GENERATING DIGESTS FROM EDUCATIONAL ARTICLES AUTOMATICALLY BASED ON SECOND ORDER HMM

CanLi Wu
School of Informatics, Zhejiang Sci-Tech University

Lican Huang[1,2]
[1]School of Informatics, Zhejiang Sci-Tech University
[2]Hangzhou Domain Zones Technology Co., Ltd
Hangzhou, China
Huang_lican@yahoo.co.uk

Abstract—***Automatically generating summary of articles is very important when we encounter explosive reading information; computers can help people on text compression, extraction, representation and obtain core text content automatically. However, computer still encounters a lot of difficulties, for example, how to divide words from ambiguity, inaccuracies, redundancy of the lengthy article, and so on. This paper presents an improved Hidden Markov Model (HMM) Word segmentation method.***

Keywords—Summary generating; HMM; word segmentation

I. INTRODUCTION

For many years, the research of automatic summarization often appeared in the important journals or academic conferences. And much work has been done in theory and practice. It is often hard for educators and academics to find useful articles from huge amount of articles. This paper is dedicated to improve the model and algorithm of the generation of automatic summary of articles in education field and to increase the quality of summary, so that the reader can quickly find the in-depth study of the articles.

The purpose of the automatic abstract is to analyze, understand and process the contents of the network-like text automatically through the computer, and to generate a simple information representation which can express the original content. The general steps are divided into three ones: preprocessing and effective word segmentation, text compression and information classification, adjusting the experimental parameters and evaluating the quality of the results. Although the research of automatic summarization has developed, the word segmentation work is still plagued by researchers due to large redundancy, low accuracy and easy to produce ambiguity and so on.

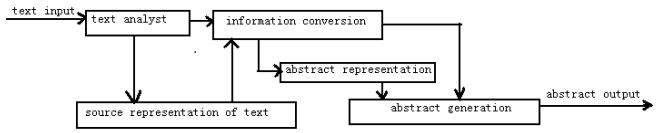

Fig.1 flow chart of automatic summary

Because the length of educational articles is too longer to read, so the pretreatment of the words and effect of the disposal of whole article are very important. This paper firstly will be contend to compare the advantages and shortcomings of various statistical models in common, then classify words on the proposed second-order HMM which will be improved on the basis of first-order HMM. However, the improvement of previous backward algorithm (Baum Welch) On Markov Chain Process will be aimed to handle the most critical problems of ambiguity processing that are low accuracy of words segmentation and unofficial operation of the stratified sampling data[1][2][3][4][5][6][7][8][9].

II. OVERVIEW OF *SECOND-ORDER HMM* AND BAUM-WELCH ALGORITHM

Hidden Markov model is aimed to determine the implicit parameter in observable parameters, and it is the most commonly statistical model to process the data. Therefore, it is also widely used in the preprocessing step of automatic summarization. But in the traditional first-order HMM, the current event probability is only related to the probability of the previous event, and the analysis of the meaning of words is not accurate enough. So this paper tries to use the second-order HMM model to enhance the ability of the computer's semantic understanding and the logic of the context and it tries to enhance the logic of the semantic understanding and context with second-order HMM model. The model is a novel mixture model based on the principle of maximum entropy and maximum adjacent information, which combined with multi-level corpus lexicon and can be determined in the hidden according to the first two moments of the state.

0122

Second-order HMM is a triple: initialize the probability vector (π_i), state transition matrix (a_{ij}) and the confusion matrix (b_{ij}). The state of the t+1 time is not only related to the t time, but also to the t-1 time:

$$a_{ij} = P(X_{t+1} = S_k \mid X_i = S_j, X_{t-1} = S_i, X_{i-2} = ...)$$
$$= P(X_{t+1} = S_k \mid X_i = S_j, X_{t-1} = S_i)$$

$$\sum_{k-1}^{N} a_{ij} = 1; a_{ij} \geq 0$$

N represents the total number of States.

The same probability of observation vector also depends on the time before the system state, i.e.:

$$b_{ij}(l) = P(y_t = v_t \mid X_t = S_j, X_{t-1} = S_i)$$

So $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ consist the parameters of second-order hidden MHH, and they representative the initial state distribution, the transfer of state distribution, and the observations of the probability distribution respectively.

The total length of observation sequence state is N, and we will use length k in a sliding window of segmentation to get the sub sequence set { X_i }; among them, $1 \leq i \leq L/k$.

The output probability of each sub sequence will be calculated:

$$\xi = \frac{candidate - subsequences}{total - subsequences}$$

Finally, ξ will be compared with a threshold value. If it will be less than the minimum likelihood value, the probability will meet the requirements, and it also can decide two state probabilities of subsequent hidden sequences with the storage method in modified algorithm.

Let $\sum_{t}^{\tau-1} \xi_t(i,j,k)$ is the entire observation sequence of state transition frequency; $\hat{\pi} = (n-1)$ **is** the state is the frequency of the following;

$$\hat{a}_{ij} = \hat{a}_{ijk} = \frac{\xi(i,j,k)}{\xi(i,j)};$$

$$\hat{b}_{ij}(l) \approx \frac{\sigma(i,j)}{\xi(i,j)};$$

Where $\xi(i,j)$ will be the state that transfer from the i state to j and $\sigma(i,j)$ will be not only the state that will transfer from the i state to j state, but also it will be the Y_t measuring frequency. In order to make the global maximum information entropy, the initial value must be asked to close to the maximum.

$$\hat{\pi}_i = \frac{P(y, x_j = i \mid \phi)}{P(y \mid \phi)} = \gamma_t(i)$$

$$\hat{a}_{ijk} = \frac{\sum_{t=2}^{\tau-1} P(y, x_{t-1} = i, x_t = j, x_{t+1} = k \mid \phi)}{\sum_{t=2}^{\tau-1} P(y, x_{t-1} = i, x_t = j, x_t = j \mid \phi)}$$

$$\approx \frac{\sum_{t=2}^{\tau-1} \xi_t(i,j,k)}{\sum_{t=2}^{\tau-1} \gamma_t(i,j)}$$

$$\hat{b}_{ij}(l) = \frac{\sum_{t=2}^{\tau} P(y, x_{t-1} = i, x_t = j \mid y, \phi)\, \delta(y_t, v_t)}{\sum_{t=2}^{\tau-1} P(x_{t-1} = i, x_t = i \mid y, \phi)}$$

$$= \frac{\sum_{t=2, y_t = v_k}^{\tau-1} \gamma_t(i,j)}{\sum_{t=2}^{\tau-1} \gamma_i(i,j)}$$

If $y_t = v_l$, then $\xi(y_t, v_t) = 1$, so the maximum information will be obtained if $y_t = v_l$.

III. THE IMPROVED ALGORITHM

Baum-Welch algorithm eventually find the maximum probability in the hidden state to determine the last node state under second-order HMM hypothesis, then forward in turn to derived results of each node according to the results of the last state, in order to obtain the effective segmentation of the whole article. But for the long and redundant content of educational articles, the traditional iterative algorithm requires a great deal of time. To improve computing speed, novel improved Baum Welch algorithm is presented, which can be combined with multi-layer recursive sampling when

0123

determining the iterative process; the system can save the data which is in a maximum probability of state, thus save the time back to traverse.

The sampling process will cause changes in sample size and quantity, so it should be adjusted according to changes in sampling factor. The transfer memory coefficient for sample x_k^{t-1} as the iterative sampling will be f_j^t / f_k^{t-1}, if memory size will be given, and the minimum first timestamp of S_j^{t-1} will be set to t_1 and maximum last timestamp of S_j^{t-1} will be set to t_2 in multi-level recursive sampling process.

IV. EVALUATION SCHEME

In the mechanized assessment of the quality of automatic paper, the measurement factors are accuracy rate, recall rate, F value, compression rate, coverage, readability and logicality and coherence, etc. And these measures are the important factors that will affect the qualities of automatic summarizations directly. The accuracy and relevant of evaluation are important for the generation of abstract. We use the semantic similarity evaluation method. The main idea is similarity comparison with the one written by specialist to judge the quality, it can not only solve the problem of experts disagree, but also solve the partial doubts of the mechanized index evaluation. It calculates artificial number of sentences N_h in the abstract extraction, the mechanized number of sentences Nm and the overlapping sentences number of the two kinds of the extraction N_{hm}. Then it will obtain Boolean value F by accuracy R and recall rate P value. Now the artificial abstracts are generally adopted by the extraction by experts in the original. In order to avoid personal bias, we ask multiple experts for the same article, and take the majority opinion collections.

V. CONCLUSIONS

This paper puts forward the use of the second order HMM, and makes the probability analysis of the meaning of the words by the former state of a word to decide to enhance the accuracy of the segmentation. The improved Baum Welch algorithm uses the function of Evaluation and analysis to record the results of the corresponding implied condition and the maximum probability on HMM. So we do not need to do back traverse for sequence statistics and accelerate the speed of execution.

ACKNOWLEDGMENT

The paper is supported by the project “Hangzhou Qinglan Plan--scientific and technological creation and development (No.20131831K99” of Hangzhou scientific and technological committee. The software copyrights is owned by Hangzhou Domain Zones Technology Co., Ltd., and Chinese patent applied is owned by Hangzhou Domain Zones Technology Company.

REFERENCES

[1] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, et al. Brain–computer interfaces for communication and control[J]. Clinical Neurophysiology. 2002,113(6):767-791.

[2] K. Talbot, M. Turner, R. Marsden. Motor neuron disease: a practical manual[M]. Oxford University Press. 2010.

[3] D. Regan, Human Brain Electrophysiology: Evoked potentials and evoked magnetic fields in science and medicine[J]. New York Elsevier. 1989:380-420.

[4] Liu J, Li Z, Guan L, et al. A Novel Parameter-Tuned Stochastic Resonator for Binary PAM Signal Processing at Low SNR[J]. IEEE Communications Letters.2014,18(3):427:430.

[5] S. Lu, Q. He, F. Hu, et al. Sequential Multiscale Noise Tuning Stochastic Resonance for Train Bearing Fault Diagnosis in an Embedded System. IEEE Transaction on instrumentation and measurement[J]. 2014,63(1):106-116.

[6] J. Vidal. Toword Direct Brain-Computer Communication [J]. Annual Review of Biophysics and Bioengineering, 1973,2(1):157-180.

[7] L. A. Farwell, E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials[J]. ElectroencephalogrClinNeurophysiol, 1988, 70(6): 510-523.

[8] W. Speier, C. Arnold, J. Lu, et al. Natural language processing with dynamic classification improves P300 speller accuracy and bit rate[J]. J Neural Eng, 2012, 9(1): 016004.

[9] R. Fazel-Rezai, B. Z. Allison, C. Guger, et al. P300 brain computer interface: current challenges and emerging trends[J]. Front Neuroeng, 2012, 5: 1-14.