SPEECH ACT MODELING IN A SPOKEN DIALOGUE SYSTEM USING FUZZY HIDDEN MARKOV MODEL AND BAYES' DECISION CRITERION

Chung-Hsien Wu, Gwo-Lang Yan and Chien-Liang Lin

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.

Abstract

In this paper, a corpus-based fuzzy hidden Markov model (FHMM) is proposed to model the speech act in a spoken dialogue system. In the training procedure, 29 FHMM's are defined and trained, each representing one speech act in our approach. In the identification process, the Viterbi algorithm is used to find the top M candidate speech acts. Then Bayes' decision criterion, which stores the relationship between the phrase and the speech act, is employed to choose the most probable speech act from the top M speech acts. In order to evaluate the proposed method, a spoken dialogue system for air travel information service is investigated. The experiments were carried out using a test database from 25 speakers (15 male and 10 female). There are 120 dialogues, which contains 725 sentences in the test database. The experimental results show that the correct response rate can achieve about 82.7% using the FHMM and the Bayes' decision criterion.

1. INTRODUCTION

In recent years, the domain of spoken dialogue has been broadly researched. Many application systems such as air travel information service, weather forecast system, automatic call manager, and railway ticket reservation have been presented [1]-[6]. But there are still many problems making the dialogue system unnatural. Especially, in the language understanding model, some approaches that use a large set of rules to explain the syntactic and semantic possibilities for spoken sentences suffer from a lack of robustness when faced with the wide variety of spoken sentences that people really use. In addition, traditional approaches just allow users to make a clear inquiry without ambiguity. Generally, they have low capability to identify the exact speech act from an erroneous sentence generated from a speech recognizer. In this paper, a statistical speech act model, called fuzzy hidden Markov model (FHMM), is proposed to identify some candidate speech acts from the phrase sequences generated from the speech recognizer. A postprocess based on Bayes' decision criterion is employed to predict and verify the identified speech act. The combination of FHMM and Bayes' decision criterion will effectively reduce the misidentification rate resulted from the speech recognition errors.

In our approach, a spoken dialogue system for air travel information service is investigated. The architecture of this system is shown in Figure 1. The input speech is recognized into a syllable lattice by a speech recognizer. The syllable lattice is then used to generate possible phrase sequences and reduces some insertion errors, such as hesitation and repetition, using a pre-processor. In the semantic analysis module, 29 FHMM's are constructed and used to determine the possible speech acts using the Viterbi algorithm. The combination of the scores from FHMMs' and the Bayes' decision criterion is used to verify the reliability of the speech act. For the operation of the dialogue system, the dialogue manager is mixed initiative. For an incomplete inquiry, the system can initiatively ask the user about the information in order to complete the semantic slots. In addition, if the user finds that the system does not acquire the correct information, the user may repair it in the next turn of the dialogue. Finally, the TTS module generates the speech response to the user.

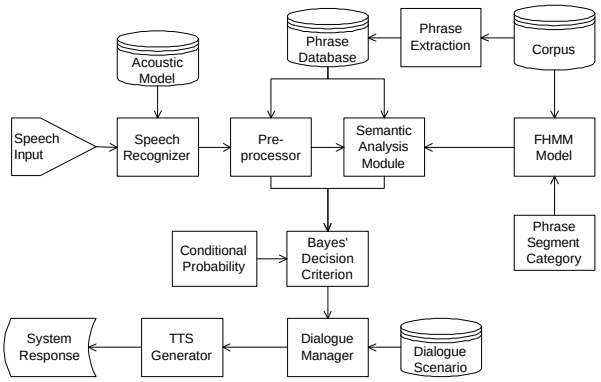

Figure 1. System Architecture of the Dialogue System

2. PHRASE ANALYSIS

2.1. Phrase Extraction

After analyzing the corpus for air travel information service, a phrase extraction method called phrase-like unit extraction [7] is employed to extract phrases from the specific corpus. There are 178 phrases extracted from the corpus automatically.

2.2. Phrase Segment Category

Based on the phrases extracted above, we build a specific dictionary and calculate a bigram relation vector for each phrase. The bigram relation vector $x_t, \{1 \le t \le T\}$, are used as the features to cluster the phrases. We assume that each phrase belongs to one or more clusters and a fuzzy membership function is used to represent the degree of the phrase belonging to a cluster. The entry u_{it} in the membership matrix U is the fuzzy membership of the vector x_t with respect to the cluster center C_i and satisfies

$$0 \le u_{it} \le 1 \quad \text{and} \quad \sum_{i=1}^{I} u_{it} = 1 \tag{1}$$

The fuzzy C-means algorithm [8] is adopted for clustering. Furthermore, the fuzzy objective function Z_m can be defined by the least-square function as

$$Z_m(U, \mathbf{m}; X) = \sum_{t=1}^{T}\sum_{i=1}^{I} u_{it}^{m} d_{it}^{2} \tag{2}$$

Where U is the membership matrix, $\mathbf{m} = (\mathbf{m}_1, .., \mathbf{m}_I)$ are cluster centers, $m(>1)$ is the exponential weight, and d_{it} is the measure of distance defined in the following

$$d_{it} = \cos(0) - \cos(\theta_{x_t, \mathbf{m}_i}) = 1 - \left(\frac{x_t \bullet \mathbf{m}_i}{|x_t| * |\mathbf{m}_i|}\right) \tag{3}$$

The distance d_{it} is defined based on the cosine value of the

10.21437/Eurospeech.1999-312

angle between two vectors x_t and m_i with the absolute cost to the cosine zero, which is the best condition between two vectors.

The basic idea in fuzzy C-means algorithm is to minimize Z_m over the membership matrix U and cluster centers m. The update functions to minimize Z_m are as follows

$$u_{it} = \frac{\left(\frac{1}{d_{it}^2}\right)^{1/(m-1)}}{\sum_{i=1}^{I}\left(\frac{1}{d_{it}^2}\right)^{1/(m-1)}} \quad (4)$$

$$m_i = \frac{\sum_{t=1}^{T} u_{it}^m x_t}{\sum_{t=1}^{T} u_{it}^m} \quad (5)$$

This algorithm is described as follows

Fuzzy C-Means Algorithm:

Step 1: Initialize I to 1 and the initial cluster c_1 to x_1. Select a radius R and m (m>1).

Step 2: For each input x_t, {t=1,2,…, T}, sequentially calculate the following equation:
If $|x_t - c_i| \le R$, {i = 1,2,...I}, then $x_t \in c_i$,
else create a new cluster c_{I+1} to x_t and $I = I + 1$

Step 3: Compute distance d_{it} by (3)

Step 4: Update matrix U using (4) and calculate new cluster centers by (5)

Step 5: Stop if the decrease in the value of fuzzy objective function Z_m at the current iteration relative to the value of the Z_m at the previous iteration is below a chosen threshold, otherwise go to step 3.

After the process of fuzzy C-means algorithm, some phrases, which have similar syntactic and semantic meaning, may have similar membership function. In total, we classify them into 27 clusters, each representing one phrase segment category (PSC), using the bigram relation vector.

2.3. Speech Act Analysis

In a spoken dialogue, the phrase is an important component for a speech act. Each speech act can be decomposed into several phrases and can be identified according to the combination of phrases. In our approach, 29 speech acts are defined automatically based on the corpus and phrases.

3. FUZZY HIDDEN MARKOV MODEL FOR SPEECH ACT

3.1. Definition of Speech Act FHMM [9]

In this approach, the speech act FHMM is used to model the sequence of PSC. Each PSC represents one state of the speech act FHMM. The membership of the phrase in the PSC represents the observation probability. The state transition is the transition from one PSC to another PSC. The discrete observation of speech act FHMM is defined as follows:

- N: the number of states in the model, each state represents one PSC. We label the individual states as {1,2,…,N} and denote the state for the λth state as S_λ.
- M: the number of distinct observation symbols per state. The observation symbols correspond to the input phrases in the task. We denote the individual symbols as
$$V = \{v_1, v_2, ..., v_M\} \quad (6)$$
- The state-transition probability distribution from state i to state j is represented by
$$a_{ij} = P(S_{\lambda+1} = j \mid S_\lambda = i),\ 1 \le i, j \le N \quad (7)$$
- The observation symbol distribution at state j is defined as
$$b_j(k) = P[O_\lambda = v_k \mid S_\lambda = j] * PhS(O_\lambda), 1 \le k \le M \quad (8)$$
Where $P[O_\lambda = v_k \mid S_\lambda = j]$ is the fuzzy membership $u_{j\lambda}$, and $PhS(O_\lambda)$ is the normalized speech recognition score O_λ.
- The initial state is $p_i = P[S_\lambda = i],\ 1 \le i \le N \quad (9)$

The type of FHMMs used in this paper is a standard Markov model. That is, each state can transit to any state. The allowable transition paths are trained by the training corpus.

3.2. Construction of Speech Act FHMM

The construction of speech act FHMM can be divided into three parts. They are phrase collection, phrase clustering, and training of FHMM. They are briefly explained below:

1. Phrase collection [7]: The main work in this step is to collect the corpus. Word segmentation is performed first to choose the keywords for the specific task. Finally, the important and meaningful keywords are combined and defined as the phrases. We collect about 178 phrases to form a task-specific dictionary.

2. Phrase clustering: This step clusters the phrases into phrase segment category. The fuzzy C-means algorithm is based on the bidirectional phrase bigram vector. The main criterion is to cluster the phrases with similar syntactic and semantic structure.

3. Training of FHMM: The training corpus is tagged with 29 speech acts. Each FHMM is trained by the subcorpus belonging to its corresponding speech act.

3.3. Speech Act Identification

In the identification of speech act, given a speech utterance S, the phrase sequence can be determined according to the following equation:

$$P_h(PS_k \mid S) = \max_{1 \le h \le H}[\log P_h(CS_k \mid S)] + a\sum_l \log P(ph_l^k \mid ph_{l-1}^k) \quad (10)$$

where PS_k is the kth phrase sequence. H is the number of the speech act FHMMs. ph_l^k is the lth phrase in the sequence of the kth phrase sequence. $P(ph_l^k \mid ph_{l-1}^k)_{i(}$ is the phrase bigram probability. For an input speech S, $P_h(CS_k \mid S)$ expresses the probability corresponding to the kth PSC sequence CS_k via the hth speech act FHMM. It can be denoted by the following equations:

$$P_h(CS_k \mid S) = \max_{1 \le i \le N} \delta_L^{k,h}(i) \quad (11)$$

$$\delta_l^{k,h}(i) = \max_{S_1 S_2 \ldots S_{l-1}} P[S_1 S_2 \ldots S_l = i, o_1 o_2 \ldots o_l \mid \lambda_h] \quad (12)$$

Where N is the number of states. $\delta_l^{k,h}(i)$ is the highest probability along a single path, for the *l*-th input phrase, which accounts for the first *l* observations $O = [o_1 o_2 \ldots o_l]$ and ends at state *i*. For example, the phrases in the sentence " "(I want to book the flight departing at two o'clock this afternoon) can be segmented into " "(I want to book)," "(this afternoon)," "(two o'clock)," "(de), and " "(flight). The corresponding word segment categories are "Action", "Time," "Time," "Filler," and "Flight," respectively. The speech act FHMM with the highest probability among the entire speech act FHMMs should be the "Booking" FHMM.

4. BAYES' DECISION CRITERION

The Bayes' decision criterion is a postprocess to verify and choose the most probable speech act from the top M candidate speech acts. Given some phrases as evidences, it is achievable to verify the speech act from the combination of phrases in a sentence.

4.1 Bayes' probability model

The Bayes' theory is a good method to deal with the problem of uncertainty. According to the Bayes' theory, the probability of hypothesis H given evidence E can be described as

$$P(H \mid E) = \frac{P(H) * P(E \mid H)}{P(E)} \quad (13)$$

For an arbitrary number of hypotheses H_i ($i = 1,.., k$), which are mutually exclusive and exhaustive, we suppose the H_i partition the universe. Equation (13) can be generalized as

$$P(H_i \mid E) = \frac{P(H_i) * P(E \mid H_i)}{\sum_{i=1}^{k} P(E \mid H_i) * P(H_i)} \quad (14)$$

Generally, to be more realistic and to accommodate multiple sources of evidence $E_1, E_2 .., E_n$, we generalize equation (14) further to obtain

$$P(H_i \mid E_1 E_2 .. E_n) = \frac{P(H_i) * P(E_1 E_2 .. E_n \mid H_i)}{\sum_{i=1}^{k} P(E_1 E_2 .. E_n \mid H_i) * P(H_i)} \quad (15)$$

Actually, the probability $P(E_1 E_2 .. E_n \mid H_i)$ is very difficult to calculate from corpus. But it is easier to get the probability $P(E_j \mid H_i)$, if we assume that every evidence is statistically independent. We can get that

$$P(E_1 E_2 .. E_n \mid H_i) = P(E_1 \mid H_i) * P(E_2 \mid H_i) * .. * P(E_n \mid H_i) \quad (16)$$

In order to avoid the misrecognition problem of phrase or statistical problem due to sparse data, we discard E_j if $P(E_j \mid H_i) \le T$, where T is a chosen threshold.

4.2 Verification using Bayes' Decision Criterion

In the verification process using Bayes' decision criterion, the phrases are used as the evidences E_j and the speech act is treated as a hypothesis H_i. The verification score for the *kth* speech act is defined as the combination of the output score of the FHMM and the score from the Bayes' decision criterion with respect to its corresponding FHMM.

$$Vscore(k) = -\alpha \log(P_k(PS \mid S)) - (1-\alpha)\log(P(H_k \mid E_1 E_2 .. E_n)) \quad (17)$$

Where α is the weight between 0 and 1. The speech act with the highest *Vscore* and above a chosen threshold is regarded as the final output. On the contrary, some candidate speech acts from FHMMs' may be rejected using the verification process.

5. DIALOGUE MANAGER

In order to make the system friendly and get complete required information, we propose some dialogue strategies when interacting with the users.

- Mixed initiative strategy: The system usually guides the users to give the required information for a specific intention. On the contrary, the user also can inquire the information that he/she needs actively. For example, the user may inquire which flight departs by three o'clock in the afternoon.
- Confirmation strategy: To make sure the information that the system gets is correct, the user must confirm the information he/she provided. But not all of the data will be reconfirmed. The system just makes sure some important semantic slots. Therefore, at the end of the dialogue, the system will make the final check of the information in the semantic slots.
- Repair strategy: The user can correct the information he/she provided at any time.
- Recovery strategy: If the user changes the content of the semantic slot, the system will update the semantic slot and respond according to the new content in the semantic slot.

6. EXPERIMENT

In order to evaluate the proposed method, a spoken dialogue system for air travel information service is investigated. The system has been implemented on an IBM personal computer with a Dialogic/ESC telephone interface card. The experiments were carried out using a test database from 25 speakers (15 male and 10 female). There are 480 dialogues, which contains 3038 sentences. .

6.1. Experiment on Speech Act Identification

In this experiment, the speech database was divided into two databases. The first one containing 2313 sentences was transcribed into text corpus and used to train the speech act FHMM's. They were also used as the inside test database. The second speech database was first transcribed into text and then used as the outside test database to evaluate the system performance. The test database contains 725 sentences. Table 1 shows the results for inside and outside tests with text input.

Table 1. The identification results for inside and outside test.

	Speech Act Accuracy (%)
Inside Test	97.8
Outside Test	92.5

6.2. Experiment on Response Accuracy

For evaluating the response capability of the system, the telephone speech recognition output was used directly as the input of our proposed system. The 725 sentences in speech form were fed to the telephone speech recognizer to output

phrase sequences. First, we have to determine the value of a in order to achieve the optimal combination of FHMM and Bayes' decision criterion. The experimental results shown in Figure 2 give the correct response rate for the different values of a. If $a = 1$, *Vscore* is totally dominated by FHMM. The correct response rate is about 80.3%. On the contrary, if $a = 0$, *Vscore* is dominated by the Bayes' decision criterion. The correct response rate is about 78.9%. When a is chosen as 0.7, the system can achieve the best correct response rate of 82.7%.

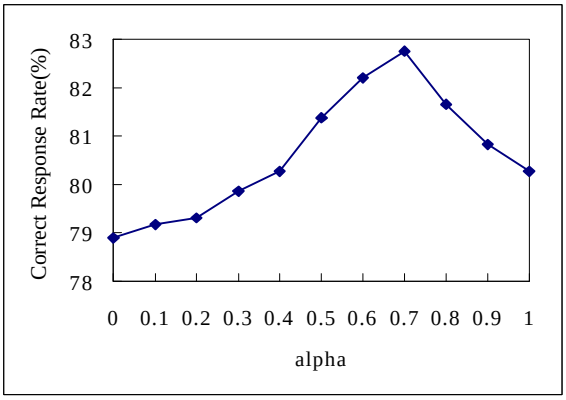

Figure 2. Correct response rate as a function of the value a.

Another experimental results shown in Table 2 list that about 80.3% of the sentences can be responded correctly using speech act FHMM at a speech recognition rate of 78%. The correct response rate can be improved by 8% compared to a baseline system. This result shows that speech act FHMM is useful to identify the speech act from a sentence. Furthermore, the correct response rate will be improved by 10% using both FHMM and Bayes' decision system. The Bayes' Decision criterion along with FHMM can further improve the response rate.

Table 3. Correct response rate at a speech recognition rate of 78%.

	Response rate(%)
Baseline	72.6
Baseline + speech act FHMM	80.3
Baseline + speech act FHMM + Bayes' Decision Criterion ($a = 0.7$)	82.7

7. CONCLUSION

In this paper, we have proposed a corpus-based FHMM to identify speech act from a sentence and the Bayes' decision criterion is used to verify the identified speech act to improve it's reliability. This FHMM model combines not only the bigram of the phrases, but also the syntactic and semantic structure of a sentence. The Bayes' decision criterion utilizes the relation between phrases and speech act to verify and choose the most probable speech act. Experimental results show that the system can achieve the correct response rate of 82.7% using both speech act FHMM and Bayes' Decision Criterion. It shows that using speech act FHMM and Bayes' Decision Criterion is capable of identifying the speech act of a sentence and achieves encouraging improvement for spoken dialogue processing.

8. REFERENCES

[1] Helen Meng., Senis Busayapongchai, and Victor Zue, et al. "WHEELS: A Conversational System in the Automobile Classification Domain," ICSLP '96 Vol. 1. pp. 542-545

[2] S. Bennacef and L. Lamel et al., "Dialog in the RAILTEL Telephone-Based System," ICSLP'96 Vol. 1. pp. 550-553

[3] Frank Seide and Andreas Kellner, "Toward an Automated Directory Information System," EuroSpeech'97 Vol.3. pp.1327-1330

[4] Chun-Jen Lee, Eng-Fong Huang, and Jung-Kuei Chen, "A Multi-Keyword Spotter for the Application of the TL Phone Directory Assistant Service," Proceedings of 1997 Workshop on Distributed System Technologies & Applications, pp. 197-202

[5] Tung-Hui Chiang, Chung-Ming Peng, Yi-Chung Lin, Huei Ming Wang and Shih-Chieh Chien, "The Desigh of A Mandarin Chinese Spoken Dialogue System," in Proceedings of COTEC'98 , Taipei 1998, pp. E2-5.1~E2-5.7

[6] Hsien-Chang Wang, Jhing-Fa Wang, and Yi-Nan Liu, "A Conversational Agent for Food ordering Dialog Based on Venus Dictate," Proceedings of ROCLING X International Conference 1997,pp.325-334

[7] Yu-Sheng Lai and Chung-Hsien Wu, "Unknown Word and Phrase Extraction Using a Phrase-Like-Unit-Based Likelihood Ratio," in Proceedings of ICCPOL99, Tokushim, Japan, 1999, Vol. 1. pp.5-9

[8] H.-J. Zimmermann, "Fuzzy Set Theory and It's Applications,"Kluwer Academic Publishers, 1991, pp230-236

[9] Chung-Hsien Wu, Gwo-Lang Yan, and Chien-Liang Lin, "Spoken Dialogue System Using Corpus-Based Hidden Markov Model," in *Proceedings of ICSLP98*, Sydney, Australia, 1998.