A Part-Of-Speech Tagging Approach for Chinese-Hmong Mixed Text

IOP Conf. Series: Materials Science and Engineering **864** (2020) 012064 doi:10.1088/1757-899X/864/1/012064

A Part-Of-Speech Tagging Approach for Chinese-Hmong Mixed Text

Hang-Cheng Li, Li-Ping Mo*, Kai Qing Zhou

College of Information Science & Engineering, Jishou University, Jishou, 416000, Hunan, China

Email : zmx89@jsu.edu.cn

Abstract. Part-of-speech (POS) tagging is a basic problem that needs to be solved in the informationization of Chinese-Hmong mixed text including square Hmong characters, so far no one has studied it. This paper proposes a POS tagging approach for Chinese-Hmong mixed text by utilizing improved Hidden Markov Model (HMM) to expand contextual information. The results of comparative experiments based on cross-validation reveal the proposed approach has perfect performance, and is able to obtain tagging results with good consistency and high coverage.

1. Introduction

POS tagging is a basic task in natural language processing, which plays an important role in removing word ambiguity, reducing query ambiguity, and improving search efficiency. The research of POS tagging technology is closely related to the construction of corpus. In the early 1960s, English POS tagging technology began to attract attention during the construction of the world's earliest machine-readable corpus, Brown Corpus [1]. English POS tagging technology was greatly developed in the 1980s and 1990s, and has matured after nearly 60 years of development. During this period, POS tagging technologies for Arabic [2], Uyghur [3] and other languages have developed rapidly. In recent years, research on POS tagging for low-resource languages [4] has also gradually been concerned. However, the research on POS tagging technology for Hmong language has not really started. This paper presents an automatic tagging method using an improved HMM to expand the context information for Chinese-Hmong mixed text, especially includes square Hmong characters.

The rest of this paper is organized as follows. Section 2 introduces HMM and its application in POS tagging. Section 3 depicts POS tagging method for Chinese-Hmong mixed text including square Hmong characters based on HMM. Section 4 shows the experimental results and analysis. Section 5 presents the conclusions.

2. HMM and its application in POS tagging

Traditional POS tagging include statistics-based methods, rule-based methods, and a combination of statistics and rules. Statistics-base methods abstract the POS tagging problem as a mathematical statistical model, calculate POS probability of a word in the context of the trained corpus, and tag the sentence using the tags sequence with the highest probability. The method based on HMM is representative of statistics-base methods, has been widely applied because of its ability to obtain tagging results with good consistency and high coverage.

According to [5], HMM is a time series probability model with two layers. One is an observation layer composed of the observation sequence to be identified, another is a hidden layer composed of a

finite automaton corresponding to the Markov process. HMM can be formally defined as a 4-tuple $\lambda=(S, O, \pi, A, B)$, where $S=\{S_1, S_2, ..., S_N\}$ is a finite set including N hidden states, $O=\{O_1, O_2, ..., O_M\}$ is a finite set including M observation symbols associated with hidden states, $\pi=\{\pi_i\}$ is the probability distribution of the initial state, $\pi_i=P(q_1=S_i)$ is the probability of selecting a hidden state $S_i(1\le i\le N)$ at the initial state q_1, $A=\{a_{ij}\}$ is the transition probability matrix of the hidden state, and $a_{ij}=P(q_t=S_j|q_{t-1}=S_i)$ $(1\le i, j\le N, 1\le t\le M)$ is the probability of the state q_t is S_j when state q_{t-1} is S_i, $B=\{b_i(O_k)\}$ is the emission probability matrix of the observation symbol, that is, the output probability matrix of the hidden state, and $b_i(k)=P(O_k \mid q_t=S_i)$ $(1\le i\le N, 1\le k\le M)$ is the probability that the symbol $O_k$can be observed when state $q_t$is S_i.

Suppose each word in the sentence to be tagged is an observation symbol, each POS in the tags sequence is regarded as a hidden state, and the correlation of the n-th POS just exists among the first n-1 (n>1) POSs, respectively. Then, HMM can be applied to POS tagging modeling for searching the optimal tags sequence based on the potential relationship between POS and its appearance probability as following steps. First, initialize S, O and π, let each tag in tag-set as a hidden state, and let each word in the sentence as an observation symbol. Second, train the HMM to obtain matrix A, and B. Final, execute the Viterbi algorithm on the trained HMM to obtain the best tags sequence of the sentence in the test text.

3. POS tagging for Chinese-Hmong mixed text based on improved HMM

3.1. Design of the square Hmong POS tag-set

Square Hmong characters are ideographic characters with a structure similar to Chinese characters, are usually mixed with Chinese characters, appear in the Chinese-Hmong songbook and script. An example of Chinese-Hmong mixed text including the square Hmong characters is shown in figure 1.

Figure 1. Chinese-Hmong mixed text including the square Hmong characters

According to the word formation, a square Hmong character represents a morpheme or word [6]. Usually, square Hmong words mainly including one character or two characters, and rare words including 3 characters or more. Considering that square Hmong words are determined by their corresponding Chinese meaning, the square Hmong POS tag-set was designed in [7] according to the *Modern Chinese Corpus Processing-Specifications and Manual of Lexical Segmentation and POS Tagging* which is edited by Institute of Computational Linguistics, Peking University, China.

3.2. Construction and improvement of HMM

Givena sentence of the Chinese-Hmong mixed text to be tagged is $W=W_1W_2\ldots W_n$, and let the tags string of the sentence to be output be $T=T_1T_2\ldots T_n$, POS tagging is the process of obtaining a sequence of POS state $T^*$with the maximum probability $P(T_{1,n}|W_{1,n})$ based on HMM. According to the Bayesian formula, T^* can be calculated by the following equation (1) in the traditional HMM.

$$T^* = argmaxP\left(T_{1,n} \mid W_{1,n}\right) = argmax\frac{P\left(W_{1,n} \mid T_{1,n}\right)P\left(T_{1,n}\right)}{P\left(W_{1,n}\right)} \tag{1}$$

Since $W_{1,n}$ is a given input and $P(W_{1,n})$ is a constant, Eq.(1) can be simplified as follows.

$$T^* = argmaxP\left(T_{1,n} \mid W_{1,n}\right) = argmaxP\left(W_{1,n} \mid T_{1,n}\right)P\left(T_{1,n}\right) \tag{2}$$

If $P(T_{1,n}|W_{1,n})$is calculated using bigram-model $P(W_{1,n}|T_{1,n}) = \prod_{i=1,n} P(W_i|T_i)$, the above equation (2) can be further expressed as $T^* = argmax \prod_{i=1,n} P(W_i|T_i)P(T_i|T_{i-1})$, where $P(W_i|T_i)$, $P(T_i|T_{i-1})$ and π_i are all calculated using the maximum likelihood estimates shown in equation (3).

$$P(T_i|T_{i-1}) = \frac{Num(T_i,T_{i-1})}{Num(T_{i-1})}, \quad P(W_i|T_i) = \frac{Num(W_i,T_i)}{Num(T_i)}, \quad \pi_i = P(q_1 = T_i) = \frac{Num(q_1 = T_i)}{Num(q_1)} \tag{3}$$

Where $Num(T_i,T_{i-1})$ is the number of simultaneous occurrences of T_i and T_{i-1}, $Num(W_i,T_i)$ is the number of W_i is tagged as T_i, $Num(T_i)$ is the number of occurrences of T_i in the training corpus, $Num(q_1=T_i)$is the number of T_i appears as the first POS in a sentence of the training corpus, and $Num(q_1)$ is the number of sentences in the training corpus.

In the traditional HMM, only the current T_i is considered in calculating the emission probability of the current word W_i. However, it is necessary to consider the influence of context on words and POS when POS tagging in real corpus. Therefore, we improve the model above so that the emission probability of the current word W_i depends not only on the current T_i, but also on the subsequent T_{i+1}. As a result, an improved model denoted as $T^* = argmax \prod_{i=1,M} P(W_i|T_i,T_{i+1})P(T_i|T_{i-1})$, where $P(W_i|T_i,T_{i+1}) = \frac{Num(W_i,T_i,T_{i+1})}{Num(T_i,T_{i+1})}$ if let $Num(W_i,T_i,T_{i+1})$ be the number of occurrences when W_i is tagged as T_i and the subsequent is T_{i+1}.

3.3. Viterbi algorithm based on improved HMM

Viterbi algorithm is often applied in HMM to search the optimal state sequence.POS tagging for Chinese-Hmong mixed text based on improved HMM is the process of searching the best POS tag sequence. So the use of Viterbi algorithm on the improved HMMis feasible. Let $\delta_i(j)$ represent the maximum probability that HMM will reach the state S_j and output the words $O_1O_2\ldots O_i$ along a certain path, and let$\psi_i(j)$ record the previous state on the path at time i-1 before the maximum probability to reach. Suppose that the total number of tags in the POS tag-set is N, and the mixed text to be tagged includes M words, Viterbi algorithm can be described as follows.

Step 1. Initialize. Let $\delta_1(j) = \pi_j b_j(O_1)$and $\psi_1(j)=0$.

Step 2. Starting from i=2, use $\delta_i(k) = \max_{1\le j\le N} \delta_{i-1}(j)a_{jk}b_k(O_i), \quad (2 \le i \le M;\ 1 \le k \le N)$ and $\psi_i(k) = \arg\max_{1\le j\le N} \delta_{i-1}(j)a_{jk}b_k(O_i), \quad (2 \le i \le M;\ 1 \le k \le N)$ recursively calculate $\delta_i(k)$ and $\psi_i(k)$. Suppose that T_k is the k-th tag, use $\delta_i(T_k) = \max_{1\le j\le N}[\delta_{i-1}(T_j) \times \frac{P(T_k|W_i) \times P(T_k|T_j)}{P(T_k|W_{i-1})}], \ (2 \le i \le M;\ 1 \le k \le N)$ and $\psi_i(T_k) = \arg\max_{1\le j\le N}[\delta_{i-1}(T_j) \times \frac{P(T_k|W_i) \times P(T_k|T_j)}{P(T_k|W_{i-1})}], \ (2 \le i \le M;\ 1 \le k \le N)$ recursively calculate $\delta_i(T_k)$ and $\psi_i(T_k)$.

Step 3. Recursion ends with i=M. At this time, let $O_M^* = \arg\max_{1\le j\le N}[\delta_M(T_j)]$, $P(O_M^*) = \max_{1\le j\le N}[\delta_M(T_j)]$.

Step 4. Backtrack the path to obtain the optimal POS tag sequence q^*, and each q^*_i is given as $q^*_i = \psi_{i+1}(q^*_{i+1}), \ (i = M-1, M-2, ..., 1)$.

Obviously, the complexity of this algorithm mainly depends on the number of tags N in the tag-sets and the number of words M in the text to be tagged. In the worst case, due to the bidirectional dependence is considered in the improved HMM, M^3 paths need to be processed when scanning the current word, the time complexity of the algorithm is O(N*M^3).

4. Experiment and analysis

Comparative experiments between the presented approach and the basis method using traditional HMM were conducted on desktop PC with Intel (R) Core (TM) i5-3470 CPU @ 3.20GHz, 4G memory, Win7 operating system andPython3.0. Most of the experimental corpora came from the *People's Daily marked corpus in January 1998*, and a small amount came from the hand-tagged Chinese-Hmong songbook and script. The number of tags in Chinese corpus tag-set and Hmong corpus tag-set is 39 and 14, respectively. Considering that the experimental data are insufficient due to the lack of Hmong corpus, we conducted four 4-cross-validation experiments. Three quarters of the corpus were used as the training set and one quarter was used as the test set. Three indicators, the accuracy P, recall rate R and F_1 shown in the following equation (4) were used to evaluate the performance of the two models.

$$P=\frac{N_t}{N_r}\times 100\%,\qquad R=\frac{N_t}{N_a}\times 100\%,\qquad F_1=\frac{2\times P\times R}{P+R}\times 100\% \tag{4}$$

Where N_t is the number of correctly tagged words, $N_r$is the total number of words recognized, N_a is the total number of words.

Comparative experimental results for corpus training sets with different word levels show that, first, as the size of the training set increases from 50000 to 200000, the three indicators P, R and $F_1$of the improved HMM increased by 12.6106%, 12.6965%, and 12.6587%, respectively, but that of traditional HMM increased by 9.1356%, 9.902%, and 9.1566%, respectively; second, as the size of the training set increases from 200000 to 800000, the three indicators P, R and $F_1$of the improved HMM increased by 2.1401%, 2.4792%, and 2.3106%, respectively, but that of traditional HMM increased by 6.1958%, 6.3801% and 6.2890% only, respectively. Based on the experimental results, the means of P, R and F_1 were calculated and shown in table 1.

Table 1. Comparison of POS tagging performance based on two models

Number of words	Mean of P (%)		Mean of R (%)		Mean of F_1 (%)	
	mproved HMM	Traditional HMM	Improved HMM	Traditional HMM	Improve HMM	Traditional HMM
50,000	79.0493	72.4946	78.3887	71.8858	78.7135	72.1865
200,000	89.0179	79.1174	88.3413	78.4203	88.6776	78.7667
800,000	90.9229	84.0193	90.5315	83.4243	90.7266	83.7204

As shown in table 1, in the case of training sets of the same size, the three indicators P, R and F_1 of the improved HMM are significantly higher than that of traditional HMM. On the other hand, the increasing of three indicators of the two models tends to decrease while the size of the training set grows. Furthermore, the three indicators P, R and $F_1$of improved HMM show better stability than that of traditional HMM when the training set reaches a certain size. Obviously, the proposed approach has more perfect performance than that of the traditional one, and is able to obtain tagging results with good consistency and high coverage in thePOS tagging of Chinese-Hmong mixed text.

5. Conclusion and future work

In this paper, we have shown that how to improve the HMM, and how to apply the improved HMM to realize the POS tagging for Chinese-Hmong mixed text including square Hmong characters. Our work laid a foundation for further research on Hmong speech recognition, information retrieval, machine translation and other technologies.

As future work, the study of optimizing the accuracy and speed of POS tagging for Chinese-Hmong mixed text using intelligent optimization algorithms such as harmony search algorithm will be conducted.

6. Acknowledge
This work was supported by the National Natural Science Foundation of Hunan Province (No.2019JJ40234), Research Foundation of Education Bureau of Hunan Province (No.19A414, No.18B317), Research Foundation of Special Project on Language Application of Hunan Provincial Language Committee (No.XYJ2019GB09), Research-based Study and Innovative Experimental Project for College Students in Hunan Province (No.20180599), and Research-based Study and Innovative Experimental Project for College Students in Jishou University (No. JDCX20180122).

References

[1] Christopher D M and Schutze H 1999 *Foundations of Statistical Natural Language Processing* (The MIT Press, Cambridge) pp 136-157

[2] Zeroual I, Lakhouaja A and Belahbib R 2017 Towards a standard Part of speech tag-set for the Arabic language *J.King Saud Univ-Comp and Infor. Sci.* **29** pp 171–178

[3] Muhetaer P, W. Silamu and Maimaitayifu 2019 Uyghur part of speech tagging method based on hybrid model *Computer Simulation* **35** pp 268-273

[4] Kim Y B, Snyder B and Sarikaya R 2015 Part-of-speech taggers for low-resource languages using CCA Feature *Proc 2015 Conf on Empirical Methods in NLP (Lisbon)* Association for Computational Linguistics pp 1292-1302

[5] Moon T, Erk K and Baldridge J 2010 Crouching Dirichlet, hidden markov model: unsupervised part-of-speech tagging with context local tag generation *Proc of the 2010 Conf on Empirical Methods in NLP (Cambridge,MA)* Association for Computational Linguistic pp 196–206

[6] Yang Z B and Luo H Y 2008 On the folk coinage of characters of the Miao people in Xiangxi area *J. Jishou University (social sciences edition)* **29** pp 130–134

[7] Zhou T, Mo L P, Zeng H, Zhi L, Wengyu Li and Ying W 2019 Design of the part-of-speech tag set for the square Hmong characters 2019 *J. Intell Comp and App* **9** pp 131-134