Part-of-Speech Tagging Based on Maximum Entropy

Hongdan Zhao, Jiangde Yu

Abstract—Part-of-speech is a fundamental step in natural language processing. This paper presents a part-of speech tagging method base on Maximum entropy. The proposed method is made up of three steps, that is, (1) Designing the context feature. (2)Training process and (3) Tagging process. Maximum Entropy estimation is able to compute Probability Density Function of the random variables, and in this paper, we solve the problem of tagging part of speech by tackling an optimization problem using maximum entropy. Closed evaluations were performed on PKU, NCC and CTB corpus from Bakeoff 2007. Experimental results showed that the context feature window including 3 words was better, and using single-word feature set were appropriate for Chinese part of speech tagging.

Keywords—Component Merchant, Formatting Analysis, Style Arch, Insert Styling, Part-of-Speech Tagging, Maximum Entropy, Context Windows, Context Feature

I. INTRODUCTION

PART-of-Speech (POS) tagging is the process of classifying and labeling words in a sentence according to their grammatical categories, i.e., verbs, nouns, particles, … etc.[1]. It is considered as an important step in many Natural Language Processing (NLP) implementations[2] as it deliver a layer of abstraction over the vast variances of the lexical, syntactic and semantic content of natural language. The input to a tagging algorithm is a string of words and a specified tag set of the kind described. The output is a single best tag for each word. For examples, "I eat an apple". The tagging result is "*I/ pronoun eat/ verb an/ quantity apple/ noun*". The difficulty of POS tagging is caused by multi-tagging words. The multi-tagging words mean that a word has many tagging. This one is almost universal. For humans, it is easy to distinguish. But for computers, it is ambiguous. That is it has more than one possible usage and part-of speech. How to solve the problem is the important difficulty that currently the part-of-speech tagging facing. In many natural language processing tasks, such as information retrieval, information extraction, text classification, machine translation, in order to achieve the better result, it is all dependents on the results of POS tagging.

The methods commonly used in POS tagging are divided into the following categories: The first is a rule-based methods [5], such as Transformation Based Learner (TBL) method[8], Statistical Decision Tree (SDT) method; The second is based on statistical methods, such as Hidden Markov Model (HMM)[7], the Maximum Entropy Model (ME), and Support Vector Machine model (SVM)[9] and Conditional Random Field Model (CRF). The rule-based methods had poor adaptation and could not give the probability of each possible classification results. So it is not used for component parts of bigger probability model. In the statistical methods, the HMM and SVM had better tagging effects, but because of the insufficient prediction information, so it has a great influence on tagging precision, especially for out of vocabulary (OOV). When using the CRF building model, the feature template may extend billions of context features. That will need more training time, and may make some CRF SDK unable run. The Maximum entropy could effectively use the context information, if the constraints conditions are satisfied. The model could be consistent with the probability distribution of training data. Especially for OOV, because of the context information, could get better tagging effect. Chen[3] proposed An English POS Tagging Approach Based on Maximum Entropy, but the result is not very good. And Kardan improved the method[6]. But that still have some problems. Singh [4] proposed a method that The Part of Speech Tagging of Marathi text using trigram method.

This paper describes a Chinese part-of-speech tagging system based on maximum entropy model and presents the influence on the training model size and tagging accuracy after considering the contextual features. Firstly, it introduces the basic principles when modeling, and then the feature templates used in the modeling process were analyzed. Finally, closed evaluations were performed on PKU, NCC and CTB corpus from Bakeoff2007. Experimental results showed that the feature window including 3 words was better, and using single-word feature collection is appropriate for Chinese pos tagging.

II. MAXIMUM ENTROPY POS TAGGING

A. Maximum Entropy Model

Maximum entropy model is a machine learning algorithm. The goal of statistical modeling is to construct a model that best accounts for some training data. More specific, for a given

This work was supported by Science and technology Innovative Research Team of Henan Educational Department of China under Grant 17IRTSTHN012, the Key Technology Projects of Henan Educational Department of China under Grant 16B520003, the National Natural Science Foundation of China under Grant 4100125, and supported by Program for Chang Jiang Scholars and Innovative Research Team in University.

Hong-dan ZHAO is with School of Computer and Information Engineering, Anyang Normal University, Anyang, China,e-mail:zhd.cn@126.com

Jiang-de Yu is with School of Computer and Information Engineering, Anyang Normal University, Anyang, China. e-mail: jiangde_yu@163.com

empirical probability distribution p_1, we want to build a model p as close to p_1 as possible. The model is more suitable to solve classification problems. When dealing with Chinese, such as word segmentation, part of speech tagging, syntax and semantic analysis, etc. These natural language problems can be formalized as a classification problem. The purpose is to estimate the probability of a class in the context. In Chinese context, x content can include Chinese characters, words, part of speech. For different tasks, the context selection is also different. The method of this kind of problem can be used to deal with statistical modeling. The first is collecting a large number of training sample, the sample represents the task knowledge and information. Sample quality determines the degree of completeness of knowledge. And then set up a statistical model, and the sample knowledge with the model, predict the future behavior of stochastic process.

In the Chinese POS tagging task, (x, y) represents some training samples, y is the POS tag assigned to a word, and x represents the contextual information regarding the word in consideration, such as the surrounding words. For example, in a sentence.

$[w_1\ v]$
$[w_2\ n]$
$[w_3\ vl]$
$[w_4\ n]$
$[w_5\ ns]$
$[w_6\ n]$
$[w_7\ Ng]$
$[w_8\ b]$
$[w_9\ n]$
$[w_{10}\ nrf\,]$
$[w_{11}\ n]$

The w_i is the word, and the w_{i-1} or the w_{i+1} is the context. These could be the training samples. From the sample we could get the context information .The building block of the model will be a set of statistics of the training sample. This model can be used to predict the probability of POS tagging. And the model of the distribution and training corpus probability empirical probability distributions should match. By the principle of maximum entropy can be shown that, x, y should be properly distributed to meet the maximum entropy models under conditions known constraints, that is the maximum entropy model, the general form of the formula 1 has the form.

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x, y)\right] \qquad (1)$$

Where z(x) is defined as formula 2

$$Z(x) = \sum_{y} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x, y)\right] \qquad (2)$$

Z(x) is normalization factor to ensure that $\sum_y p(y|x) = 1$. $f_i(x,y)$ are known as feature functions, which the function value is 0 or 1. λ_i is a weighting parameter corresponds to the features. k is the number of feature function.

Given a sequence of Chinese words $\{x_1, x_2,\ldots x_n\}$ and tags $\{y_1, y_2,\ldots y_n\}$ as training data, define x_i as the history available when predicting y_i. Then was chosen the maximize of the likelihood of the training data.

This model also can be interpreted under the Maximum Entropy formalism, in which the goal is to maximize the entropy of a distribution subject to certain constraints. Here, the entropy of the distribution p is defined as formula 3:

$$H(p) = -\sum_{x,y} \tilde{p}(x) p(y \mid x) \log p(y \mid x) \qquad (3)$$

And the constraints are given by formula 4:

$$\sum_{x,y} \tilde{p}(x, y) f(x, y) = \sum_{x,y} \tilde{p}(x, y) p(y \mid x) f(x, y) \qquad (4)$$

The observed feature expectation is defined as formula 5:

$$\tilde{p}(f) = \sum_{x,y} \tilde{p}(x, y) f(x, y) \qquad (5)$$

And the model's feature expectation is formula 6:

$$p(f) = \sum_{x,y} \tilde{p}(x, y) p(y \mid x) f(x, y) \qquad (6)$$

It can be shown (Darroch and Ratcliff, 1972) that if p has the form (1) and satisfies the constraints, it uniquely maximizes the entropy H(p) over distributions that satisfy , and uniquely maximizes the likelihood over distributions of the form (1) The model parameters for the distribution p are obtained via Generalized Iterative Scaling(Darroch and Ratcliff, 1972).

B.Contextual Features

To achieve a successful mode for any task by using the maximum entropy model, an important step is to select a set of useful features for the task. In the following, the feature sets used in the tasks are discussed. Generally, the context which is selected is based on a certain range, which is in the current word around. The range is called "context window". This window represents that when conducted POS tagging, the context range size. Figure 1 shows that the possible context window when POS tagging.

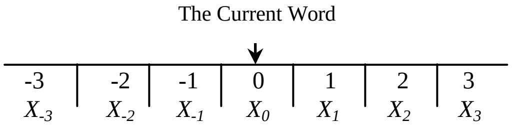

Figure 1. CONTEXT WINDOW

If using two words before and after each word as the range of the current context, the context of the range can be considered 5 word windows. If only one word, the context of the range can be considered 3 word windows. In the paper, it is According to the current word in the text sequence and its context to determine the word's tag. So in the feature sets, it mainly contains words that appear before and after the current word, the word string and other language elements. The whole features are divided into 10 categories.

TABLE I. FEATURE DENOTATION

Feature	Feature denotation
X_{-2}	The second word to the left
X_{-1}	The previous word
X_0	The current word
X_1	The next word
X_2	The second word to the right of the current word
$X_{-2}X_{-1}$	The combination of X_{-2} and X_{-1}
$X_{-1}X_0$	The combination of X_{-1} and X_0
X_0X_1	The combination of X_0 and X_1
X_1X_2	The combination of X_1 and X_2
$X_{-1}X_1$	The combination of X_{-1} and X_1
$T_{-1}T_0$	part-of-speech tag assign to the word X_0 and X_{-1}

In table I represents that the feature template used in our part-of-speech tagging. "X_n" is on behalf of the current word or the current word apart from the several words. For example, X_0 is the current word. X_{-1} is the previous word. X_1 is the next word. And so on. In the table, the last feature $T_{-1}T_0$ template is used for part-of-speech tag assign to the word x_0 and X_{-1}.

When generating feature, scan each word in the text, Each template in the template library circulation,, Each information function in the template get the value from the word context. Each information function values combined to get feature premise, get the feature action by mark the words, so as to obtain feature.

The generator of feature algorithm as follows:

```
First:
   scan the corpus;
Second: loop templates.
   The current template matching using start feature;
   If the generation feature is already exists in the library
   features: feature count plus one.
  Else Add new features into the feature library;
Third: repeat the first and second step.
```

C. Estimation Algorithm

The POS tagging Based on the maximum entropy model generation process is in fact as mentioned above develop templates, training of Idioms. Matching the generated and selected feature set, feature set using the maximum entropy parameter estimation algorithm to generate the tagging model, used for tagging.

Algorithm is described as follows：

First: from the text at the beginning start scanning;

Second: circular matching template generation characteristics, add roughing feature set;

Third: back to update the current word for each word, the current is executive the second step

Forth: circular matching template generation characteristics, add roughing feature set;

Fifth: for roughing select features, generate the selected feature set into Sixth;

Sixth: using the GIS algorithm to estimate the selected feature set parameters, access to part of speech tagging model;

By using the maximum entropy model training, the parameters of feature sets can be obtained, each corresponding to the feature. To judge a word mark what is to look at the features of the current word context meet, According to the obtained parameters, the maximum marking probability as the part of speech tag of the current word marking probability as the current word part of speech tag.

III. Experiments

A.Experimental design

In the part of speech tagging, maximum entropy feature is generated by matching the feature template in the corpus. Feature template selection is a hard thing, to waste time and energy. If the feature templates all be enumerated. It is very big and difficult to complete. Considering the situation, On the basis of previous studies, we use the method of artificial selection, only consider some simple features, after comparing several rounds of adjustment, the final selection of the 6 different templates.

By combining different features, we have participated in multi-group experiment and investigated the influence of Feature combinations in different context window. The feature of different combination information is shown in table II.

TABLE II. FEATURE TEMPLATE SETS

No	The feature of different combination
1	$X_{-1},X_0,X_1,X_{-1}X_0,X_0X_1,X_{-1}X_1,T_{-1}T_0$
2	$X_{-1},X_0,X_1,T_{-1}T_0$
3	$X_{-1}X_0,X_0X_1,X_{-1}X_1,T_{-1}T_0$
4	$X_{-2},X_{-1},X_0,X_1,X_2,X_{-2}X_{-1},X_{-1}X_0,X_0X_1X_1X_2,X_{-1}X_1,T_{-1}T_0$
5	$X_{-2},X_{-1},X_0,X_1,X_2,T_{-1}T_0$
6	$X_{-1}X_0,X_0X_1,X_1X_2,X_{-1}X_1,T_{-1}T_0$

According to the number of words involved in feature template, the feature template is divided into single word feature templates and double word feature template, combined with the generator of feature algorithm and the feature template number 3, the training sample extend feature such as follow:

[v U$_{03}$-_B/w$_1$ U$_{04}$-w$_1$/w$_2$ U$_{05}$-_B/w$_2$ E]
[n U$_{03}$-w$_1$/w$_2$ U$_{04}$-w$_2$/w$_3$ U$_{05}$-w$_1$/w$_3$ v]
[vl U$_{03}$-w$_2$/w$_3$ U$_{04}$-w$_3$/w$_4$ U$_{05}$-w$_2$/w$_4$ n]
[n U$_{03}$-w$_3$/w$_4$ U$_{04}$-w$_4$/w$_5$ U$_{05}$-w$_3$/w$_5$ vl]
[ns U$_{03}$-w$_4$/w$_5$ U$_{04}$-w$_5$/w$_6$ U$_{05}$-w$_4$/w$_6$ n]
[n U$_{03}$-w$_5$/w$_6$ U$_{04}$-w$_6$/w$_7$ U$_{05}$-w$_5$/w$_7$ ns]
[Ng U$_{03}$-w$_6$/w$_7$ U$_{04}$-w$_7$/w$_8$ U$_{05}$-w$_6$/w$_8$ n]
[b U$_{03}$-w$_7$/w$_8$ U$_{04}$-w$_8$/w$_9$ U$_{05}$-w$_7$/w$_9$ Ng]
[n U$_{03}$-w$_8$/w$_9$ U$_{04}$-w$_9$/w$_{10}$ U$_{05}$-w$_8$/w$_{10}$ b]

In order to feature template on the Chinese part of speech tagging in recognize that there is a "quantity", this research carries on the quantitative analysis from multiple angles And the design of relevant experimental. Table 2 lists several groups used in the experiment Feature template set. Among them, the serial number 1 to 3 feature template set is five word window template set, 4 to 6 sets of feature template is three word window template set. In the feature template only a single word feature template concentration set and only double word combinations constitute a set. In addition, the suffix "Single" and "Double" respectively.

For comparison, this paper follows the Bakeoff closed track rules, that the test model. Focus on learning from the corresponding training corpus tagging knowledge. The three training corpus we received for the Chinese part-of-speech tagging task include the PKU, NCC, CTB of the Bake off 2007. The training corpus size and testing corpus size is shown in table 3. When using the maximum toolkit training the model, it must be format the corpus suitable for the toolkit.

TABLE III. TRAINING CORPUS SIZE

Corpus	Training Corpus Size	Testing Corpus Size
PKU	8377KB	1976KB
NCC	3680KB	911KB
CTB	4995KB	1235KB

B. Evaluation

When evaluate the performance of Chinese POS tagging, commonly used evaluation indicators: tagging accuracy. Tagging accuracy represents that in the words of all the parts of speech tagging, the correct word part of speech tagging the share ratio. The formula 7 is as follows.

$$\text{Tagging accuray} = \frac{\textit{the correct word of tagging}}{\textit{all the parts of the tagging}} \quad (7)$$

C. Results

The experiment used the feature set that listed in Table 2. For the combination of features from No 1 to No 6, which were carried out for training, the training process of recording data is shown in table 4. This research designed two experiments, from different angles. First the model training process reflects the "quantity" attribute. Pay attention to this group of experiments are different. The feature window and feature template set influence on model training, the main extend the number of features, model of training time, training from a different set of templates and the model size of several "quantity" of the index to investigate. Second. Feature of different size of window opening effect on the performance of Chinese part of speech tagging, And different feature template set of Chinese part of speech tagging performance influence. This experimental group is concerned about the feature window and different characteristic modes of different sizes.

TABLE Ⅳ. TRAINING PROCESS OF RECORDING DATA

	PKU			CTB			NCC		
NO.	Feature number	Model size	Training time	Feature number	Model size	Training time	Feature number	Model size	Training time
1	1418810	70632KB	1276s	959057	45617KB	343s	898572	42341KB	401s
2	147600	11508KB	1198s	115045	7372KB	292s	117933	7382KB	351s
3	1271314	57712KB	1142s	844048	37576KB	277s	780698	34127KB	344s
4	2295484	124868KB	1909s	1563311	78912KB	501s	1476224	73564KB	532s
5	245929	22660KB	1740s	191718	13766KB	421s	196514	13770KB	460s
6	2049658	99633KB	1397s	1371630	63928KB	363s	1279768	58205KB	399s

Experimental data shows that:

(1) The feature number generated during model training proportional to the training time of the model. The feature is bigger. The training time needs more long. Because of the PKU corpus having more part of speech tagging sets, so the training corpus has more feature number.

(2) There is no necessary correlation between the feature number and model.

(3) There is much more feature number that the double-word feature collection generated than the single-word feature collection generated.

The first group of experiments using all 6 groups of feature template set respectively in three corpuses for Chinese part of speech tagging training. Then get the model. The second group of experiments is to using these models to test corpus. For part of speech tagging, Mainly compares the feature window is set to "5 word window" and "3 word window" of the part of speech tagging. The experiment number 1, 2, 3 feature template set is based on the 5 word feature window and Serial number for 4, 5, 6 feature template set is based on the 3 word feature window. The second concern is the performance using different feature template set training model for part of speech tagging. Secondly mainly compared the tagging performance feature template feature template set single word and double word set. In Table 4 shows the performance of Chinese part of speech of the 6 groups of feature template set training model in the test corpus on the corresponding tagging.

The experiments test the different testing corpus using the training model. Finally, evaluation results are shown in table 5. The experiment result shows that:

TABLE V. THE SCORES OF DIFFERENT TRACKS

Features No	PKU	NCC	CTB
1	93.82%	91.13%	91.97%
2	93.98%	91.53%	91.92%
3	81.51%	73.48%	79.14%
4	92.91%	90.34%	91.64%
5	93.18%	90.64%	91.06%
6	79.61%	72.59%	77.99%

(1) It could get better accuracy using feature set No 1, which represents the context window is three, than using feature set No 4, which represents the context window is five.

(2) No matter how big the window, consisting of a single-word feature set such as No 2 and number 5 could get better accuracy than double-word feature.

Overall, the maximum entropy model can achieve better accuracy.

IV. Conclusions And Further Work

The Chinese POS tagging is the foundation task in the natural language processing. It is also the foundation of the syntactic parsing and the chunk analysis. If there is some error when POS tagging, it will be enlarged and that should affect the results of further processing. So it has very important significance to the natural language processing. This paper presents a method to POS tagging for Chinese based on Maximum entropy and designs the context feature, also conducts some experiments. Experimental results showed that the feature window including 3 words was better, and using single-word feature collection is appropriate for Chinese POS tagging. The highest tagging accuracy can reach about 94%.

Our future work includes two aspects; How to improve recognition of OOV. In addition, Consider the impact of the introduction of more contextual information marked effect on.

References

[1] Fasha M, “A Proposed Adaptive Scheme for Arabic Part-of Speech Tagging” . International Journal of Advanced Computer Science & Applications, 2017, 8(7).

[2] R. A. Abumalloh, H. M. Al-Sarhan, O. Bin Ibrahim, and W. Abu-Ulbeh, “Arabic Part-of-Speech Tagging” J. Soft Comput. Decis. Support Syst.,vol. 3, no. 2, pp. 45–52, 2016.

[3] Chen Y, “An English POS Tagging Approach Based on Maximum Entropy” International Conference on Intelligent Transportation, Big Data and Smart City. IEEE, 2016:81-84.

[4] Singh J, Joshi N, Mathur I. “Part of Speech Tagging of Marathi Text Using Trigram Method” . International Journal of Advanced Information Technology, 2013(2).

[5] Peng T, Dai Y, Zhu F, et al. “Rule-Based Method for Unsupervised Part-of-Speech Tagging”. Journal of Jilin University, 2015.

[6] Kardan A A, Imani M B. “Improving Persian POS tagging using the maximum entropy model” Intelligent Systems. IEEE, 2014:1-5.

[7] Hu Chun-jing, Han Zhao-qiang, “Application Study of Hidden Markov Model Based Part-of-speech Tagging”. Computer Engineering and Applications, 2002, 38(6):62-64

[8] QU Gang, LU Ru-Zhan, “A Feature-Based Chinese POS Tagging Model”. Journal of Computer Research and Development. [J].2003,40(4):556-561

[9] Jesús Giménez and Lluís Márquez. “SVMTool: A general POS tagger generator based on Support Vector Machines”. In proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. 2004: 43-46

Hongdan ZHAO was born on Sep. 15, 1982. The birth place is AN Yang city, HE Nan province, China, birth date: Educational background: 1997-2000 HUANG He Science and Technology College , major in Computer and Application, Bachelor; 2004-2007 ZHENG Zhou University, major in computer application Technology, Master; major field of study: natural language processing.

Jiangde Yu was born in 1971, he graduated from Beijing institute of technology, associate professor, mainly engaged in natural language processing, information extraction, text data mining, etc.