Proceedings of the 35th Chinese Control Conference
July 27-29, 2016, Chengdu, China

Part-of-speech Tagging Based on Dictionary and Statistical Machine Learning

YE Zhonglin[1], JIA Zhen[1+], HUANG Junfu[1], YIN Hongfeng[2]

1.School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
E-mail: zhonglin_ye@foxmail.com, zjia@swjtu.edu.cn, 990504422@qq.com

2.DOCOMO Innovations Incorporation, Palo Alto, USA
E-mail: hongfeng_yin@yahoo.com

Abstract: Part-of-speech tagging is the basis of Natural Language Processing, and is widely used in information retrieval, text processing and machine translation fields. The traditional statistical machine learning methods of POS tagging rely on the high quality training data, but obtaining the training data is very time-consuming. The methods of POS tagging based on dictionaries ignore the context information, which lead to lower performance. This paper proposed a POS tagging approach which combines methods based on dictionaries and traditional statistical machine learning. The experimental results show that the approach not only can solve the problem that the training data are insufficient in statistical methods, but also can improve the performance of the methods based on dictionaries. The People's Daily corpus in January 1998 is used as testing data, and the accurate rate of POS tagging achieves 95.80%. For the ambiguity word POS tagging, the accuracy achieves 88%.

Key Words: part-of-speech tagging, ambiguity word, word segmentation dictionary, maximum entropy, big data

1 Introduction

Part-of-speech (POS) tagging procedure is to give a POS (lexical category) to each word in texts [1].

The Southwest Jiaotong University (SWJTU) Chinese words segmentation system is based on big data and the word dictionary is used to label POS tag for words. The dictionary used in the system consists of words, POS, word frequency and POS frequency. For ambiguity word POS tagging, the system chooses the tag with the highest POS frequency. The POS tagging accuracy of the system is 93.51% and the test data is People's Daily in January 1998. The purpose of the work in this paper is to improve the performance of POS tagging in the SWJTU words segmentation system. The approach based on the combination of statistical machine learning and dictionary is proposed in this paper. Firstly, the original words segmentation dictionary is extended, and then the dictionary is reprocessed and revised according to the corpus of People's Daily in January 1998, which makes the POS tagging results more accurate. Finally, we use both statistical machine learning and dictionary to label POS tag for words. The method proposed in this paper solves the problem of ambiguity word POS tagging, lack of training corpus of statistical methods and ignoring sentence context information based on method using dictionary.

2 Related Works

The POS tagging methods using dictionary and rules firstly split sentences according to the words in the dictionary and then assign a POS tag to each word according to POS tag rules which are determined by the context information [2-5]. In this method, the size and coverage degree of rules have direct influence on the POS tagging performance [6].

The POS tagging methods based on statistical machine learning mainly use some machine learning algorithms, such as Conditional Random Field (CRF), Hidden Markov Model (HMM) and Maximum Entropy (ME) etc. HMM would bring about the problem of the labeling bias [7]. In 2010, Moon [8] proposed the non-boundary HMM model. In 2012, Yuan Litchi [9] used HMM algorithm and semantic parsing results to establish the prediction model, and the accuracy was 94.64%. ME algorithm can make full use of context information. Liu Yaofeng [10] used the ME algorithm to train model and got the accuracy of 93.99% when carrying on the closed test and gained the accuracy of 84.60% when carrying on the open test [11]. CRF model is also able to use

*This work is supported by National Natural Science Foundation of China under Grant No.61572407, No.61262058, National Key Science and Technology Program of China under Grants No.2015BAH19F02, No.2016G04001, National Social Science Fund Major Project of China under Grant No.15ZDB107, and Fundamental Research Funds for the Central Universities under Grants No. 2682015CX070. Corresponding author: zjia@swjtu.edu.cn

context information, and it can solve the problem of labeling bias by probability model [12]. In 2006, Hong Mingcai[13] used the news data of People's Daily from February to December in 1989 to establish the CRF prediction model, and the accuracy reached 98.56%. In 2011, Sun Jing[14] used the dictionary to label POS tagging, and then she used the CRF iteration model to optimize the tagging results. The POS tagging accuracy raised about 1.88%～2.26%.

There are other methods based on the combination of statistical machine learning and rules. Eric[15] proposed the POS tagging method based on the error-driven in 1995. He found out the best POS tagging rules from the result of initial tagger, and then he combined the rules with the result of tagger. Wang Guangzheng[16] added priority for the tagging rules, which solved the problem of ambiguity word POS tagging in 2008, and the tagging accuracy reached at 96.4%. Jiang Shangpu[17] put forward a POS tagging approach which combined statistics and rules in 2010, which adopted joint words segmentation and POS tagging algorithm as the basic framework, and the POS tagging F-Measure value reached 94.8%.

3 POS Tagging Set

There are many POS tagging sets, such as 863 POS tagging set[18], which divides POS into 28 categories. The People's Daily tagging set has 26 basic POS tags[19]: noun, verb, adjective etc.. In addition, from the view of application needs, it also has added some proper nouns: name, address, punctuation mark, etc., which forms 46 POS tag in total. ICTCLAS POS tagging set[20] is based on the People's Daily tagging set, and it subdivides some tags and finally get 99 POS tags. There are 22 one-level POS tags, 66 second-level POS tags and 11 third-level POS tags in 99 tags. The SWJTU Chinese word segmentation system is further subdivide the ICTCLAS POS tagging set, and it has 26 one-level POS tags, 50 second-level POS tags and 50 third-level POS tags, which are 126 POS tags in total.

SWJTU POS tagging set subdivides Chinese name into Chinese surname(nr1), Chinese first name(nr2), Japanese name(nrj). Meanwhile, the organization name also is subdivided into corporation name(ntc), government agency(ntc), university name(ntu) and middle school or primary school name(nts) etc. The subdividing is helpful for entity recognition, text classification, knowledge extraction etc.

4 Dictionary Expansion and POS Subdivision

Based on the original SWJTU segmentation dictionary, this paper researched how to extend this dictionary, and meanwhile explored how to subdivide the POS of some words.

We extracted the words from online Encyclopedia and other knowledge Web sites, such as HowNet, CNKI, Human Cube, PIINGSHI Site and so on, to extend the original dictionary. In order to extend the dictionary more easily, we set different dictionary storage format, and the words can be appended POS or not. The word frequency statistic varies in different data sets, and the word and word frequency statistic is helpful for entity recognition. In Addition, word frequency for ambiguity words needs to be sorted, which means that the POS with higher frequency must be placed ahead of the POS with lower frequency. The dictionary format of the dictionary is shown in Table 1.

Table 1: SWJTU Segmentation Dictionary Format

Format	Example
Word Word Frequency POS1: POS1 Fequency1, POS2: POS2 Frequency2	特聘|distinguished employment[1] 222 vn:17605,v:9098
Word Word Frequency POS1, POS2	苏丹| Sudan 21811 nsf,nrf,nmc
Word Word Frequency POS: POS Frequency	国家机关| government 21499 n:21472
Word Word Frequency POS	隋唐|sui and tang dynasties 21468 t
Word Word Frequency	开放分类|open catetory 22332

The data downloaded from online Encyclopedia and other sites contain entry names and entry tags (entry category information). According to entry tags, entries can be classified. For example, the tags of entry "Tsinghua university" are "学校|school/大学|university","组织机构|organization", "教育|education", "高等教育|senior education", "大学|college" etc. Thus, we can extract all entries which are Universities according to the "大学|university" tag. We can achieve the extension of dictionary and POS subdivision. The specific rules are as follows:

(1) The organization nouns are subdivided into corporation (ntc), government agency (nto), university (ntu), middle school or primary school (nts) etc. In the original dictionary, all proper nouns are tagged as "nz".

(2) The academic nouns are subdivided into: mathematics (gm), physics (gp), chemistry (gc), biology (gb)

[1] The words following the symbol "|" are translation of Chinese characters.

etc.

(3) The people names (nr) are subdivided into Chinese name (nr2), Japanese name (nrj), transliteration name (nrf).

(4) The POS tag "nn"[2] is a job relevant POS, and is subdivided the tag into professional titles (nnt) and occupation (nnd).

(5) The POS tag "ni" is an organization relevant POS, and is subdivided into second-level institutional (nit) and institutional suffix (nis).

(6) The POS tag "nb" is a biological name relevant POS, and the tag is subdivided into animal names (nba) and plant names (nbp).

(7) The POS tag "nh" is a medical health relevant POS, and is subdivided into drugs (nhm) and disease (nhd).

(8) The POS tag "nm" is an item relevant POS, and the tag is subdivided into second-level chemical name (nmc).

There are about 7,375,803 entities in the extended dictionary, including 22,929 ambiguity words and 532,874 non-ambiguity words. The total of word frequency statistic is up to 7 billion times. The SWJTU segmentation dictionary can cover most of Chinese words, which is helpful for POS tagging.

5 Dictionary Reprocess and Modification

People's Daily corpus is labeled manually, so it is an important standard for researching Chinese words segmentation and POS tagging. We firstly statistic the ambiguity words and non-ambiguity words in People's Daily, and then we use them to revise the POS tag in SWJTU segmentation dictionary. The procedure contributes to POS tagging accuracy improvement. The statistical result of the ambiguity words and non-ambiguity words in People's Daily is shown in Table 2.

Table 2: Ambiguity Words and Non-ambiguity Statistic

Dictionary	Ambiguity	Non-ambiguity	Sum
People's Daily	4555	51456	56011
Segmentation dictionary	22929	5352874	5375803

It can be found from Table 2 that the non-ambiguity words in the established dictionary are up to more than 500 million. Therefore, the accuracy of non-ambiguity words POS tagging plays a decisive role in the POS tagging accuracy. If the accuracy of non-ambiguity words is improved, the POS tagging will gain great improvement.

There are 46 tags in People's Daily tagging set and 126 tags in SWJTU tagging set. And there are about seven tags in People's Daily tagging set never occur in SWJTU tagging set because the two tagging sets are different. The seven POS tags are shown in Table 3 as follows.

Table 3: POS Tagging Set Differences Comparison

Tags	POS description	Example
Bg	distinction morpheme	次|times/Bg
Dg	adverb morpheme	痛|pain/ Dg
g	morephere	No use in People's Daily
Mq	numeral morephere	甲|first/ Mq
Qg	measure word morephere	No use in People's Daily
Ug	auxiliary morephere	No use in People's Daily
Yg	modal morephere	耳|ear/ Yg

The remain 39 POS tags in People's Daily tagging set can be found in the SWJTU POS tagging set, thus we can correct the SWJTU segmentation dictionary according to People's Daily news data. The method of correcting the POS tag is shown in Table 4.

Table 4: Dictionary and POS Correcting Approach

People's Daily	Segmentation dictionary	Frequency	Proportion	Modification
Non-ambiguity Word	non-ambiguity word	32148	62.47%	replace
Non-ambiguity Word	ambiguity word	19308	37.52%	append
Ambiguity Word	ambiguity word	3785	83.09%	append
Ambiguity Word	non-ambiguity word	770	16.9%	append

The frequency is the occurrence time of words in both People's Daily and the segmentation dictionary. The proportion equals to the frequency divided by the total number of words in People's Daily. The total number of words in People's Daily is 56,011 in which the number of non-ambiguity words is 51,456 and the number of ambiguity words is 4,555.

Modification methods include "replace" and "append". "Replace" is to replace the POS in the dictionary using the POS in People's Daily set only when the words are non-ambiguity both in the dictionary and People's Daily. Because the POS tag of word in People's Daily is tagged manually and more accurate than the tag in SWJTU segmentation dictionary. In other cases, we add the POS tag following the last tag of word.

[2] SWJTU POS tagging set is introduced on the website "ics.swjtu.edu.cn."

6 POS Tagging Based on Statistical Machine Learning and Dictionary

Ambiguity word POS tagging is the key problem. Ambiguity word is a word that in the different context has different POS. For example, "改善|improvement", "提高|increase", "讲话|speak" in some contexts are verbs, but in some contexts are gerunds. Non-ambiguity word refers to a word that has only one POS in any contexts.

In this paper, the segmentation dictionary is used for non-ambiguity words POS tagging, and the statistical machine learning method is used for ambiguity words POS tagging. Meanwhile, the statistical machine learning method used in this paper is the maximum entropy (ME) POS tagging algorithm of Stanford parser.

We firstly use the dictionary to tag the words in sentences and then use the ME model to tag the words so that we can get two different tagging results for the same sentence. Next, we need to find out the word with different tags in sentences. If the word is a non-ambiguity word, we label the word using the tag in the dictionary which does not take the context information into account. If the word is an ambiguity word, we label the ambiguity word using all possible POS. Then we use ME model of Stanford parser to choose the best tag. The model considers the contexts, so the combination between dictionary and ME algorithm can improve the accuracy of POS tag. The POS tagging uses Penn Treebank tagging set and it needs to establish the corresponding relationship with SWJTU segmentation POS tag set because the two sets are different.

POS tagging is done after segmentation. The words in sentences are separated and there are blank spaces between each word. Blank space represents the segmentation symbol of words. After that, the POS tagging will be done. The process of POS tagging flow is as follows:

(1) Tag the words using the segmentation dictionary. The ambiguity words contains all its tags occurred in People's Daily news data, and the non-ambiguity has only one POS tag (in this step, ambiguity words has one result at least).

(2) Tag the words using ME model.

(3) Map the POS tag of words in sentences into SWJTU tag from ME tagging result.

(4) Compare the tagging results of step (1) with step (2) for ambiguity words. If the tag in the step (2) is one of the tag in the step (1), or the tag in the step (1) contains the tag in the step (2), we choose the tag in the step (2) as the final result.

(5) Process all ambiguity words in sentences according to step (4).

(6) Output the POS tagging results for the entire sentence.

7 Experiments

7.1 Longitudinal Comparison Experiments

This experiment is a comparison experiment about the different POS tagging algorithms.

Experiment 1: Use the ME model to tag the POS for the news data of People's Daily. For ambiguity words, if the results of ME tagging are different from the results of tagging which are based on dictionary, the results of ME are preserved. For non-ambiguity words, use original SWJTU segmentation dictionary to tag the POS of the words.

Experiment 2: Use the original SWJTU segmentation dictionary to tag the POS for the news data of People's Daily.

Experiment 3: Use the new SWJTU segmentation dictionary to tag the POS for the news data of People's Daily. In the dictionary the POS tags are revised using the approach illustrated in Table 4.

Experiment 4: Use the algorithm in the Chapter 6 and the new SWJTU segmentation dictionary. In the dictionary, the POS tags are replaced respectively according to the POS tags in the People's Daily.

Experiment 5: Use the algorithm in the Chapter 6 and the new SWJTU segmentation dictionary. In the dictionary, the POS tags in People's Daily are appended respectively to the tail of the tags in the original dictionary.

Experiment 6: Use the algorithm in the Chapter 6 and the new SWJTU segmentation dictionary. In the dictionary, the POS tag is replaced by the POS tag in People's Daily if the word is an ambiguity word in SWJTU original dictionary. Otherwise, the POS tag in People's Daily is appended to the tail of the current POS tag in SWJTU original dictionary.

Experiment 7: Use the algorithm in the Chapter 6 and the new SWJTU segmentation dictionary. In the dictionary, the POS tag is replaced by the POS tag in People's Daily if the word is an ambiguity word both in SWJTU original dictionary and People's Daily. Otherwise, the POS tag in People's Daily is appended to the tail of the current POS tag

in SWJTU original dictionary.

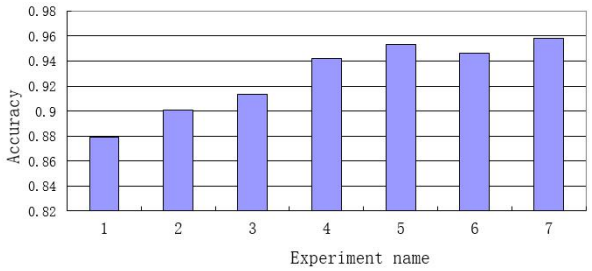

Fig. 1: Longitudinal Comparison Result for POS Tagging

Figure 1 shows that in experiment 1, the accuracy is 87.91%. In this experiment, ambiguity words depend on the results of ME POS tagging totally. In experiment 2, the accuracy is 90.06%. In this experiment, we only use existing POS tagging method based on SWJTU segmentation system. In experiment 3, the accuracy is 91.33% and we both use existing SWJTU segmentation system and ME algorithm.

In experiment 4, 5, 6, and 7, we employ different dictionary correction methods and the experiments show the results change according to different methods. In experiment 4, the accuracy is 85.13% when all POS tags are replaced by People's Daily POS tag. In experiment 5, the accuracy is 94.02% when People's Daily POS tags are appended to the tail of the existing POS tag. In experiment 6, the accuracy is 94.35% when the tags of ambiguity words are replaced by the POS tags in SWJTU segmentation dictionary, and non-ambiguity words are appended to tail of the current POS tag. In experiment 7, the accuracy is 95.80% when POS tag of the words are replaced by People's Daily POS tag if word is non-ambiguity words in both People's Daily and SWJTU dictionary, and in other conditions, POS tag of the word is appended to the tail of the current POS tag.

7.2 Horizontal Comparison Experiments

In horizontal comparison experiment, People's Daily news data is tagged with Stanford, Jieba and ICTCLAS tagging tool, and compared with the method proposed in this paper. Results are shown in Figure 2.

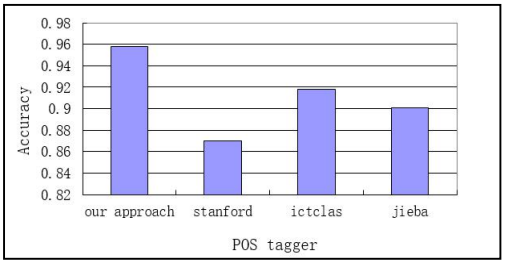

Fig. 2: Horizontal comparisons result for POS tagging

Shown in Figure 2, the accuracy of POS tag method presented in this paper is 95.8%, and the accuracy of Stanford is 87.01%, and the accuracy of ICTCLAS is 91.83%, and the accuracy of Jieba is 90.11%. This result indicates the POS tag method presented in this paper is better than other existing POS tagging method.

7.3 POS Tagging Comparison for Ambiguity Words

An experiment of POS tagging of ambiguity words is done, comparing the POS tag method presented in this paper with ICTCLAS2015, jieba online words segmentation system, Stanford Parser. Fifty sentences are selected which contain ambiguity words such as "提高|enhance", "改善|improvement", "跑步|run", "西安|xi'an" and so on. The experiment results are shown in Figure 3.

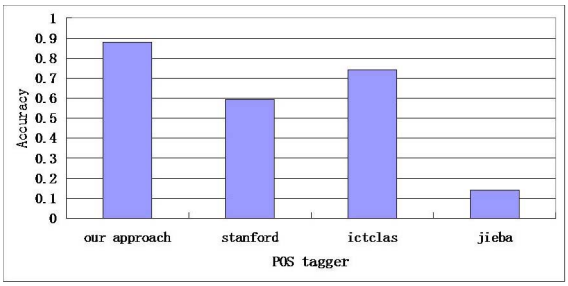

Fig. 3: POS tagging comparison for ambiguity words

As shown in Figure 3, the accuracy of our approach is 88%, and the accuracy of Stanford Parser, Ictclas and Jieba are 59.5%, 74%, 14% respectively. It shows that the performance of the method proposed in this paper is much better than other methods in ambiguity POS tagging testing,

8 Conclusions

In this paper, a POS tagging method is proposed based on statistical machine learning and SWJTU segmentation dictionary, and the method optimizes the POS tagging results of SWJTU Chinese words segmentation system. The accuracy achieves 95.80% when the method is tested in People's Daily January 1989 news data. The POS tagging algorithm approach is integrated into SWJTU Chinese words segmentation system. Compared with other POS tagging methods mentioned in this paper, the performance of the POS tagging method proposed in this paper is better, and the accuracy is 88% in ambiguity words POS tagging experiment. Additionally, our approach solves three problems as follows: 1) POS tag errors in the dictionary; 2) insufficient training corpus in the statistic machine learning method; 3) ignoring context information of words. Although

some POS tags of non-ambiguity words in SWJTU segmentation dictionary are amended, POS tags of many other words are still not corrected. The correction of POS tags of non-ambiguity words and choosing POS tags of ambiguity words are still the focus of the study in afterwards research.

References

[1] Kantor, P.: Foundations of statistical natural language processing. Information Retrieval, 4(1): 80-81, 2001.

[2] Li, H.D., Jia Z., Yin H F., et al.: Rule-based tagging method of Chinese ambiguity words. Journal of Computer Applications, 34(8):2197-2201, 2014.

[3] Hu, W.T., Yang, Y., Yin, H.F., et al.: Organization name recognition based on word frequency statistics. Application Research of Computers, 30(7):2014-2016, 2013.

[4] Yin, K., Yin, H.F., Yang, Y., et al.: Semantic similarity computation of Baidu encyclopedia entries based on SimRank. Journal of Shandong University (Engineering science), 44(3):29-34, 2014.

[5] Schmitz, S.: A note on sequential rule-based POS tagging. In: Proceedings of the 9th international workshop on finite state methods and natural language processing. Portland: Association for Computational Linguistics, 2011: 83-87.

[6] Saharia, N., Das D., Sharma U, et al.: POS tagger for Assamese text. In: Proceedings of the 47th annual meeting of the association for computational linguistics and the 4th international joint conference on natural language processing of the AFNLP. Singapore: Association for Computational Linguistics, 2009: 33-36.

[7] Zin, K.K., Thein, N.L.: POS tagging for Myanmar using hidden Markov model. In: Proc of international conference on the current trends in information technology. DC, USA: IEEE, 2009: 1-6.

[8] Moon, T., Erk, K., Baldridge, J.: Crouching dirichlet hidden Markov model: unsupervised POS tagging with context local tag generation. In: Proceedings of the 2010 conference on empirical methods in natural language processing. Uppsala, Sweden: Association for Computational Linguistics, 2006: 196-206.

[9] Yuan, L.C.: A part-of-speech tagging method based on improved hidden Markov model. Journal of Central South University (Science and Technology), 43(8): 3053-3057, 2012.(in Chinese)

[10] Liu, Y.F., Wang, Z.L., Wang C J.: Model of Chinese Words Segmentation and Part-of-Word Tagging. Computer Engineering, 36(4): 17-19, 2010.(in Chinese)

[11] Song, H.Y, Yao, T.F. Active learning based corpus annotation. In: Proceedings of IPS-SIGHAN joint conference on Chinese language processing. 2010: 28-29.

[12] Huang, D., Jiao, S.D., Zhou H W.: Dual-Layer CRFs Based on Subword for Chinese Word Segmentation. Journal of Computer Research and Development, 47(5): 962-968, 2010.(in Chinese)

[13] Hong, M.C., Zhang, K., Tang, J., et al.: A Chinese Part-of-speech Tagging Approach Using Condition al Random Fields . Computer Science, 33(10): 148-155, 2007.(in Chinese)

[14] Sun, J., Li, J.H., Zhou, G.D.: An Unsupervised Chines Part-of-speech Tagging Approach Using Conditional Random Fields. Computer Applications and Software, 28(4):21-23, 2011.(in Chinese)

[15] Brill, E.: Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Computational Linguistics, 21(4):543-565, 1995.

[16] Wang, G.Z., Wang X F.: A Method of POS Tagging Based on Priority of Rules. Journal of Anhui University of Technology(Science and Technology), 25(4):426- 429, 2008. (in Chinese)

[17] Jiang, S.P., Cheng, Q.X.: Study on Japanese Word Segmentation and P OS Tagging Based on Rules and Statistic. Journal of Chinese Information Processing, 24(1): 117-122, 2010(in Chinese)

[18] Liu, J.T., Song, Y., Xia, F.: The Construction of A Segmented and Part-of-speech Tagged Archaic Chinese Corpus: A Case Study on Huaiananz. Journal of Chinese Information Processing, 27(6): 6-15, 2013. (in Chinese)

[19] Yu, S.W., Duan, H.M., Zhu, X.F, et al.: The Basic Processing of Contemporary Chinese Coupus at Peking University SPECIFICATION. Journal of Chinese Information Processing, 2(05): 49-64, 2002. (in Chinese)

[20] Xia, T., Fan, X.Z., Lin, L: Java Call for ICTCLAS by JNI [J]. Journal of Computer Applications, 24(12):177-182, 2004.(in Chinese)