Improvement for the automatic Part-of-speech Tagging Based on Hidden Markov Model

Lichi Yuan
School of Information Technology, Jiangxi University of Finance & Economics
Nanchang 330013, China
E-MAIL: yuanlichi@sohu.com

Abstract—**In this paper, the Markov Family Models, a kind of statistical Models was firstly introduced. Under the assumption that the probability of a word depends both on its own tag and previous word, but its own tag and previous word are independent if the word is known, we simplify the Markov Family Model and use for part-of-speech tagging successfully. Experimental results show that this part-of-speech tagging method based on Markov Family Model has greatly improved the precision comparing the conventional POS tagging method based on Hidden Markov Model under the same testing conditions. The Markov Family Model is also very useful in other natural language processing technologies such as word segmentation, statistical parsing, text-to-speech, optical character recognition, etc.**

Keywords- Markov Family model; Part-of-Speech tagging; Hidden Markov model; Viterbi algorithm

I. INTRODUCTION

Tagging words with their correct part-of-speech (singular proper noun, predeterminer, etc) is an important precursor to further automatic natural language processing. Part-of-speech tagging is used as an early stage of linguistic text analysis in many applications, including subcategorization acquisition, text-to-speech synthesis, and corpus indexing. Two prominent distinct approaches to be found in previous work are rule-based morphological analysis on the one hand, and stochastic Model such as Hidden Markov Models (HMMs) on the other hand.

Rule-based morphological analysis relies on hand-crafted rules to decompose input tokens into their morphological components, computing the resultant lexical category as a function of those components. Such systems incorporate the linguistic competence of their human authors, to the extent that such competence can be and is expressed in the systems' rule sets. Unfortunately, the construction of hand-crafted rule set for unrestricted input tokens of a given language is a time-consuming and labor-intensive task. Another common problem for token-wise rule-based approaches is that of ambiguity-in order to determine which of multiple possible analyses of a single token is the correct one, some reference in the context in which the token occurs is usually required.

Stochastic tagging techniques such as Hidden Markov Models rely on both lexical and bigram probabilities estimated from a tagged training corpus in order to computer the most likely PoS tag sequence for each sequence of input tokens. The existence of hand-tagged training corpora for many languages and the robustness of the resulting Models have made stochastic taggers quite popular. Disadvantages for HMM taggers include the large amount of training data required to achieve high levels of accuracy, as well as the fact that no clear allowance is made in traditional HMM tagging architectures for prior linguistic knowledge.

The Hidden Markov Models used for tagging have three assumptions[1]: (1) limited horizon, (2) time invariant (stationary), (3) simplifying assumption: probability of a word depends only on its own tag, but these assumptions (especially the third assumption) are too crude. In this paper, the Markov Family Model, a kind of statistical Models was firstly introduced. Under the assumption that the probability of a word depends both on its own tag and previous word, but its own tag and previous word are independent if the word is known, we simplify the Markov Family Model and use for part-of-speech tagging successfully. Experimental results show that this part-of-speech tagging method based on Markov Family Model has greatly improved the precision comparing the conventional POS tagging method based on Hidden Markov Model under the same testing conditions. The Markov Family Model is also very useful in other natural language processing technologies such as word segmentation, statistical parsing, text-to-speech, optical character recognition, etc.

II. HMM AND ITS APPLICATIONS IN TAGGING

2.1 Hidden Markov model

Definition 2.1 Hidden Markov model

A hidden Markov model[1] (HMM) is a five-tuple (S, A, V, B, π) where:

$S=\{s_1,\cdots,s_N\}$ is a finite set of states;

$V=\{v_1,\cdots,v_M\}$ is a finite observation alphabet;

$\pi=\{\pi_1,\cdots,\pi_N\}$ is the distribution of initial states, where

$$\pi_i = P(X_1 = s_i\}\ 1 \le i \le N \qquad (1)$$

$A=(a_{i,j})_{N\times N}$ is a probability distribution on state transitions, where

$$a_{i,j}=P(X_{t+1}=s_j \mid X_t=s_i) \quad (2)$$

is the probability of a transition to state s_j from s_i;

$B=(b_{j,k})_{N\times M}$ is a probability distribution on state symbol emissions, where

$$b_{i,k}=P(o_t=v_k \mid X_t=s_i)\ 1\le k\le M,\ 1\le i\le N \quad (3)$$

is the probability of observing the symbol v_k when in state s_i.

From the definition of hidden Markov model, it can be seen that hidden Markov model is based on a double stochastic process: a finite state Markov chain is the hidden stochastic process, the other stochastic process is the observation sequence related to the state Markov chain. A major unrealistic assumption with HMM is that successive observations are independent and identically distribution within a state. In order to cope with the deficiencies of the classical HMM, Markov Family model, a new statistical model was introduced.

2.2. Using HMM for tagging

For a tagset (T) and a finite set of word (W), it is customary to define a bigram HMM part-of-speech tagger (T, A, W, B, π), where the probability functions A, B, and π are estimated from a tagged training corpus. Under such a model, part-of-speech tags are represented as states of the model, and the task of finding the most likely tag sequence $t_{1,n}$ for an input word sequence $w_{1,n}$ can be formulated as a search for the most likely sequence of HMM states given the observation sequence $w_{1,n}$:

$$\arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n}) = \arg\max_{t_{1,n}} \frac{P(w_{1,n} \mid t_{1,n})P(t_{1,n})}{P(w_{1,n})}$$

$$= \arg\max_{t_{1,n}} P(w_{1,n} \mid t_{1,n})P(t_{1,n}) \quad (4)$$

We now introduce this expression to parameters that can be estimated from the training corpus. In addition to the Limited Horizon assumption (3), we make two assumptions about words: words are independent of each other, and a word's identity only depends on its tag.

$$P(w_{1,n} \mid t_{1,n})P(t_{1,n}) = \prod_{i=1}^{n} P(w_i \mid t_{1,n}) \times P(t_n \mid t_{1,n-1}) \times P(t_{n-1} \mid t_{1,n-2}) \times \cdots \times P(t_2 \mid t_1)$$

$$= \prod_{i=1}^{n} P(w_i \mid t_i) \times P(t_n \mid t_{n-1}) \times P(t_{n-1} \mid t_{n-2}) \times \cdots \times P(t_2 \mid t_1)$$

$$= \prod_{i=1}^{n} P(w_i \mid t_i) \times P(t_i \mid t_{i-1}) \quad (5)$$

(We define $p(t_1 \mid t_0) = 1.0$ to simplify our notation.)

III. MARKOV FAMILY MODEL AND ITS APPLICATION IN POS TAGGING

3.1 Markov Family model

Definition 3.1 (Markov Family model)

Let $\overrightarrow{\{X_t\}}_{t\ge1} = \{x_{1,t}, \cdots x_{m,t}\}_{t\ge1}$ is a m-dimensional stochastic vector, whose componential variable $X_i=\{x_{i,t}\}_{t\ge1}, 1\le i\le m$ taking values in finite set $S_i, 1\le i\le m$. It can be said that these componential variables $X_i, 1\le i\le m$ construct a m-dimensional Markov Family model if satisfying the following conditions:

1) Each componential variable $X_i, 1\le i\le m$ is a n_i-order Markov chain.

$$P(x_{i,t} \mid x_{i,1}, \cdots, x_{i,t-1}) = P(x_{i,t} \mid x_{i,t-n_i+1}, \cdots, x_{i,t-1}) \quad (6)$$

2) What value a variable will take at time t is only related to its previous values before time t and the values that the rest variables take at time t.

$$P(x_{i,t} \mid x_{1,1}, \cdots, x_{1,t}, \cdots x_{i,1} \cdots, x_{i,t-1} \cdots, x_{m,1}, \cdots x_{m,t}) =$$
$$P(x_{i,t} \mid x_{i,t-n_i+1}, \cdots, x_{i,t-1}, x_{1,t}, \cdots x_{i-1,t}, x_{i+1,t}, \cdots x_{m,t}) \quad (7)$$

3) Conditional independence:

$$P(x_{i,t-n_i+1}, \cdots, x_{i,t-1}, x_{1,t}, \cdots x_{i-1,t}, x_{i+1,t}, \cdots x_{m,t} \mid x_{i,t}) =$$
$$P(x_{i,t-n_i+1}, \cdots, x_{i,t-1} \mid x_{i,t}) \cdot P(x_{1,t} \mid x_{i,t}) \cdots P(x_{m,t} \mid x_{i,t}) \quad (8)$$

Condition 1 means that Markov Family model is constructed on a multiple stochastic process. From this point, it can be said that the standard HMM is a special case of MFM. Condition 2 demonstrates the relations among these Markov chains of MFM, and it can also simplify the calculation of model. According to Condition 3, the previous n_i-1 values that a variable X_i will take before time t and the values that the rest variables take at time t are independent if the value of the variable X_i takes at time t is known. From the view of the statistics, the assumption of independence is stronger than the assumption of conditional independence, and it can be inferred from independence to conditional independence. So the assumption of conditional independence in Markov Family model is more realistic than the assumption of independence in HMM.

3.2 Using Markov Family model for Tagging

A major unrealistic assumption with HMM tagging model is that successive words (observations) are independent and identical distribution within a tag (state).

Under the assumption that the probability of a word depends both on its own tag and previous word, but its own tag and previous word are independent if the word is known, Markov Family model has been successfully applied to Part-of-speech tagging.

Let S_1 be the finite set of Part-of-Speech tags, S_2 be the finite set of words, and Markov chain the properties of Markov Family model, a word's tag and its previous word are independent if the word is known:

$$P(w_{i-1},t_i \mid w_i)=P(w_{i-1} \mid w_i)\cdot P(t_i \mid w_i) \qquad (9)$$

For simplicity, also suppose that word sequence $\{w_i\}_{i\geq 1}$ and tag sequence $\{t_i\}_{i\geq 1}$ are all 2-order Markov chain, thus can find the sequence of tags $t_{1,n}=t_1,\cdots t_n$ that maximizes the probability of the tag sequence given the word sequence $w_{1,n}=w_1,\cdots w_n$.

$$\begin{aligned}\arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n}) &= \arg\max_{t_{1,n}} \frac{P(w_{1,n} \mid t_{1,n})P(t_{1,n})}{P(w_{1,n})} \\ &= \arg\max_{t_{1,n}} P(w_{1,n} \mid t_{1,n})P(t_{1,n})\end{aligned} \qquad (10)$$

Where

$$P(w_{1,n} \mid t_{1,n})=P(w_n \mid w_1,\cdots,w_{n-1},t_1,\cdots,t_{n-1},t_n)\cdot P(w_{1,n-1} \mid t_{1,n-1}) \qquad (11)$$

According to the properties of Markov Family model, have:

$$P(w_{1,n} \mid t_{1,n})=P(w_n \mid w_{n-1},t_n)\cdot P(w_{1,n-1} \mid t_{1,n-1}) \qquad (12)$$

From the equation (9), can get

$$\begin{aligned}P(w_n \mid w_{n-1},t_n) &= \frac{P(w_{n-1},t_n \mid w_n)\cdot P(w_n)}{P(w_{n-1},t_n)} \\ &= \frac{P(t_n \mid w_n)\cdot P(w_{n-1} \mid w_n)\cdot P(w_n)}{P(t_n \mid w_{n-1})\cdot P(w_{n-1})} \\ &= \frac{P(t_n \mid w_n)\cdot P(w_n \mid w_{n-1})}{P(t_n \mid w_{n-1})}\end{aligned} \qquad (13)$$

So

$$\begin{aligned}&\arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n}) \\ &= \arg\max_{t_{1,n}} P(w_1 \mid t_1)\cdot P(t_1)\prod_{i=2}^{n}\frac{P(t_i \mid w_i)\cdot P(t_i \mid t_{i-1})}{P(t_i \mid w_{i-1})}\end{aligned} \qquad (14)$$

Once have a probabilistic model, the next challenge is to find an effective algorithm for finding the maximum probability tag sequence given an input. The *Viterbi Algorithm*[1] is a dynamic programming method which efficiently computers for a given word sequence $w_1,\cdots w_n$ most likely to generate the tag sequence $t_1,\cdots,t_n$ according to the model parameters. The computer proceeds as follows:

comment: Given: a sentence of length n, the number of the tag set is T.

comment: initialization

$\delta_1(t^j)=P(w_1 \mid t^j)\cdot P(t^j), 1\leq j\leq T$

$\Psi(t^j)=0 \quad 1\leq j\leq T$

comment: Induction

for *i*: = 1 to *n*-1 step 1 do

for all tags t^j do

$\delta_{i+1}(t^j)=\max_{1\leq k\leq T}[\delta_i(t^j)\times P(t^j \mid w_{i+1})\times P(t^j \mid t^k) / P(t^j \mid w_i)]$

$\Psi_{i+1}(t^j)=\arg\max_{1\leq j\leq T}[\delta_i(t^j)\times P(t^j \mid w_{i+1})\times P(t^j \mid t^k) / P(t^j \mid w_i)]$

end

end

comment: Termination and path-readout, $X_1,\cdots,X_n$ are the tags choose for words $w_1,\cdots w_n$

$X_n = \arg\max_{1\leq j\leq T} \delta_n(j)$

for j: =n-1 to 1 step -1 do

$X_j = \Psi_{j+1}(X_{j+1})$

end

$P(X_1,\cdots,X_n)=\arg\max_{1\leq j\leq T}\delta_n(j)$

Figure 1. Algorithm for tagging

IV. EXPERIMENTAL RESULTS

We use an annotated corpus selected from People's Daily newspaper 1998 for training and testing. The corpus uses 42 tags, and has about 244974 tokens. Some statistical properties about the annotated corpus are as follows:

42 tags	22345 types		244974 tokens	
1	20048	89.720%	162246	66.230%
2	1934	8.655%	50243	20.510%
3	297	1.329%	21419	8.743%
4	51	0.228%	9901	4.042%
5	10	0.045%	424	0.173%
6	4	0.018%	155	0.063%
7	1	0.004%	586	0.239%

The experimental results are demonstrated in table 1.

TABLE I. TAGGING EXPERIMENTAL RESULTS

model	Hidden Markov model	Markov Family model
accuracy	94.642%	96.214%

From table 1, it can be seen that tagging method based on Markov Family model has higher performance than the conventional POS tagging method based on Hidden Markov model under the same testing conditions; the precision is enhanced from 94.642% to 96.214%.

V. CONCLUSIONS

The advent of hidden Markov model (HMM) has brought about a considerable progress in natural language processing and speech recognition technology. However a number of unrealistic assumptions with HMMs are still regarded as obstacles for its potential effectiveness. A major one is the inherent assumption that successive observations are independent and identical distribution (IID) within a state. In order to overcome the defects of the classical HMM, Markov Family model, a new statistical model is introduced in this paper and it overcomes the defects of unrealistic assumptions about HMM. Tagging experimental results have verified the efficacy of the proposed model. Certainly, theory about Markov Family model should be progressed right along, and the applications of MFM in speech recognition will be studied in the future.

ACKNOWLEDGMENT

This work was supported by the National Nature Science Foundation of China(60763001) and the Nature Science Foundation of jiangxi province(2009GZS0027).

REFERENCES

[1] Christopher D Manning, Hinrich Schutze. Foundations of Statistical Natural Language Processing [M]. London: the MIT Press, 1999.

[2] Sharon Aviran, Paul H. Siegel, Jack K. wolf. Optimal Parsing Trees for Run-length Coding of Biased Data. IEEE Transaction on information Theory. 2008, 2, 54(2):841-849.

[3] Deyu Zhou, Yulan He. Discriminative Training of the Hidden Vectors State Model for Semantic Parsing. IEEE Transaction on Knowledge And Data Engineering. 2009,1, 21(1): 66-77.

[4] J. Gimenez, L. Marquez. Fast and accurate part-of-speech tagging: The SVM approach revisited. The 4th RANL P, Bulgaria, 2003.

[5] Eugene Charniak, Curtis Hendricson, Neil Jacobson, and Mike Perkowitz. Equations for Part-of-Speech tagging. Proceedings of the Eleventh National Conference on Artificial intelligence, Menlo Park, AAAI Press/MIT Press, 1993: 784-789.

[6] B. Turish. Part-of-speech tagging with finite-state morphology. Poster presented at the conference Collocations and Idioms: linguistic, Computational, and Psycholinguistic perspectives, Berlin, 18-20 September, 2003.

[7] T. Brants. A statistical Part-of-Speech tagger. In Proceeding of the sixth Applied Natural Language Processing Conference, ANLP-2000, Seattle, WA, April 29, 2000.

[8] Rabiner L. R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Feb 1989, vol.77(2): 257 - 285.

[9]YUAN Li-chi. A speech recognition method based on improved hidden Markov model[J]. Journal of Central South University: Natural Science, 2008, 39(6): 1303-1308.

[10] LIU Shui, LI Sheng, ZHAO Tie-Jun, et al. Directly Smooth Interpolation Algorithm in Head-Driven Parsing[J]. Journal of Software, 2009, 20(11):2915-2924.

[11] LI Zhenghua, CHE wanxiang, LIU Ting. Beam-Search Based High – Order Dependency Parser[J]. Journal of Chinese Information Processing, 2010, 24(1): 37-41.