the word order is different between Naxi Language and Chinese, the dependencies is consistent. So we can map the dependencies of Chinese to Naxi Language easily, the following examples is shown the specific method.

“ ” Translated into Chinese is “我喜欢吃鱼”.

When it is handled by word alignment, the translation result is (|我 (i) (1) 喜欢 (like) (4) 吃 (eat) (3) 鱼 (fish) (2)).

Syntactic Analysis on Chinese sentences, As shown in Figure3：

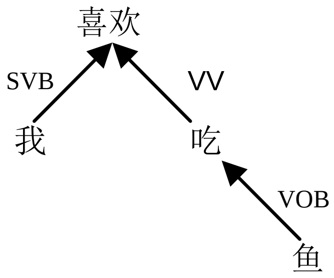


Figure 3. Chinese Dependency Tree

Here we will build Naxi Language Dependency Tree based on the aligned word and Chinese Dependency Tree. The main method is shown in Figure 4:

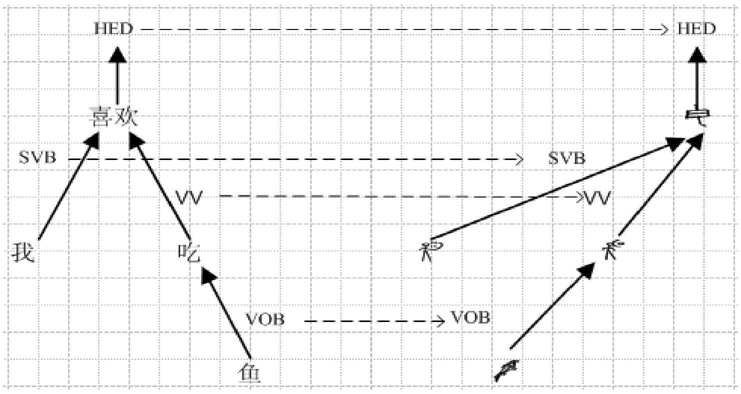
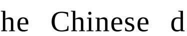


Figure 4. The Generation Methods of Naxi Dependency Tree (1)

Figure4 shows that although the order of Chinese word “喜欢(like)”“吃(eat)”“鱼(fish)”are different from the Naxi word “”, it has no effect on the dependencies. So, we could mapping the Chinese dependencies to the Naxi sentences, and generate Naxi Language Dependency Tree.

For the second point of the differences, thinking of that the number of vocabulary Naxi language is relatively small, we conclude a special dictionary of Special Naxi Word and it is shown in Table 2.

Table 2. The Table of Chinese Phrase-Naxi Word Mapping

Particular Naxi Word	Chinese Phrase
	男子吹笛
	月出东山
	马蹄陷入泥
	开辟田地
	听到好消息
...	...

In the dictionary, there are 144 Naxi words and each of them corresponds to a Chinese phrase. Except these special Naxi words, the relationship between Naxi word and Chinese word is one-for-one corresponding.

In the experiment, we use the core words of Chinese phrases to determine the dependencies of Special Naxi Word. And we assume, in the text, the core word of Chinese phrase is the root node of phrase in Dependency Tree. We take " " as a mapping instance. " " in Chinese means "以油抹发(use oil to wipe hair)". The generation method of Naxi Dependency Tree is shown in Figure 5.

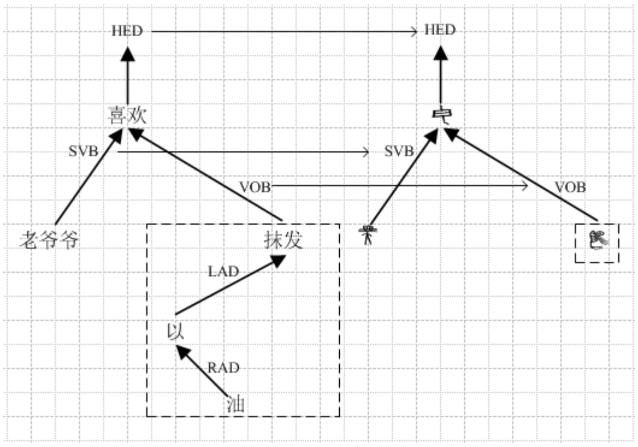


Figure 5. The Generation Methods of Naxi Dependency Tree (2)

In Figure 5, the Special Naxi word " " corresponds to Chinese phrase "以油抹发(use oil to wipe hair)". The dependent node of Special Naxi Word " " is consistent with the core word "抹发 (wipe hair)" of the Chinese phrase. And the dependency of " " is consistent with the core word "抹发 (wipe hair)", too. After the contrast between Chinese sentence and Naxi sentence, we found that most of dependencies and dependent node of Special Naxi words are consistent with the core words of the Chinese phrases, so we can use the core words of the Chinese phrases to determine the dependent node and dependency of Special Naxi Word.

3. Experimental Results and Analysis

The method that we use Chinese as intermediary to build Naxi Dependency Tree use Naxi dictionary and 10,000 Chinese - Naxi Sentence pairs. Chinese dependency parsing is done by CTBparser tool kit, CTBparser tagging set is changed follow the requirements of our research team and the characteristics of the Naxi language. Finally, dependency tree library which contains 10,000 Naxi sentences generates based on Chinese – Naxi mapping.

Meanwhile, we take 3,000 manual annotated Naxi sentences as the initial set on which we use CRFparser machine learning tools to model, and then generate Naxi dependency tree model which we use to extend Naxi sentence. During experiment, a 10,000 Naxi sentence dependency tree library is extended. In this way, we obtained a dependency tree library based on statistical machine learning method. In the testing experiment we use dependency tree library obtained above as contrast to Naxi dependency tree library generated based on using Chinese as an intermediary.

The selection of sentence dependency parsing evaluation indexes is as follows. Dependency arc accuracy (Unlabeled Attachment Score, UAS), identification accuracy (Labeled Attachment Score, LAS) and the root node correct rate (Root Accuracy, RA) are defined as follows.

$$UAS = \frac{\text{arc correct words}}{\text{all words}} \times 100\%$$

$$LAS = \frac{\text{arc correct and dependencies words}}{\text{all words}} \times 100\%$$

$$RA = \frac{\text{the number of correct root in sentences}}{\text{all sentences}} \times 100\%$$

The testing results of Naxi Dependency Treebank Constructed using Chinese as an intermediary are shown in Table 3:

Table 3. The Experimental Results

	UAS	LAS	RA
CRFparser Building dependency tree library of Naxi language	75.56%	75.32%	81.79 %
Built using Chinese as an intermediary library of Naxi language dependency tree	79.13%	77.31%	85.23 %

We can see from Table3, obviously, Naxi language dependency Treebank which is generated by rule-based mapping method based on Chinese have a higher accuracy rate.

Naxi language structure is simple. We can conclude from the experiments that introducing some of rules when generating the Naxi language dependency Treebank can avoid the process of human-annotated corpus and improve the accuracy of dependency tree library rate compared with statistical machine learning. However, the current methods used in the experiments still contain some shortcomings. By analyzing of the error, we found that the errors mainly focus on special Naxi words and the possible reason is that the process of these specific words is rough. In the future research, we will focus on studying special Naxi word, and do classification of special Naxi words, in order to improve Naxi Dependency Treebank accuracy.

4. Conclusion

In this paper, it is proposed a method of building Naxi language dependency Treebank based on Chinese-Naxi Language relationship alignment. This method avoids the artificial tagging process, also improves the accuracy rate in comparison to the traditional statistical machine learning methods, and solves some difficult problems during constructing Naxi language dependency Treebank resources. In the future research work, it will be conducted some experiments about constructing Naxi dependency Treebank based on alignment relationship between Naxi language and different languages. And then the experiment results will be comparatively analyzed with those from Chinese-Naxi language alignment relationship. Finally, it may be implemented to construct Naxi Dependency Treebank by integrating Multilingual-Naxi alignment features together.

Acknowledgements

This work was supported by National Natural Science Foundation of China, Contract No. 61163022.

References

[1] J. Hajic, “Building a Syntactically Annotated Corpus: The Prague Dependency Treebank”, Issues of Valency and Meaning, **(1998)**, pp. 106-132.

[2] T. H. King, R. Crouch, S. Riezler, M. Dalrymple, and R. Kaplan, “The PARC700 dependency bank”, Proceedings of the EACL03, 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), **(2003**) April 13-14, pp.1-8, Budapest.

[3] I. Boguslavsky, S. Grigorieva, N. Grigoriev, L. Kreidlin, and N. Frid, “Dependency Treebank for Russian”, Concept, Tools, Types of Information. Proceedings of the 18th International Conference on Computational Linguistics (COLING), **(2000)** July 31-August 4, pp.987-991, Germany.

[4] C. Bosco and V. Lombardo, “Dependency and relational structure in Treebank annotation”, Proceedings of Workshop on Recent Advances in Dependency Grammar at COLING'04, **(2004)**, Geneve, Switzerland. Pp.1-8.

[5] T. B. Y. Lai and C. N. Huang, “Dependency-based Syntactic Analysis of Chinese and Annotation of Parsed Corpus”, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, **(2000)** October 1-8, pp.255-262, Hong Kong, China

[6] Q. Zhou, “Annotation Scheme for Chinese Treebank. Journal of Chinese Information Processing”, **(2004)**

[7] F. Xia, “The Segmentation Guidelines for the Penn Chinese Treebank (3.0)”, IRCS Technical Reports Series, **(2000)**.

[8] C.Huang, F.Yi Chen, KeJian Chen,“Sinica Treebank: Design Criteria”,Annotation Guidelines, and Onli ne Interface, Proceedings of the Second Chinese Language Processing Workshop, (2000) October; Hong Kong, China. pp. 29-37

[9] Xin Chen. Active Learning for Chinese Dependency Treebank Building. J. Harbin Institute of Technology. **(2011)**

[10] X. Yang, Z. Yu, J. Guo, X. Pan and C. Mao, “Naxi-Chinese Bilingual Word Alignment Method Based on Entity Constraint”, Lecture Notes in Computer Science, vol. 8229, no. 3, **(2013)**, pp.378-386.

[11] L. Li, Z. Yu, C. Mao and J. Guo, “The Extracting Method of Chinese-Naxi Translation Template Based on Improved Dependency Tree-To-Strin”,. Lecture Notes in Computer Science,vol. 8229, no. 3, **(2013)**, pp.350-358.

[12] Z. Yu, Y. Xian, T. Wei, J. Guo and Z.Tao, “Naxi-English Bilingual Word Alignment Based on Language Characteristics and Log-linear Model. China Communications, no. 3, **(2012)**, pp.78-86.

[13] Z. Yu, T. Zhang, J. Guo and C. Mao, “Naxi-English Bilingual Word Alignment Based on Language Feature Model of Naxi’, Proceedings of 7th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE2011), **(2011)** November 27-29, Tokushima, Japan.

[14] Z. Tao, Y. Zhengtao, G. Jianyi and C. Xianbi, “A bilingual word alignment algorithm of Naxi-Chinese based on feature constraint models”, Journal of Xi'an Jiaotong University, vol. 45, no. 10, **(2011)**, pp.48-53.

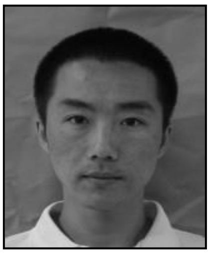
[15] Z. Huihui, Y. Zhengtao, S. Longhua, G. Jianyi and H. Xudong, “Naxi Sentence Similarity Calculation Based on Improved Chunking Edit-distance”, International Journal of Wireless and Mobile Computing, vol. 7, no. 1, **(2014)**, pp.48-53.

[16] P. F. Brown *et al*. “A Statistical Approach to Machine Translation. J. Computational Linguistics” **(1990)**.

[17] S. Xiang and L. Yu-jian, “Computational Performance Analysis of GIZA”, Computer Engineering & Science. **(2010)**.

Authors

Gao Sheng-Xiang, is currently A Ph.D. candidate at Kunming University of Science and Technology, Kunming, China. She has been a CCF member since 2013. She received her M.S. degree in Pattern Recognition and Intelligent System from Kunming University of Science and Technology in 2005. Her research interests include nature language processing, machine translation and information retrieval.



An Ming-Jia, is currently A M.S. candidate at Kunming University of Science and Technology, Kunming, China. His research interests focus on nature language processing and machine translation.

Yu Zheng-Tao, is currently a professor and Ph. D supervisor at the School of Information Engineering and Automation, and he is also the chairman of Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, China. He received his Ph.D. degree in computer application technology from Beijing Institute of Technology, Beijing, China, in 2005. His main research interests include natural language processing, machine translation and information retrieval.

Mao Cun-Li, is currently A Ph.D. candidate at Kunming University of Science and Technology, Kunming, China. He has been a CCF member since 2011. He received his M.S. degree in Technology of Computer Application from Kunming University of Science and Technology in 2011. His research interests include nature language processing, machine translation and information retrieval.

Xian Yan-Tuan, is currently A Ph.D. candidate at Kunming University of Science and Technology, Kunming, China. He received his M.S. degree in Pattern recognition and intelligent system from Shengyang Institute of Automation (SIA), Chinese Academy of Science, in 2006. His research interests are, machine translation and information retrieval.