International Journal of Multimedia and Ubiquitous Engineering
Vol.12, No.2 (2017), pp.29-40
http://dx.doi.org/10.14257/ijmue.2017.12.2.03

Building Laos Dependency Treebank by Means of Chinese-Laos Bilingual Corpus of Word Alignment

Ruochen Yin[1,2], Lanjiang Zhou[*1,2], Feng Zhou[1,2] and Fajie Li[1,2]

[1]School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China
[2]Intelligent Information Processing Key Laboratory, Kunming University of Science and Technology, Kunming 650500, China; or Affiliation(s)
915090822@qq.com

Abstract

As few studies on Laos, there has not built relatively large dependency Treebank. Compared with the rich and mature Chinese corpus, the syntactic analysis of Laos is more difficult and still a urgently controversial issue. The existing machine learning methods need a lot of training corpus, these methods are not fully applicable to Laos and the accuracy of Laos processing results is also low. This paper presents an approach of Chinese-Laos bilingual corpus of word alignment to built Laos Dependency Treebank method. Firstly, the aligned word processing was made by Chinese-Laos sentence pairs. Secondly, the dependency parsing was done with Chinese sentences. Finally, Laos Dependency Parsing Treebank was generated by Chinese-Laos Languages align relationship and Chinese Dependency Tree. The results show that the accuracy of this method has been improved significantly compared with machine learning methods. This approach not only simplifies the process of manual collection and annotation of Laos Treebank, but saves the manpower and time in building the Treebank as well. In the case of the scarcity of Laos corpus, this approach can automatically build the Laos Dependency Treebank with high quality.

***Keywords**: Laos Dependency Treebank; Chinese Dependency Parsing; Word Alignment*

1. Introduction

Lao is China's neighboring country, the exchanges between the two peoples has a long history. Language plays a very important role during the communication. So it has important practical significance for the research of Chinese and Laos bilingual. As a fundamental work in the translation process between Laos and Chinese, Laos syntactic parsing requires a series of analytical procedures to obtain the complete syntax tree of the sentences. Shallow parsing doesn't require the complete parsing tree but the recognition of some components with simple structures, namely, it turns the syntactic parsing into two sub-tasks: identification and analysis of the lexical chunks, and the dependency analysis among lexical chunks. Because of the relatively large difficulty of complete syntactic parsing, the shallow parsing becomes the mainstream of syntactic analysis [1]. Dependency parsing is a very effective way for the machine language syntax analysis. This paper uses the dependency tree method for the Laos syntactic parsing. The construction of Laos dependent labeling system and dependency Treebank has become the core work of the whole Laos dependency analysis. Effective and reasonable solution to this problem can provide a strong support for the upper applications such as syntactic parsing, machine translation, information acquisition, *etc.* The researches of dependency

Lanjiang Zhou, * Corresponding Author

ISSN: 1975-0080 IJMUE

parsing and dependency Treebank construction have been performed in China and other countries. There are some famous dependency Treebank such as Czech Prague Treebank [2], English Parc Treebank [3], Russian Treebank, and Italian Treebank [4-5]. There is also some influential Chinese dependency Treebank, such as Harbin Institute of Technology Social Computing and Information Retrieval Research Center Chinese Dependency Treebank (HIT-CIR-CDT) [6] which constructed by Harbin Institute of Technology Social Computing and Information Retrieval Research Center (HIT-SCIR) with a size of 1.2 million words and 6 million sentences. In the construction of Laos dependency Treebank, there is no people involved in this field around the world. The research in this field is basically a blank.

From the above analysis, we can see that the construction of large languages Treebank has made some achievements. But for Laos, its research work is very little and don't have a dependency Treebank with a substantial scale. Just like Chinese, labeled dependency Treebank is the essential resource for the Laos statistical dependency syntactic structure analysis. How to construct the Laos Dependency Treebank is the key problem solved in this paper. According to the characteristics of Laos, this paper presents an approach to built Laos Dependency Treebank on the basis of Chinese-Laos bilingual corpus of word alignment. Comparing with machine learning method, the experimental results show that the method proposed in this paper have some improvement in Unlabeled Attachment Score(UAS), Labeled Attachment Score(LAS), Root Accuracy(RA).

2. The Differences between Chinese and Laos

Some grammatically structural differences between Laos and Chinese via the comparative studies are shown as follows:

(1) In Laos, both the descriptive attribute and the restrictive attribute are usually placed after the modified central words which they modified(except for some idiomatic usage). The order of the sentence elements in Chinese is: (Attributive) Subject+ Predicate+ (Attributive) Object. And the order in Laos sentence is: Subject (Attributive) + Predicate + Object (Attributive). For example, the sentence "他父亲开新车(His father drives a new car)" in Laos is "ພໍ່(父亲)ເຂົາ(他)ຂັບ(开)ລົດ(车)ໃຫມ່(新)". Hence, as modifiers, the attributive are placed at the back of central word which could be a subject or predicate.

(2)The most basic usage of quantifiers is to be placed at the back of numeral act as a noun classifier (use nouns as classifiers) and modified noun as a attributive. If the numeral itself is two or greater than two, the order is noun + numeral + quantifier (*e.g.*, the order of the phrase "两张桌子(two desk)" in Laos is "桌子两张", "ໂຕະ(桌子)ສອງ(两)ຫນ່ວຍ(张)"). If the numeral itself is one, the order is: numeral + quantifier + noun *e.g.*, the order of the phrase "一只鸭(a duck)" in Laos is"鸭只一", "ເປັດ(鸭)ຫນຶ່ງ(一)". The quantifiers could also be put in front of the adjective to compose a phrase. This phrase usually modifies the noun in front of it and the order of this kind of sentence is: noun + quantifier + adjective.

(3)The auxiliary word we use to connect adverbial in Chinese is "地" and in Laos is mainly "ຢ່າງ" and sometimes "ດວຍ". In Chinese, the adverbial elements are generally placed in the back of the subject and in front of the predicate verb. But in Laos, the time, cause, purpose, degree adverbial are generally be placed in front of the whole sentence or sometimes in the end of the sentence.

(4)The adverbial connected by the auxiliary word "ຢ່າງ" is generally placed in the back of the predicate verb in Laos. But if there is an object in the sentence then the adverbial should be placed in the back of the object, that is, at the end of the sentence. For example, to “ນ້ອງຊາຍຂອງຂ້ອຍຮຽນໜັງສື[ຢ່າງດຸໝັ່ນຂະຫຍັນພຽນ].” “我弟弟[勤奋]学习(My brother studies [hard].).”. The predicate of this sentence is the verb-object structure, therefore, the adverbial must be placed at the end of the sentence, which is placed in the back of the object.

(5)The manner adverbial generally placed at the end of the sentence in Laos. For example, to “ເພິ່ນມາເຮັດວຽກຢູ່ທີ່ປະເທດລາວ[ຕາມຄໍາເຊີນ].” “他是[受邀]来老挝工作的(He was [invited] to work in Laos.)”.

3. Chinese and Laos Word Alignment

Word alignment is a very important concept in statistical machine translation. An example of the word alignment between a Chinese sentence and a Laos sentence is shown in Figure 1.

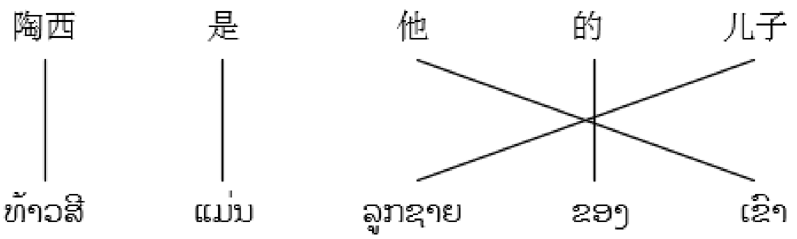

Figure 1. Example of Word Alignment

In this example, there are 5 pairs of words have to be aligned: ທ້າວສີ(陶西)ແມ່ນ(是)ລູກຊາຍ(儿子)ຂອງ(的)ເຂົາ(他). In this paper, the Chinese - Laos sentence on word alignment can be expressed as follows:(ທ້າວສີ ແມ່ນ ລູກຊາຍ ຂອງ ເຂົາ|陶西(1)是(2)他(5)的(4)儿子(3)). The figures behind the Chinese words indicate the position of the aligned words in Laos sentence (*e.g.*, “儿子(3)” means that the word “儿子” is aligned with the third word “ລູກຊາຍ” in the Laos sentence.). In this paper, we use the open source tool GIZA++ [7] to deal with the word alignment of Chinese-Laos parallel sentence pairs. The accuracy of word alignment result we got was 45.37%, so we need to adjust the result manually. The workload is about 400 thousand words and finally we get the parallel sentence pairs which have been word aligned with high quality. GIZA software package was first realized by The Johns Hopkins University machine translation summer camp, after that, Och *et al.* optimized the GIZA software package and called it GIZA++. GIZA++ realized the 5 machine translation model proposed by IBM Company, its main idea is to use bilingual parallel corpus to carry out word alignment training and finally gets the result of the word alignment. Today, GIZA++ is still the core component of most statistical machine translation systems, and has a wide range of applications in terms of word alignment.

4. Chinese Dependency Parsing

The mission of syntactic parsing is to derive the syntax structure of the sentence automatically according to given grammars. At present, the research of syntactic parsing mainly includes phrase structure grammar and dependency grammar. The phrase structure tree consists of three symbols (*i.e.*, the terminator, the non-terminal and the phrase mark) which are constituted with specific grammatical rules. In phrase structure grammar, some terminals constitute a phrase and act as non-terminals to involve in the next reduction until the whole sentence has been reduced to a root node. In dependency grammar, the predicate verb which is not affected by any other components is considered as the center to control other ingredients, and all the dominated parts are subordinated to their dominants in a certain dependency relation. From the above, we can see that dependency grammar has gradually become the research topic of today's researchers because of simple form, easy to label, convenient and so on.. So the research of dependency grammar has been carried out in many languages. In this paper, we use the dependency grammar as the grammar system of syntactic parsing [8]. A Chinese dependency syntax tree is shown in Figure 2.

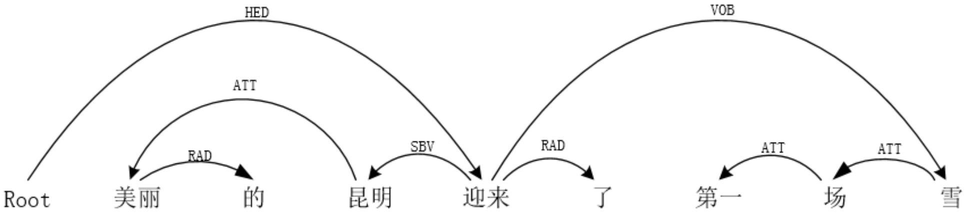

Figure 2. Structure of Chinese Dependency Tree

From Figure 2, we can see that the expression form of dependency grammar is simple and easy to be understood. Dependency grammar is a direct representation of the relationship for words, and there is no additional syntax notation. So even non-professional people are quite easy to understand the syntax form, this is very beneficial to the construction of Treebank. The main semantic relations between Chinese and Laos are similar. Chinese dependency parsing is the premise of Laos dependency Treebank construction. According to the structural characteristics and semantic relations of Laos and to avoid the data sparse problem, this paper defines the dependency relations as shown in Table 1. The experiment mainly based on the 14 main dependency relations in the table.

Table 1. Laos Dependency Relations

Type of relationship	Description	Example
SBV	subject-verb	我送他一本书(我 <-- 送)
VOB	verb-object, direct-object	我送他一本书(送 --> 书)
IOB	indirect-object	我送他一本书(送 --> 他)
FOB	fronting-object	他什么水果都吃（水果 <-- 吃）
DBL	Double	妈妈喊我吃饭（喊 --> 我）
ATT	Attribute	小狗（小 <--狗 ）
ADV	Adverbial	非常优秀（非常 <-- 优秀）
CMP	Complement	吃完了饭（吃 --> 完）
COO	Coordinate	大象和蚂蚁(大象 --> 蚂蚁)
POB	preposition-object	在教室里（在 --> 里）
LAD	left adjunct	大象和蚂蚁(和 <-- 蚂蚁)
RAD	right adjunct	同学们（同学 --> 们）
IS	independent structure	两个单句在结构上彼此独立
HED	Head	指整个句子的核心

5. The Mapping from Chinese to Laos Syntax Tree

Based on the aforementioned Chinese-Laos word alignment and the Chinese syntax analysis, the next task is mapping the dependency relationship from Chinese to Laos. Then the Laos dependency syntax tree is generated according to Chinese dependency syntax tree and the relationship between Chinese and Laos word alignment. Although the word order in Laos sentence is different with that in the Chinese sentence, the dependency relation is just the same. So we can directly map the Chinese sentence dependency relation to the Laos sentence, the specific method is shown in the following example.

Laos: ເຈົ້າ ເຂົາ ເຮັດວຽກ ຢູ່ ສະຖານີວິທະຍຸ (1-1)

Chinese: 他的妈妈在广播电台工作(His mother works in a radio station) (1-2)

After word alignment, the result is:

Laos: ເຈົ້າ(1)ເຂົາ(2)ເຮັດວຽກ(3)ຢູ່(4)ສະຖານີວິທະຍຸ(5) (2-1)

Chinese: 他(2)的妈妈(1)在(4)广播电台(5)工作(3) (2-2)

In Sentence (2-1) and (2-2), the words with the same number in the brackets behind them mapped with each other. After syntactic parsing for the Chinese sentence, we get the syntactic parsing tree, as shown in Figure 3.

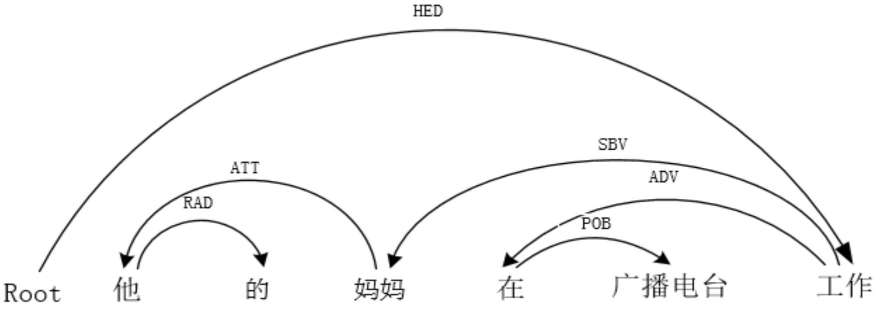

Figure 3. Dependency Tree of Chinese Sentences

The combination of the Laos grammatical features, and the generation of dependency parsing tree of the Laos sentence based on the previous word alignment and the Chinese dependency parsing tree are shown in Figure 4. The dependency relation is not affected although the position of word “ເຮັດວຽກ” in the Laos sentence is not consistent with word “工作” in the Chinese sentence. After the study of Laos and Chinese grammar structure, it is found that the dependence structure of the two languages is equivalent. So we can generate the Laos dependency parsing tree by mapping the dependency relation of Chinese sentence directly to the Laos sentence. However, the ambiguity is generated considering the differences of these two languages. As the word alignment shown in Figure 4, the word “的” in the Chinese sentence has no mapping object in the Laos sentence. The analysis of the Laos sentence is not affected because the dependency relation of the Laos sentence has been clearly analyzed.

Some words in the Laos language correspond to a Chinese phrase. This paper summarizes a special Laos dictionary, as shown in Table 2. This dictionary contains 113 words and each Laos word corresponds to a Chinese phrase. Except these relatively special Laos words, the other Laos words and Chinese words are basically mapped with each other.

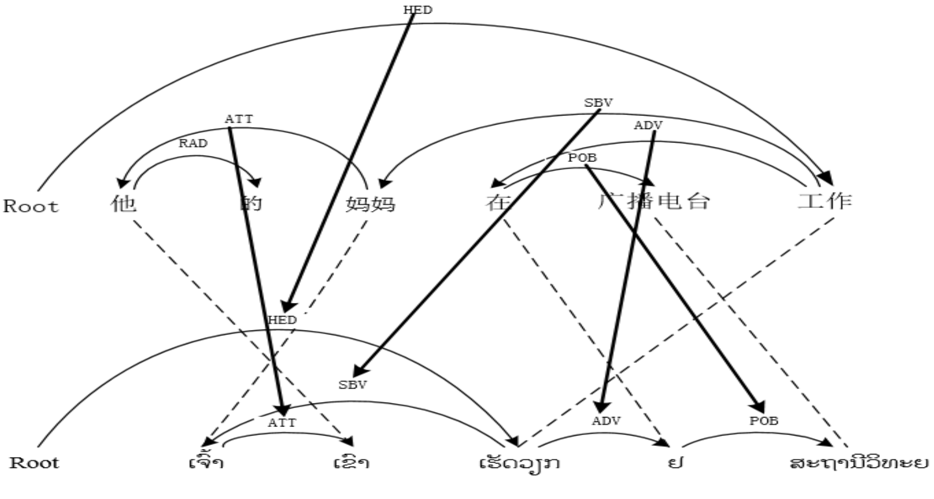

Figure 4. The First Way to Generate Laos Dependency Tree

In the experiment, the dependency relation of these special Laos words is determined according to the key words in Chinese phrases. This paper stipulates that the key word of Chinese phrase is the root node of the phrase in the dependency tree. For example, we use the Laos word “ໂທລະສັບ” to make a map, the word “ໂທລະສັບ” means “打电话 (make a call)” in Chinese. The result is shown in Figure 5.

Table 2. The Comparison of Chinese Phrases in Laos Words

Special Laos words	Chinese phrase
ໂທລະສັບ	打电话
ກົດເທື່ອດຽວ	点击一下
ຈະກັບມາ	马上回来
ໄຂມັນຫຸ້ມ	脂肪变性
ສະບາຍດີ	打个招呼
...	...

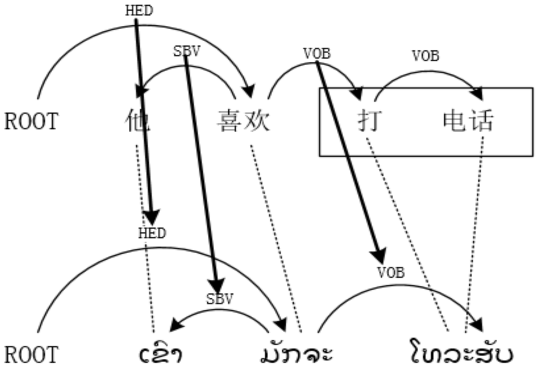

Figure 5. The Second Way to Generate Laos Dependency Tree

In Figure 5, the special Laos word “ໂທລະສັບ” corresponds to the Chinese phrase “打电话”, both the dependency node and the dependency relationship of the word “ໂທລະສັບ” are the same with the key word “打” in Chinese phrase “打电话”. After the comparative study of Chinese sentences and Laos sentences, we found that the dependency relationships and dependency nodes of most special Laos words are consistent with the corresponding Chinese phrases. So we can identify the dependency nodes and dependency relationships of the special Laos words through the key word in Chinese phrases.

6. Experiment and Interpretation of the Result

6.1. Experimental Data

Experimental data is the news from international channel on 7 news sites. These sites cover the major mainstream news sites and contain news with sports, politics, entertainment, military and other aspects, therefore, we can ensure the diversity of the experimental data.

6.2. Evaluation Method

The evaluation index of sentence dependency parsing contains Unlabeled Attachment Score(UAS), Labeled Attachment Score(LAS), Root Accuracy(RA). The definition is as follows.

$$UAS = \frac{N_a}{N} \qquad LAS = \frac{N_{al}}{N} \qquad RA = \frac{N_{sr}}{N_s}$$

N is the number of words in the test corpus. The dependency relation of all words is represented by the triad (Word, Headword, Relation). The “Word” stands for word itself, the “Word” is dependent on “Headword” with the relation of “Relation”. N_a is the number of the words which their “Headword” are right. N_{al} is the number of the words which their “Headword” and “Relation” are both right. N_s is the number of the sentences. N_{sr} stands for the number of the sentences which their root nodes are right.

6.3. Result Analysis

The test corpus contains 30,000 pairs of sentences. This paper use Harbin Institute of Technology's LTP [9] platform to complete the work of Chinese dependency parsing. First, this paper adjusts the Label set to meet the requirements of the experiment and the characteristics of Laos, then generates dependency Treebank by the Chinese-Laos mapping. The following are the experimental results when the Statistical quantities are 10000, 20000, 30000, respectively. As shown in Table 3.

Table 3. The Experimental Result of Building Laos Dependency Treebank by Using Chinese as a Medium

Corpus number	UAS%	LAS%	RA%
10,000	72.13	70.27	80.93
20,000	75.36	73.51	84.31
30,000	76.22	73.34	84.64

At the same time, this paper use the MaltParser [10] and the MSTParser [11] to realize the machine learning modeling with an initial set contains 5000 manually labeled Laos sentences, then generate the dependency tree model. After that, the dependency tree model is applied to expand the Laos sentences. In the experiment, 30000 Laos sentences were expanded. The dependency Treebank is generated by this methods based on the statistical machine learning, then applied to compare with the experimental method of building Laos Dependency Treebank by using Chinese as a medium. The results of the experiment are shown in Table 4.

Table 4. Comparison of Other Methods and the Method of This Paper

Method	UAS%	LAS%	RA%
MaltParser	73.13	69.31	81.26
MSTParser	72.65	70.68	79.93
The means of Chinese-Laos Bilingual Corpus of Word Alignment	76.22	73.34	84.64

From Table 3 and Table 4, it can be seen that compared with the methods of machine learning the accuracy of method in this paper is obviously improved when the Laos corpus is relatively small. In the case of the scarcity of Laos corpus, machine learning methods can’t generate an excellent dependency tree model, but the rule based mapping method is not affected by this.

Laos language structure is similar with Chinese language structure to a certain extent, but also has its special language characteristics. So on the basis of Chinese dependency Treebank we can use the rule based mapping method to generate the Laos dependency Treebank. This can avoid the manual tagging process of the Laos corpus, and can get higher accuracy than the machine learning under the condition of relatively few corpus. Through the analysis of error cases, we find that the performance of this method is very good to deal with the short sentence but its treatment effect of long sentences is relatively poor. It is attributed to the complex structure of the long sentence and the difference between these two languages, which are still needed to be investigated combing with the analysis of language structures in a deeper level. There is also a part of the error caused by the wrong results of the Chinese dependency automatic analysis.

7. Conclusion

This paper proposes a construction method of the Laos dependency Treebank based on the word alignment between Chinese and Laos. This method avoids the process of manually labeling Laos dependency Treebank. In the case of the scarcity of Laos corpus, machine learning methods can’t generate an excellent dependency tree model, but the rule based mapping method is not affected by this. Compared with the traditional statistical machine learning method, this method is simpler, and the accuracies of UAS, LAS and RA are respectively increased by 3.57%, 4.03% and 4.71%. In the next step of research, we will study the dependency relation of the long sentences and the differences between the two kinds of languages. At the same time, we will adjust the Chinese dependency structure, and constantly improve the accuracy of Chinese dependency Treebank, automatically build high quality Laos Dependency Treebank finally. After solving the problem of the resource scarcity of Laos dependency Treebank construction, we will carry out the experimentation of the Laos dependency Treebank construction based on the alignment relationship between different languages and Laos, and compare them with the Laos dependency Treebank constructed by the mapping of Chinese to Laos. Finally We will realize the Laos dependency Treebank's construction experiment combing with the alignment characteristics between several languages and Laos.

Acknowledgments

This paper is supported by National Natural Science Foundation Project (61562049).

References

[1] M. Jinshan, "Research on Chinese Dependency Parsing Based on Statistical Methods", Harbin Institute of Technology, **(2007)**.

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

[3] T. H. King, R. Crouch, S. Riezler, M. Dalrymple and R. M. Kaplan, "The PRAC 700 dependency bank", Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora(LINC-03), Saarbrucken, Germany, **(2003)**, pp. 1-8.

[4] I. Boguslavsky, S. Grigorieva, N. Grigoriev, L. Kreidlin and N. Frid, "Dependency treebank for Russian: concept, tools, types of information", Conference on Computational Linguistics. Association for Computational Linguistics, vol. 2, **(2000)**, pp. 987-991.

[5] C. Bosco and V. Lombardo, "Dependency and relational structure in treebank annotation", Coling Recent Advances in Dependency Grammar, **(2004)**, pp. 1-8.

[6] C. Xin, "Active Learning for Chinese Dependency Treebank Building", Harbin Institute of Technology, **(2011)**.

[7] S. Xiang and L. Y. Jian, "Computational Performance Analysis of GIZA++", Computer Engineering & Science, vol. 32, no. 5, **(2010)**, pp. 147-149.

[8] C. Wanxiang, Z. Meishan and L. Ting, "Active Learning for Chinese Dependency Parsing", Journal of Chinese Information Processing, vol. 2, no. 6, **(2012)**, pp. 18-22.

[9] http://ir.hit.edu.cn

[10] J. Nivre, J. Hall and J. Nilsson, "MaltParser: A Data-Driven Parser-Generator for Dependency Parsing", In Proceeding of LREC-2006, **(2006)**, pp. 2216-2219.

[11] R. McDonald, K. Lerman and F. Pereira., "Multilingual Dependency Analysis with a Two-Stage Discriminative Parser", Tenth Conference on Computational Natural Language Learning, New York City, USA, **(2006)**.

Authors

Ruochen Yin, he is a student in the department of Computer Application Technology from Kunming University of Science and Technology. He got the Master's degree in 2014. His research direction is natural language processing.

Lanjiang Zhou, he is an associate professor in Intelligent Information Processing Key Laboratory from Kunming University of Science and Technology. His research interests include natural language processing and embedded system research.

Feng Zhou, he is an associate professor in Intelligent Information Processing Key Laboratory from Kunming University of Science and Technology. His research interests include natural language processing and software engineering application technology.

Fajie Li, he is a student in the department of Software Engineering from Kunming University of Science and Technology. He got the Master’s degree in 2013. His research direction is natural language processing.