Vietnamese Part of speech tagging based on multi-category words Disambiguation Model

Zhao Chen[1]，Liu Yanchao[1], Guo Jianyi[1,2,†]，Chen Wei[1,2]， YanXin[1,2], Yu Zhengtao[1,2] ，Chen Xiuqin[3]

1. The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China; 2.The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;3. Kunming University of Science and Technology of International Education

†Corresponding author, E-mail: gjade86@hotmail.com

Abstract. POS tagging is a fundamental work in Natural Language Processing, which determines the subsequent processing quality, and the ambiguity of multi-category words directly affects the accuracy of Vietnamese POS tagging. At present, the POS tagging of English and Chinese has achieved better results, but the accuracy of Vietnamese POS tagging is still to be improved. For address this problem, this paper proposes a novel method of Vietnamese POS tagging based on multi-category words disambiguation model and Part of Speech dictionary，the multi-category words dictionary and the non-multi-category words dictionary are generated from the Vietnamese dictionary, which are used to build POS tagging corpus. 396,946 multi-category words have been extracted from the corpus, by using statistical method, the maximum entropy disambiguation model of Vietnamese part of speech is constructed, based on it, the multi-category words and the non-multi-category words are tagged. Experimental results show that the method proposed in the paper is higher than the existing model, which is proved that the method is feasible and effective.

Keywords: multi-category words disambiguation；Vietnamese; part of speech dictionary；POS tagging

1 Introduction

Part of speech tagging (POS tagging) is a typical sequence annotation task in natural language processing, which is to give a correct lexical mark for each word in the sentence. It plays a very important role and is widely used in many aspects of natural language processing, such as chunk parsing, syntactic parsing, named entity recognition, noun phrase recognition, semantic analysis and machine translation etc. The study of the POS tagging in Vietnamese can effectively support the research work of language information processing in Vietnamese, which can be applied to Vietnamese machine translation, information retrieval and speech recognition, which is also the indispensable basis of block recognizer and syntactic analysis of Vietnamese. Therefore, the POS tagging in Vietnamese is one of the key problems and a difficult point in the field of Vietnamese information processing.

The study of the POS tagging had been studied earlier in English and Chinese and has achieved good results. There are three main methods for POS tagging: (1) Rule-based method. Unsupervised learning methods for disambiguation rules are proposed by Brill, E (1999) to solve English POS tagging[1]; Hu Guanlong (2007) proposed an improved conversion method applied to the Latin Mongolian tagging task, and achieved good results[2]; Wang Guangzheng (2008) proposed a priority rule-based POS tagging method, which is applied to the Chinese, the tagging accuracy rate can reach 96.4%[3]; The rule-based method is difficult to count all the rules of language completely, and the complexity of the language phenomenon also led to its rules difficult to develop; (2) Statistics-based method. BernardMeraldo (1994) proposed the use of probability model in English POS tagging[4]; Wang Lijie（2009）put forward the method for Chinese POS tagging based on SVMTool[5]; Binulal, G, S（2009） proposed the method for Telugu POS tagging based on SVM[6]; Nongmeikapam K (2012) used SVM model to do the Manipuri POS tagging[7]; Generally speaking, statistics-based method of POS tagging needs a large scale of training corpus to support machine learning, moreover, if there are fewer frequent part of speech in the training corpus, the marking effect is poor. (3) Hybrid method. Jiang Shangpu (2010) put forward the method of combining rules with statistics to do POS tagging in Japanese [8]. For the POS tagging in Vietnamese, there are already some related work, but the correct rate still needs to be improved, such as, Minh NGHIEM-Dien DINH(2008) integrated the common features (lexical features, word context features, POS features and spelling features) and special features (repeat feature, prefix and suffix features) into SVM model for POS tagging, the correct rate is 93.51%[9]; Oanh Thi Tran (2009) proposed a new feature based syllable, combining the features based word, and integrated them respectively into the model of MEM (maximum entropy), SVM and CRF (conditional random field), modeling and segmentation, comparing the results of the three models[10]; Phuong, Le-Hong (2010) proposed the method of maximum entropy incorporated with two kinds of characteristics, one kind is the basic characteristic, which including the relation between the current word and the front and back word, and the relation between the part of speech of the current word and the part of speech of the front and back words, the other is syllable feature. The correct rate of the method is 93.40% [11]. In addition, the above studies ignored the influence of multi-category words on the quality of POS tagging. Vietnamese is the same as English and Chinese, the phenomenon of multi-category words are common/ubiquitous and occupy a large proportion, which brings great difficulties to the lexical analysis and syntactic parsing. Multi-category words processing is very important in POS tagging in lexical analysis

This paper especially studies the influence of multi-category words in the POS tagging in Vietnamese, the Vietnamese dictionary is used to get multi-category words dictionary and non-multi-category words dictionary, then the corpus is divided into multi-category words and non-multi-category words, and then they are marked separately. The benefit of doing so is, for non-multi-category words, that the POS tagging based on the part of speech dictionary can achieved good experimental results of near 100% accuracy, which is much better than the experimental results based on statistical algorithms and can avoid the possibility of tagging errors in the manual tagging of corpus, thus reducing the workload of tagging the corpus; For multi-category words, we learned from the existing research, combined with the characteristics of the Vietnamese language, constructed the category words corpus, specially selected multi-category words characteristic, which is not considered in the above study. And the method of Vietnamese POS tagging based on multi-category words disambiguation model and Part of Speech dictionary was proposed, which can effectively solve the problem of fewer parts of speech type tagging effect in training corpus and improve the accuracy of POS tagging in Vietnamese.

2 Linguistic Features of Vietnamese and Construction of Part of Speech Set

Vietnamese has its own unique language features, not only in the POS tagging, but also in different tasks. Up till now there has been no standardized POS tagging corpus in Vietnamese language, therefore, a POS tagging set has been built firstly to construct POS tagging corpus in order to verify the effectiveness of the proposed method.

2.1 Linguistic features of Vietnamese

Vietnamese are the mother tongue of the Vietnamese state and belong to the South Asian language. It has the following characteristics [12]:

⑴ It is a language with a fairly fixed word order, in which consists of subject + predicate + object constitutes subject-verb-object (SVO).

⑵ The Vietnamese language is written in a variant of the Latin alphabet, a kind of isolated language, which lack of morphological changes; every morpheme is a simple, isolated syllable;

⑶ Vietnamese is a tortuous form of language, its syllable morphology does not produce any change;

⑷ The greatest impact on Vietnamese is the phenomenon of linguistics: mutation type, that is, some words have multiple parts of speech, but itself has not changed；

Leading to Vietnamese language contains ambiguous part of speech. For example: words: "yêu" as a noun when the "devil", when the verb is "love";

⑸ Each word is composed of one or more syllables (morphemes), and the spaces between syllables and syllables are separated by spaces, for example “cơm”(cook) is composed of one syllable of the word; “dưa chuột” (cucumber) is composed of two syllables; “vội vội vàng vàng” (quickly) is composed of four syllable words. Vietnamese does not have a word delimiter, spaces are generally used to distinguish syllables, and no special markings to distinguish words.

3 Construction of Vietnamese Part of Speech Tagging Set

The effective POS tagging set plays an extremely important role in the work of POS tagging. Firstly, a large POS tagging set will increase the markup difficulties, not only in the notes on the corpus, but also on the final training of the POS tagging model; Secondly, a small POS tagging set can’t provide enough text information. Therefore, we consider the above two factors and select the appropriate annotation set, which can express the information in the sentence decently, this is extremely important. For Vietnamese, it is difficult to design a good POS tagging set, because the classification of words is highly controversial [13-14]. Diep Quang Ban and Hoang Ban [14] found in their research that the classification of words usually requires three conditions: meaning, group word ability and syntactic function. By analyzing and examining the linguistic features of these three aspects, we designed a POS tagging sets with nineteen tags in this paper, as shown in table 1.

Table 1. POS tagging sets

POS	The meaning of POS	examples
N	Common noun	Hiện nay，trung_ương，pháp_chế
E	Preposition	Ở，của，tại
CH	Punctuation	，！？
L	Numeral	Những，mọi，các
A	Adjective	Hươu，khoẻ_mạnh，phì
V	Verb	Bầu，được，uỷ_nhiệm
Ny	Proper noun abbreviation	GDP，USA，CHN
Cc	Alternative conjunction and Coordinate conjunction	Và，hay，hoặc
M	Number	Trăm，4，2008
R	Adverb	Lại，cũng，ngay
C	Conjunction	Nếu，mà，là
Nc	Unit noun	Ngôi，mảnh，thửa
Np	Proper noun	Hưng_Yên，Bắc_Bộ，Đức
Nu	Metric unit word	USD，ha，kg
X	Idioms, sayings, foreign languages	tại_sao,đến_nỗi,nhất_là

P	Pronoun	chúng_ta,tôi,gì
T	Adverb of degree and modal adverbs	Chính,thôi,đi
I	Interjections and modal words	ạ,ơi,hả
Z	Sino-Vietnamese words	Phó,nguyên,tái

4 Construction of Vietnamese POS tagging model

4.1 Dictionary of multi-category words and non-multi-category words

In this paper, we base on the Vietnamese dictionary and get multi-category words dictionary and non-multi-category words dictionary, in which the multi-category dictionary is 1659 words. The Vietnamese dictionary is from 131071 entries in the dictionary made and checked by our laboratory and 30565 entries crawled from the website (http://vdict.com/).

4.2 Multi-categories Words Corpus

In this paper, we collected a lot of news, entertainment, economy and other types of articles from the Vietnamese news website. The corpus should be treated as follows: first, after sorting, to noise and other operations, a text sentence level corpus was formed; secondly, the Vietnamese word segmentation tool was used to segment the text sentence, and manual proofreading by the Vietnamese language expert, thus forming a sentence-level word segmentation corpus; then making the POS tagging and chunking analysis; finally, using the Vietnamese dictionary , by selecting and extracting, to get multi-category words dictionary, in which 1659 words are included. Based on this dictionary, we program and extract 396946 Vietnamese multi-category words from the POS corpus constructed in advance, which will be used to construct the part of speech disambiguation model.

4.3 Building POS tagging model

The method of Vietnamese POS tagging based on multi-category words disambiguation model and Part of Speech dictionary was proposed in this paper, by dividing corpus into multi-category words and non-multi-category words, marking them respectively and to construct a Vietnamese Disambiguation Model, with a view to improving the correct rate of Vietnamese POS tagging. The system block diagram is shown in Figure 2.

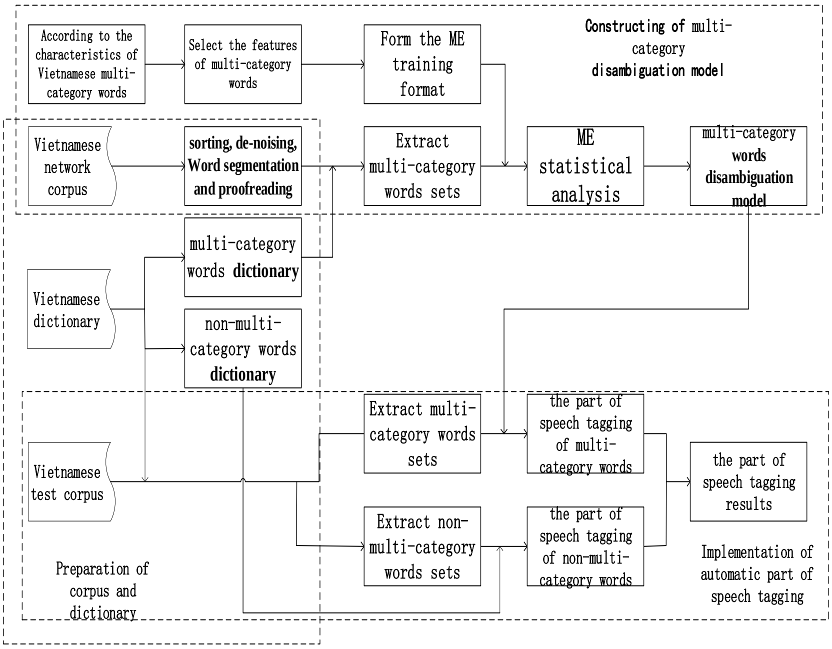
According to the characteristics of Vietnamese multi-category words
Select the features of multi-category words
Form the ME training format
Constructing of multi-category disambiguation model
Vietnamese network corpus
sorting, de-noising, Word segmentation and proofreading
Extract multi-category words sets
ME statistical analysis
multi-category words disambiguation model
multi-category words dictionary
Vietnamese dictionary
non-multi-category words dictionary
Extract multi-category words sets
the part of speech tagging of multi-category words
Vietnamese test corpus
the part of speech tagging results
Extract non-multi-category words sets
the part of speech tagging of non-multi-category words
Preparation of corpus and dictionary
Implementation of automatic part of speech tagging

Fig. 1. POS tagging model building system diagram

The principle of the proposed system is shown as in Figure 2.

(1) Preparation of corpus and dictionary. We based on the Vietnamese dictionary and get multi-category words dictionary and non-multi-category words dictionary; and collected a lot of data from the Vietnamese website. By sorting, de-noising, and other operations, we got the sentences to be marked with parts of speech.

(2) Construction of the POS disambiguation model. Combined with Vietnamese characteristics, main Vietnamese POS tagging feature are selected, and the ME model is used to build the part of speech disambiguation model.

(3) Extraction of the multi-category words and non-multi-category words. We extract multi-category words and non-multi-category words from the test corpus for POS tagging based on the Vietnamese multi-category words dictionary;

(4) Automatic part of speech tagging. We match the non-multi-category words based on the Vietnamese non-multi-category words dictionary, match the multi-category words based on the disambiguation model, and then get the part of speech tagging result; finally, we combine the two results together to get the final result.

5 Construction of the part of speech disambiguation model

Multi-functional words are prevalent in various languages, for the Vietnamese language, which also occupy a large proportion; whether or not the part of speech tag is correct directly affect the sentence POS tagging results, but also affect the surrounding POS tagging; In addition, multi-category word has great influence on text segmentation and Machine Translation. Therefore, to build the disambiguation model for multi-functional words is very important. The existing method to solve the problem is as follows: Zhi Tianyun[15] put forward a POS tagging method in Chinese multi-category words based on Rough Sets a nd fuzzy neural network; Li Huadong[16] proposed a rules-based method for Chinese words Tagging, combined with Chinese language characteristics. Although above methods have achieved very good results, no research has been found in the Vietnamese language. In this paper, we trying to use the maximum entropy (http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html) model to solve the problem, this is because that, when modeling, you only need to focus on selecting features without having to spend much time considering how to use these features. This method has a flexible selection of features, no additional independent assumptions, strong portability, and a combination of rich information. It can make the unknown part of the amount of information to maximize (entropy to maximize) in the case of ensuring that the existing knowledge is not violated, and thus more suitable for the construction of the disambiguation model.

5.1 Selecting the features for multi-category words

The model effect depends mainly on the quality of features, so it is very important to select good features. The main features in this paper are as follows: (1) Context information features of word or between words (word type contains rich form of information); (2) Contextual information features of Part of Speech (the part of speech can represent the decorate relationship between the part of speech); (3) Context information features of chunk or between chunks, which indicates the role of the word in the sentence, modify the relationship and other information; (4) Sentence component features (subject, predicate, adverbial, etc.). As shown in Figure 3.

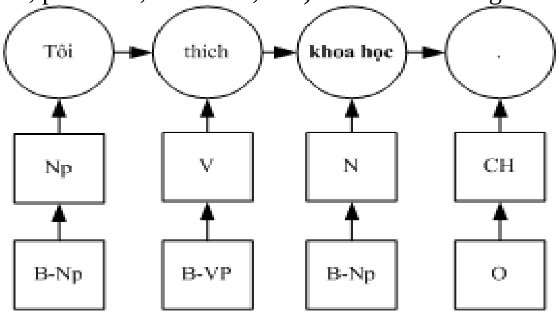

Fig. 2. Example of feature selection

5.2 Constructing of disambiguation model

The construction of Disambiguation Model is a key part of POS tagging model, which can improve the accuracy of POS tagging and provide basic support for follow-up work. The construction principle of the model is shown in figure 4.

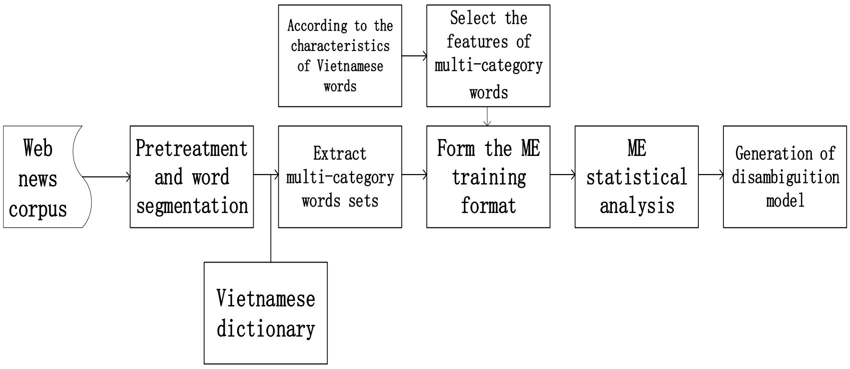
According to the characteristics of Vietnamese words
Select the features of multi-category words
Web news corpus
Pretreatment and word segmentation
Extract multi-category words sets
Form the ME training format
ME statistical analysis
Generation of disambiguition model
Vietnamese dictionary

Fig. 3. multi-category words disambiguation model

Firstly, using the web crawler program designed by us, we get the news and other different types of corpus, and made a pretreatment to them, such as: de-noising, word segmentation and other operations; secondly, based on the Vietnamese dictionary, the obtained corpus is matched to identify the multi-category word in the corpus set. Then, selecting the characteristics of the Vietnamese multi-category words and then integrating them into the training corpus. Finally, uses the maximum entropy model for statistical analysis, and generates the Vietnamese multi-category words disambiguation model.

5.3 Realizing of part of speech tagging in Vietnamese

The detailed process of Vietnamese POS tagging is as follows: (1) Based on the Vietnamese words dictionary, extracting multi-category words and non-multi-category. (2) Using the disambiguation model get POS tagging result. (3) Extracting non-multi-category results from the non-multi-category words dictionary. (4) Combining the two results together, so as to get POS tagging result.

6 Experiment and result analysis

6.1 Data sets

The data sets used in this paper are some Vietnamese sentence which was token up from the Vietnam News website, they are used as the training corpus and the test corpus. These crawled pages form a text corpus through rules extraction, duplicate removal, manual annotation, etc. The total size is the 27878 sentences and 396946 multi-category words field library, and they are encoded by UTF-8.

6.2 Experimental evaluation

Precision is a field metric widely used in information retrieval and statistical classification to evaluate the quality of results. Similarly, we can use this evaluation method to use the word tagging task. With the help of the Vietnamese linguists, 27878 Vietnamese sentences were marked and 396,946 multi-category word fields were extracted by program, they are all for experimentation. The results of the word attribute is evaluated using the precision (P), as shown in formula (1).

$$P = \frac{\text{The number of correct parts of speech tagging}}{\text{The number of parts of speech tagging}} \quad (1)$$

The precision is between 0 and 1, and the closer the value is to 1, the higher the precision, the better the effect.

6.3 Experimental design

In order to verify the performance of proposed system, we designed two sets of experiments to test it. Experiment 1 is mainly using MEM, CRF++, SVM-multiclass model and the proposed method (dictionary + disambiguation model) for comparative experiments; Experiment 2 is mainly used VietTagger developed by Hanoi University (http://vlsp.vietlp.org:8080/demo/?page=resources) and the proposed method in this paper for comparative experiments.

Experiment 1: the 27878 POS tagging of the corpus were divided into ten copies, and then tenfold the cross-validation test was made, i.e., using the popular model of MEM, CRF, SVM as well as the proposed model (dictionary + disambiguation model) for the experiments, comparing the average accuracy rate. The experimental results are shown in figure 5.

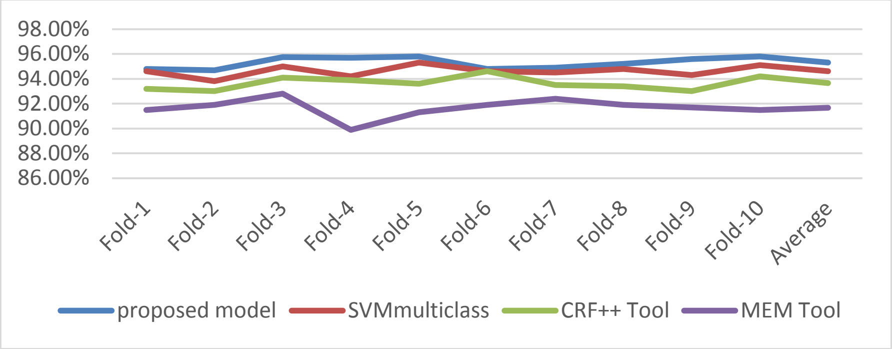

Fig. 4. Ten-fold cross experiment and Model comparison

As can be Seen from Figure 5, the average accuracy rate of MEM, CRF++, SVM-multiclass, the proposed model is respectively 91.62%, 93.71%, 94.67% and 95.22%; in detail, the accuracy rate of SVM-multiclass model is 0.96% higher than that of CRF++, and CRF++ is 2.09% higher than that of MEM. It is worth noting the accuracy rate of the proposed model is higher than SVM-multiclass model 0.55%. The experimental results shown that multi-category words disambiguation model can effectively improve the segmentation accuracy of Vietnamese.

Experiment 2: in order to verify the validity of the proposed system, the proposed model with existing POS tagging tools VietTagger and SVM-multiclass model were compared, the experimental results are shown in Table 6 and figure 7.

Table 2. POS tagging experimental results

System	Precision
VietTagger	92.13%
SVM-multiclass	94.67%
proposed method	95.22%

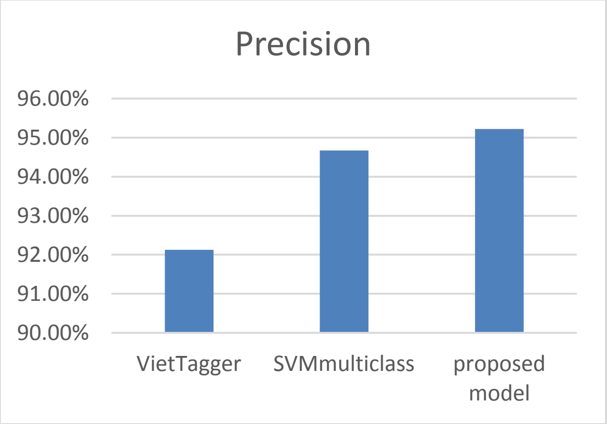

Fig. 5. POS tagging experimental results comparison

As we can be seen from Figure 7, the proposed method has achieved good results, 3.09% higher than that of the VietTagger, and 0.55% higher than that of the SVM multiclass, which proves that our method is effective and feasible.

7 Summary

In this paper, we propose a new method for Vietnamese part of speech tagging based on multi-category words disambiguation model. We first used the crawler program to get corpus on the Vietnam news website, including economic, political, cultural and military fields. Then these obtained corpus are processed to form a text corpus. In order to make better use of the corpus, these texts are divided into sentences as training corpus and test corpus, whose total size is the 27878 sentences and 396946 multi-category words field library, and they are encoded by UTF-8. At the same time, 19 kinds of annotation sets are collated and defined, and 27878 sentences were manually annotated; Furthermore, we select the characteristics of common features, special features and disambiguation of chunks as an effective feature of the proposed model. Finally, we prove the POS tagging effect of the proposed method by comparing with the existing main method. The experimental results show that the proposed method can effectively make the Vietnamese POS tagging with an accuracy rate of 95.22%. We hope our study could lead to more future works.

ACKNOWLEDGMENT

This work was supported in part by the key project of National Natural Science Foundation of China (Grant No. 61732005) and the National Natural Science Foundation of China (Grant Nos. 61262041, 61562052 and 61472168).

References

1. Brill E, Pop M. Unsupervised Learning of Disambiguation Rules for Part-of-Speech Tagging[M]// Natural Language Processing Using Very Large Corpora. Springer Netherlands, 1999:27-42.
2. Hu Guanlong,Zhang Jian，Li Miao.Improved transformation based POS tagging of Latin Mongolian[J].Computer application,2007,27(4):963-965.(in Chinese)
3. Wang Guangzheng,Wang Xifeng.POS tagging method based on rule priority[J].Journal of Anhui University of Technology:natural science,2008,25(4):426-429.(in Chinese)
4. Bernard Meraldo.Tagging English Text with a Probabilistic Model.Computational Linguistics.1994.20(2):l-29.
5. Wang Lijie, Che Wanxiang, Liu Ting.Chinese POS tagging based on SVMTool[J].JOURNAL OF CHINESE INFORMATION PROCESSING,2009,23(4):16-21.(in Chinese)
6. Binulal G S, Goud P A, Soman K P. A SVM based approach to Telugu Parts Of Speech Tagging using SVMTool[J]. International Journal of Recent Trends in Engineering, 2009, 1(2):183-185.
7. Nongmeikapam K, Nonglenjaoba L, Roshan A, et al. Transliterated SVM based Manipuri POS tagging[M]// Advances in Computer Science, Engineering & Applications. Springer Berlin Heidelberg, 2012:989-999.
8. Jiang Shangpu,Chen Qunxiu.Research on Japanese word segmentation and POS tagging based on rules and statistics[J].JOURNAL OF CHINESE INFORMATION PROCESSING,2010,24(1):117-122.(in Chinese)
9. Minh NGHIEM-Dien DINH,Mai NGUYEN.Improving Vietnamese POS tagging by integrating a rich feature set and Support Vector Machines. In Proceedings of Research, Innovation and, Vision for the Future,RIVF,p.128-133.2008.
10. Oanh Thi Tran,Cuong Anh Le,Thuy Quang Ha,Quynh Hoang Le.An Experimental Study on Vietnamese POS tagging. In Proceedings of the International Conference on Asian Language Processing,IALP,Singapore.2009.

11. Phuong Le-Hong,Azim Roussanaly.An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts. In Proceedings of TALN 2010,Montreal,Canada,2010.
12. Xiong Mingming. Research on Vietnamese lexical analysis method [D]. Kunming University of Science and Technology, 2016.
13. Diep Quang Ban, Hoang Ban. Vietnamese Grammar, Book published by the Education Publisher, Hanoi, 2004.
14. Nguyen Chi Hoa. Practical Vietnamese Grammar. Book published by Vietname National University Publisher, Hanoi, 2001.
15. Zhi Tianyun, Zhang Yangsen.The Acquiring Method of Chinese Ambiguity Word POS tagging Rules based on Rough Sets and Fuzzy Neural Network[J].Computer Engineering and Applications,2002,38(12):89-91.(in Chinese)
16. Li Huadong,Jia Zhen,Yin Hongfeng etc.Chinese Ambiguity Word’s annotation based on Rules[J],Computer application,2014,34(8):2197-2201.(in Chinese)