Journal of Theoretical and Applied Information Technology
20th February 2013. Vol. 48 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HIDDEN MARKOV MODEL AND ITS APPLICATION IN NATURAL LANGUAGE PROCESSING

[1]XUEXIA GAO, [2]NAN ZHU
[1]Asstt Prof., College of Computer Engineering, Xinxiang University, Xinxiang,China
[2]Lecture, College of Computer Engineering, Xinxiang University, Xinxiang,China
E-mail: [1]xuexiagao123@126.com

ABSTRACT

This paper describes Hidden Markov Model and its application in natural language process, first introduces the basic concept of Hidden Markov Model, then introduces the three basic issues and the basic algorithm to solve the problems, finally gives the demonstration of the application of Chinese part-of-speech tagging and speech recognition via Hidden Markov Model.

Keywords: *HMM, Natural Language Processing, Chinese Part-Of-Speech Tagging*

1. INTRODUCTION

Hidden Markov Model is the mathematical statistical model which is used to describe the statistical characteristics of random process; it is developed by the Markov chain. For the actual problem is more complex than the description of Markov chain model, the observed events is not corresponding to the status one-to-one, but related through a set of probability distribution, that model is Hidden Markov Model. It is a double random process, one of which is the Markov Chain; it describes the status transition probability. What another random process described is the statistics relationship between status and observation value., The observers perceive the existence and characteristics of hidden status, to form "hidden" Markov Model through the observation value . Hidden Markov Model in the 1970s in speech recognition field attained great success, after that it is widely applied in each domain of natural language processing, which becomes the important method of natural language processing based on statistics; it is one of important achievements in last century in the field of statistics natural language processing.

Hidden Markov Model is a quintuple group: (Q_x, Q_o, A, B, π)

(1) The limited set of status:

$Q_x = \{q_1, q_2, q_3, q_4, \cdots, q_N\}$, which represents the model status (i.e. output), the number of status is N. For some practical applications, even the status is hidden, but each status of model is related with some physical meaning, meanwhile the status is connected with each other, and which can transfer from one status to another status. X_t is used to represent the status of t moment.

(2) The limited set of observation values: $Q_o = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \cdots, \sigma_M)$, the number of observations value corresponding to each status is M, observation value is corresponding to the actual output of system model. O_t is used to express the observation of time t.

(3) Transition probability: $A = \{a_{ij}\}$, $a_{ij} = P(X_{t+1} = q_j | X_t = q_i), 1 \le i, j \le N$. a_{ij} meets $a_{ij} \ge 0, \ \forall\ i, j$ and $\sum_i a_{ij} = 1, \ \forall\ i_o$.

(4) Output probability: $B = \{b_{ik}\}$ b_{ik} represents in the status q_i, the probability of v_k appears at t moment, that is $b_{ik} = P(O_t = \sigma_k | X_t = q_i),\ 1 \le i \le N, 1 \le k \le M$. b_{ik} meets $b_{ik} \ge 0$, and $\sum_k b_{ik} = 1, \ \forall i_0$.

(5) The initial status distribution: $\pi = (\pi_i),\ \pi_i = P(X_1 = q_i),\ 1 \le i \le N$, that is the probability of q_i at moment t=1 , π_i meets

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$\sum_i \pi_i = 1$; starts from initial status to transfer to the end status so far.

All the status experienced in such a transferring process according to the order which is arranged into the vectors, this is called status chain and recorded as: Q, Q_t represents the source status of the No. t time; In such a transfer process, the vectors according to the output descending order in such a transferring process is called output chain, which is recorded as: O, O_t represents the output of No. t transfer. The output chain of Hidden Markov Model can be observed, but status chain is not visible.

In this paper, section 2 brings forward three fundamental questions of HMM; section 3 presents the basic application of HMM in speech tagging; section 4 presents the application of HMM in speech recognition, and the related algorithms are also introduced, in the section 5, the conclusion is obtained, HMM is not only widely used in natural language processing, but also has already been used in many other areas,

2. THE THREE FUNDAMENTAL QUESTIONS OF HMM

In the quintuple group (Q_x, Q_o, A, B, π) of HMM, if $\lambda = \{A, B, \pi\}$ is given HMM parameter, $O = \{\sigma_{k1}, \sigma_{k2}, \sigma_{k3}, \cdots, \sigma_{kT}\}$ is observation sequence, if its status set is $S = \{q_{i1}, q_{i2}, q_{i3}, \cdots, q_{iT}\}$, then HMM can solve the three following fundamental questions:

Problem (1), based on a given observation sequence, the probability algorithm of forward probability can be solved, making observation sequence appearing before T moment to extrapolate the probability of some observation value at current moment T.

Calculation algorithm can be used for forward probability: forward variable is defined as $a_t(i) = P(O \le t, X_t = q_i) = P(O_1 = \sigma_{k1}, \cdots, O_t = \sigma_{k1}; X_t = q_i)$, which is the observed probability of $O \le t$ at t moment arrive status q_i.

Initialize:
$$\alpha_1(i) = P(O_1 = \sigma_{k1}; X_1 = q_i) = b_{ik1} \bullet \pi_i, 1 \le i \le N$$

Recursive:
$$\alpha_{t+1}(j) = \left\{\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right\} b_{ikt+1}, 1 \le t \le T-1, 1 \le j \le N$$

Termination:
$$P(O) = \sum_{i=1}^{N} P(O_1 = \sigma_{k1}, \cdots, O_t = \sigma_{k1}; X_t = q_i) = \sum_{i=1}^{N} \alpha_T(i)$$

Backward probability algorithm: backward and forward probability algorithms are very similar, backward variables are defined as
$$\beta_t(i) = P(O > t \mid X_t = q_i) = P(O_{t+1} = \sigma_{kt+1}, \cdots, O_T = \sigma_{k_T}, X_t = q_i).$$

The calculation algorithm of backward probability is:

Initialize: $\beta_T(i) = 1, 1 \le i \le N$

Recursive: $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_{jkt+1} \beta_{t+1}(j)$,
$$t = T-1, T-2, \ldots, 1, 1, \le i \le N$$

Termination:
$$P(O) = \sum_{i=1}^{N} P(O_1 = \sigma_{k1}, X_1 = q_i),$$
$$P(O_2 = \sigma_{k2}, \cdots, O_T = \sigma_{kT} \mid X_1 = q_i) = \sum_{i=1}^{N} a_{ik1} \bullet \pi_i \bullet \beta_1(i)$$

After defining the forward probability, backward probability and their algorithms, the output probability P(O) can be calculated as

$$P(O) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i)$$

Problem (2), an observation sequence is given; according to the existing Hidden Markov Model, a status sequence with the largest probability is found. For HMM, some observation sequence O seen from outside is not the only one inside the system corresponding to the status sequence Q, but different sequences of status Q will produce different probability of O. The task of biggest status sequence is to look for the most likely status sequence Q according to the system's observation sequence, which makes the chance of producing O become the biggest. To solve the problem the most common method is Veterbi algorithm.Veterbi is a kind of deformation of dynamic programming algorithm, the summary is presented as follows:

Initialize:

Journal of Theoretical and Applied Information Technology
20th February 2013. Vol. 48 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\delta_1(i) = \pi_i b_{ik1}, 1 \le i \le N ,$$
$$\varphi_i(i) = 0, 1 \le i \le N$$

Recursive:
$$\delta_t(j) = \max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right]b_{ikt}, 2 \le t \le T, 1 \le j \le N$$
$$\varphi_t(j) = \arg\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right], 2 \le t \le T, 1 \le j \le N$$

Termination: $P^* = \max_{1\le i\le N}\left[\delta_T(i)\right]$, $P^{t^*} = \arg\max_{1\le i\le N}\left[\delta_T(i)\right]$

Status series solution is: $P^*t = \phi_{t+1}\left(q_{t+1}^*\right), t = T-1, T-2, \cdots, 1$

From this the best status sequence of $P(O|\lambda)$ can be obtained: $q_1^*, q_2^*, \cdots, q_t^*$.

Problem (3), the estimation of model parameters is the training problem of HMM, that is, how to determine the model $\lambda = \{A, B, \pi\}$ according to system output, makes the probability $P(O|\lambda)$ being the largest observation value. Generally Baum-Welch algorithm is used to evaluate all parameters in the model.

3. THE APPLICATION OF HIDDEN MARKOV MODEL IN PART-OF-SPEECH TAGGING

3.1 Principle of Part-of-speech Tagging

Hidden Markov Model in natural language processing in the beginning is used in Chinese word segmentation and part-of-speech tagging. In part-of-speech tagging, part-of-speech sequence is equivalent to the status sequence hidden in HMM, because part-of-speech sequence is hidden before mark, it is the goal of need to be solved, the given word string is the sequence of observation symbols and the known condition before mark. If the part-of-speech tagging problem model is regarded as a HMM, the set of part-of-speech tagging is determinate (the status number of HMM is determinate), each word corresponding to part-of-speech is also determinate, in the dictionary, every word has a definite one or several part-of-speech tags.

In HMM, part-of-speech tagging problem can be expressed as: in a given word sequence (observation value) $W = (w_1, w_2, w_3, w_4, \cdots, w_m)$, solving the most probable part-of-speech (status) sequence $T = (t_1, t_2, t_3, t_4, \cdots, t_m)$ makes the conditional probability $P(T|W)$ is the largest. $P(T|W)$ is hard to estimate currently, generally it uses the Bayesian principle for conversion, that is:

$$P(T|W) = \frac{P(T)P(T|W)}{P(W)}$$

In part-of-speech tagging, W is given, P(W) does not rely on T, therefore in the calculation of P(T|W), P(W) needs not to be considered, meanwhile applying the joint probability formula $P(A,B) = P(A)P(B|A)$, there is $P(T|W) = P(T)P(W|T) = P(W,T)$. further application is being done by the probability multiplication formula is doing, there is

$$P(T|W) = \frac{P(T)P(T|W)}{P(W)}$$

$$= \prod_{i=1\cdots m} P(W_i, t_i | W_1, \cdots, i-1, t_1, \cdots, i-1)$$

$$= \prod_{i=1\cdots m} P(W_i | W_1, \cdots, i-1, t_1, \cdots, i-1) \times P(t_i | W_1, \cdots, i-1, t_1, \cdots, i-1)$$

In the expression,

$$W_1, \cdots, i = W_1, W_2, \cdots, W_i, \quad t_1, \cdots, i = t_1, t_2, \cdots, t_i, 1 \le i \le m.$$

What needs to point out is the above model reflects in ideal condition is the probability distribution of related part-of-speech tagging, but for the estimated parameters space is too big, the model can not actually be calculated. Therefore, in the actual part-of-speech tagging, the model usually needs to be simplified to reduce the parameter space. Thus some independence hypotheses are introduced (that is Markov assumption) as follows:

The emergence of t_i part-of-speech tags only depends on limited prior N-1 part-of-speech tagging, that is N-POS model.

$$P(t_i | w_1, \cdots, i-1, t_1, \cdots, i-1)$$
$$\approx P(t_i | t_1, \cdots, i-1)$$
$$\approx P(t_i | t_{i-N+1}, t_{i-N+2} \cdots t_{i-1})$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A word appearance does not depend on any words, only rely on prior part-of-speech tags, and further suppose words w_i only depends on part-of-speech tagging t_i, that is :

$$P(w_i|w_1,\cdots,i,t_1,\cdots,i)\approx P(w_i|t_1,\cdots,i)\approx P(w_i|t_i)$$

After the above assumptions, a HMM with N-1 order of part-of-speech tagging can be obtained:

$$P(T|W)\approx\prod_{i=1,\cdots,m}P(t_i|t_{i-N+1},t_{i-N+2},\cdots,t_{i-1})P(W_i|t_i)$$

In the expression, $P(W_i|t_i)$ is called emission probability; Parameter $P(t_i|t_{i-N+1},t_{i-N+2}\cdots t_{i-1})$ becomes the status transition probability.

If the emergence of part-of-speech tagging t_i depends on just a limited prior part-of-speech t_{i-1}, the emergence of word w_i only depends on part-of-speech tagging t_i, thus $P(T|W)\approx\prod_{i=1,\cdots,m}P(t_i|t_{i-1})P(W_i|t_i)$. $P(t_i|t_{i-1})$ and $P(W_i|t_i)$ are two key parameters in the expression, hereinto $P(W_i|t_i)$ refers to the probability of part-of speech word w_i in t_i ; $P(t_i|t_{i-1})$ represents the times from part-of speech t_{i-1} to next t_i.

In the large-scale corpus, according to law of large number, the expression can be obtained, $P(W_i|t_i)\approx C(W_i,t_i)/C(t_i)$, $C(W_i,t_i)$ represents the times while the part-of-speech W_i is t_i; $C(t_i)$ represents the times of part-of-speech t_i,

$$P(t_i|t_{i-1})\approx C(t_{i-1},t_i)/C(t_{i-1})$$

In the expression, $C(t_{i-1},t_i)$ represents the times from part-of－speech t_{i-1} to next part-of－speech t_i , $C(W_i,t_i),C(t_{t-1}),C(t_i)$ are all obtained from corpus library with good mark and segmentation.

3.2 The Experimental Data in Part-of-speech Tagging

The corpus library in "People's Daily" in 1998 with part labels are used to test and train corpus to obtain transfer matrix table and frequency table of part-of-speech, and are used to do HMM tagging, the samples are shown as follows:

Bring (v.) the parents (noun) hope (verbs or noun) and (conjunction) entrust (noun) to arrive (v.) Beijing (noun)

Tagging process:

P (noun | auxiliary) = C (auxiliary, the noun) / C (auxiliary) = 45214/74792 = 0.6

P (verb | auxiliary) = C (auxiliary verbs) / C (auxiliary) = 5235/74792 = 0.07

P (hope | noun) = C (hope, noun) / C (noun) = 364/385325 = 0.0009

P (hope | verbs) = C (hope, verbs) / C (v.) = 402/201424 = 0.002

P (noun | auxiliary) P (hope | noun) = 0.6 * 0.0009 = 0.00054

P (verb | auxiliary) P (hope | verbs) = 0.07 * 0.002 = 0.00014

Tagging results: bring (v.) the parents (noun) hope (noun) and (conjunction) entrust (noun) to arrive(v.) Beijing (noun)

Conclusion: lots of experiment results show that based on the HMM speech tagging, the correct rate of mark results can reach 92%.

4. THE APPLICATION OF HIDDEN MARKOV MODEL IN SPEECH RECOGNITION

In the 1980s, J.K.Baker of CMU university applied the HMM to speech recognition field, in the speech recognition it achieves great success, and becomes the main method of speech recognition.

According to the introduction of the hidden markov model principle, we known it is a double random process in speech recognition, these two random processes commonly describe the statistical properties of the speech signal, one is an implicit stochastic process that uses markov chain with finite state number to analog voice signal change, and the other is an observation vector random process which is closely associated with each state of markov chain. So a certain period of characteristics of speech and time-varying signals are described by corresponding state observation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

symbol of stochastic process, and the change of signal at any time is described by transfer probability of hidden markov chain.

Based on hidden markov model, a speech recognition algorithm count based on a large number of speech data, and establishes the recognition statistical model, then extracts features from voice for identification, measures the similarity through comparision with the model, and outputs the recognition result which is the category of highest similarity model. In general, hidden markov model constitutes a speech recognition system or speaker's recognition system, which needs to solve three basic problems.

4.1 The Calculation Of Observing Output Probability $P(o|\lambda)$

For the given observed sequence O $(O_1, O_2, O_3, \ldots, O_T)$ and model $\lambda = \{\pi + A + B\}$, the probability of model λ to produce zero O can adopt forward probability and backward probability, in order to make the calculation reduced to N^2T operations.

Definition 1 Forward probability uses observation sequence before T moment to calculate the probability of an observed value at current time T, that is,vusing the probability of $O_1, O_2, \ldots O_{t-1}$ to calculate the probability of appearing $O_1, O_2, \ldots O_{t-1}, O_t$. It can be presented by $\alpha_t(i)$.

Forward probability calculation algorithm is:

(1)Initialization:

$$\alpha_1(i) = \pi_i b_i(o_1), 1 \le i \le N$$

(2)Recursion:

$$\alpha_{t+1}(j) = \{\sum_{i=1}^{N} \alpha_t(i) a_{ij}\} b_j(O_{t+1}), 1 \le t \le T-1,$$

$$1 \le j \le N$$

(3) End:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Definition 2 Backward probability uses $O_{t+2}, O_{t+3}, \ldots, O_N$ to calculate the probability of $O_{t+1}, O_{t+2}, \ldots, O_N$, which is presented by $\beta_t(i)$.

Backward ratio algorithm is listed as following:

(1)Initialization:

$$\beta_T(i) = 1, 1 \le i \le N$$

(2) Recursion:

$$\beta_t(i) = \sum_{j=1}^{N} \alpha_{ij} b_j(o_{t+1}) \beta_{t+1}(j), t = T-1, T-1, \ldots 1, 1 \le i \le N$$

End:

$$P(o|\lambda) = \sum_{i=1}^{N} \beta_T(i)$$

After defined forward probability、backward probability and their algorithms, observation output probability $P(o|\lambda)$ is easy to be obtained:

$$P(o|\lambda) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i)$$

4.2 The Search Of Best State Sequence

For HMM system, some sequence O observed from external is not the only one in the internal system corresponding to state sequence Q, but the probability of different Q producing O is not the same. The searching task of best status sequence is looking for the most likely state sequence Q according to the system output O, making the possibility of state sequence to produce O become the maximum. The commonly used algorithm is Viterbi algorithm. Viterbi algorithm is a kind of deformation of dynamic programming algorithm, it can be got via recursion algorithm as follows:

(1) Initialization:

$$\delta_1(i) = \pi_i b_i(o_i), 1 \le i \le N$$

$$\varphi_1(i) = 0, 1 \le i \le N$$

(2) Recurse:

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), 1 \le i \le N$$

$$2 \le t \le T, 1 \le j \le N$$

$$\varphi_t(i) = \arg \max_{1 \le t \le N} [\delta_{t-1}(i) a_{ij}]$$

$$2 \le t \le T, 1 \le i \le N$$

(3) End:

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$q_T{}^* = \arg\max_{1\le i\le N}[\delta_T(i)]$$

(4) State sequence solving:

$$q_t{}^* = \varphi_{t+1}(q_{t+1}{}^*), t = T-1, T-2, ..., 1$$

Therefore the best state sequence of $P(O\,|\,\lambda)$ can be obtained: $q_1{}^*, q_2{}^*, ..., q_3{}^*$.

4.3 Model Parameter Estimation

Model parameter estimation is a training problem of HMM model, that is, how to determine the model $\lambda = \{\pi, A, B\}$ and make $P(O\,|\,\lambda)$ maximum according to the given system output. The researcher generally adopts Baum-Welch re-estimation method to do HMM model training.

Baum - Welch algorithm can be described as follows:

Thus $$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O\,|\,\lambda)},$$

$$\gamma_t(i) = \sum_{j=1}\xi_t(i,j)$$

Thus the famous re-estimate formula in Baum-Welch algorithm:

$$\overline{\pi}_i = \gamma 1(t), \overline{a_{ij}} = \frac{\sum_{t=1}^{T-1}\xi t(i,j)}{\sum_{t-1}^{T-1}\gamma_t(i)}, \overline{b_{ij}} = \frac{\sum_{t=1, O_T=0}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}$$

$\overline{\lambda} = \{\overline{\pi}, \overline{A}, \overline{B}\}$ is the model parameter after estimation, and $P(O\,|\,\overline{\lambda}) \ge P(O\,|\,\lambda)$.

Complex speech recognition problem is simply expressed and solved through hidden markov model. Hidden markov model recognition system is better than any other system, there exists more training data statistical information in hidden markov model recognition system, and solving the difficulties in training and classification, therefore the application of the hidden markov model in speech recognition is a great success.

5. CONCLUSION

The application of Hidden markov model in speech recognition and speech tagging still needs to improve, and the application of the model is not only in these two aspects. HMM as a good math’s model, its research is relative thorough, algorithm is mature, efficiency is high, effect is good, and it is easy to train, in the model. If the observation value is taken as Chinese word; the status is regarded as English, which can solve the machine translation problems. HMM is not only widely used in natural language processing, but also has already been used in many other areas, such as signal processing, image processing, machine vision, even genetic engineering and other disciplines.

REFRENCES:

[1] Xiaolong Wang, Yi Guan. *Computer natural language processing* . Beijing: tsinghua university press, 2005.

[2] Daniel Jurafsky, James H Martin. *The overview of natural language processing.* Beijing: electronic industry press, 2005.

[3] Kaiying Li, Bingyan Guo. *Natural language processing*. Beijing: science press, 1991.

[4] Rui Wen, Qiaoming Zhu, Pei Li. The application of HMM and negative feedback model in part-of-speech tagging. J*ournal of Suzhou University*, 2005(7):40-42

[5] Wei Lu. The application of Hidden Markov in natural language understanding. *Computer and information technology*, 2007, 15(1):33-34

[6] Jiaheng Zheng, Hui Zhang . Chinese organization name automatic identification Based on the HMM. *Computer Application*，2002，22(11)：l-2.

[7] Xiaoming Wu, Changxin Song . Hidden markov model used for protein sequence analysis. *Biomedical engineering*,2002，19(3)：455-458.

[8] Kristie Seymore，Andrew McCallum, Ronald Rosenfeld. *Learning hidden markov model structure for information extraction.*

[9] FreitagD ， MeCallum A . Information extraction with HMM structures learned by stochastic optimization. *Proceedings of the Eighteenth Conference on Artificial Intelligence*. Edmonton：AAAI Press. 2002：584-589.

[10] Freitag D ， McCallurn A . Information extraction with HMM and shrinkage[C] . *Proceedings of the AAAI Workshop on Machine Learning for Information Extraction* . Orlando：AAAI Press，1999：31-36.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[11] Cutting D, Kupiec J, Pedersen J, etal. A practical part-of-speech tagger. Trento, Italy: *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL,1992.133-140.

[12] Schutze H, Singer Y. Part-of-speech tagging using a variable memory Markov model. Las Cruces, New Mexico: *Proceedings of ACL 32'nd*, 1994. 181-187.

[13] Fine S, Singer Y, Tishby N. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 1998, 32(1):41-62.