Spoken Dialogue System Using Corpus-Based Hidden Markov Model

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 a spoken dialogue system, the intention is the most important component for speech understanding. In this paper, we propose a corpus-based hidden Markov model (HMM) to model the intention of a sentence. Each intention is represented by a sequence of word segment categories determined by a task-specific lexicon and a corpus. In the training procedure, five intention HMM's are defined, each representing one intention in our approach. In the intention identification process, the phrase sequence is fed to each intention HMM. Given a speech utterance, the Viterbi algorithm is used to find the most likely intention sequences. The intention HMM considers not only the phrase frequency but also the syntactic and semantic structure in a phrase sequence. In order to evaluate the proposed method, a spoken dialogue model 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 contain 725 sentences in the test database. The experimental results show that the correct response rate can achieve about 80.3% using intention HMM.

1. INTRODUCTION

In this decade, spoken dialog systems have been broadly researched [1]-[6]. Many application systems such as air travel information service, automatic call manager[3][4], and railway ticket reservation[2] have been presented. But there still remain many problems in spoken dialogue modeling. Traditional approaches just allow the user to make a clear inquiry without ambiguity. Generally, they have low capability to identify the exact intention from an erroneous sentence generated from a speech recognizer. In this paper, a corpus-based intention hidden Markov model (HMM) is proposed to choose a meaningful and grammatical phrase sequence from the phrase sequences generated from the speech recognizer. This will effectively reduce the misidentification rate resulted from the speech recognition errors.

In a spoken dialogue system, the intention is the most important component for speech understanding. In this paper, a corpus-based HMM is used to model the intention. The phrases used in a specific task are determined using a universal lexicon and the mutual information between phrases in the corpus. These phrases are then classified into M word segment categories(WSC) in which phrases have similar semantic meaning or syntactic structure based on the bigram probabilities between two phrases. In the construction of intention HMM, each state represents a WSC. The mixture component in the state is represented by the frequency distribution of the phrase in this WSC. The transition probability between two states is determined using the corpus. In the training procedure, five intention HMM's are defined, each representing one intention in our approach. In the intention identification process, the phrase sequence is fed to each intention HMM. The Viterbi algorithm is used to find the best intention sequences. The intention HMM considers not only the phrase frequency but also the syntactic and semantic structure in a phrase sequence. Eventually, most illegal intention sequences can be rejected using the intention HMM's.

In our approach, a spoken dialogue model for air travel information service is investigated. The architecture of this system contains four modules. They are telephone speech recognition module, semantic analysis module, dialogue module, and text-to-speech (TTS) module. In the speech recognition module, the input telephone speech is recognized into syllable lattice. The syllable lattice is then used to generate possible phrase sequences using a task-specific lexicon. In the semantic analysis module, five intention HMM's including greeting, inquiry, booking, ending, and filler are constructed and used to determine the possible intention sequences using the Viterbi algorithm. The dialogue module is mixed initiative. For an incomplete inquiry, the system can initiatively asks the user about the information in order to complete the semantic slots. Besides, 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 responding speech to the user.

The architecture of this system is shown in Figure 1. It contains four modules described below.

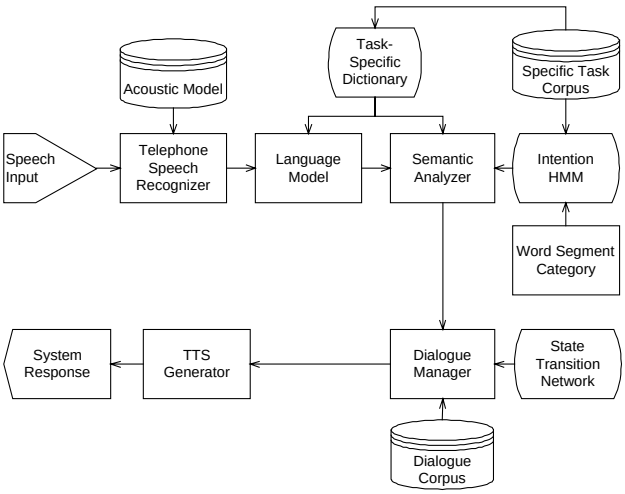

Figure 1. System Architecture of the Dialogue System

- Telephone speech recognition module: Transcribe the telephone speech into syllable lattice and convert them into a sentence based on a task-specific lexicon.
- Semantic analysis module: Five intention HMM's which are greeting, inquiry, booking, ending, and filler are employed to identify the intention from a sentence.
- Dialogue module: Respond to the user according to the intentions from semantic analysis module.
- TTS module: Convert the text response into speech .

2. DIALOGUE ANALYSIS

2.1. Dialogue Feature

After analyzing many dialogues between customers and the bookers of the airlines, the process of booking can be divided into three parts. They are explained below.

1. Greeting: the greetings between customers and the bookers.
2. Information Exchange:
   a. Inquiry: Inquire flight date, time, and number.
   b. Booking: Book a flight.
   c. Others: Computer-generated codes or repetition.
3. Ending: Ending conversation when booking or inquiry is completed such as thank you or good-bye.

2.2. Word Semantic Category

According to the dialogues described above, we use this corpus to build a specific dictionary and calculate the bigram probabilities of phrases. Some phrases may have similar syntactic and semantic meaning. We classify them into the same category using the bigram probabilities followed by the manual examination. Phrases with similar syntactic and semantic structure are assigned with a word segment category . In total there are 21 WSC's in our system

3. INTENTION HIDDEN MARKOV MODEL

3.1. Definition of Intention HMM

Every intention HMM is used to model the sequence of word segment categories. Each WSC represents one state of the intention HMM. The occurrence probability of the phrase in the WSC represents the observation probability. The state transition is the transition from one WSC to another WSC. The discrete observation of intention HMM is defined as follows:

- N : the number of states in the model, each state represents one WSC. We label the individual states as {1,2,…,N} and denote the state for the ℓ-th state as S_ℓ.
- 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

$$V = \{v_1, v_2, ..., v_M\} \quad (1)$$

- The state-transition probability distribution from state i to state j is represented by

$$a_{ij} = P(S_{\ell+1} = j \mid S_\ell = i) \ , \ 1 \le i, j \le N \quad (2)$$

- The observation symbol distribution at state j is defined as

$$b_j(k) = P[O_\ell = v_k \mid S_\ell = j] * PhS(O_\ell), 1 \le k \le M \quad (3)$$

  Where $PhS(O_\ell)$ is the normalized speech recognition score of O_ℓ.

- The initial state is $\pi_i = P[S_\ell = i], \ 1 \le i \le N \quad (4)$

The type of HMMs 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 Intention HMM

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

1. Phrase collection: 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 chosen as the phrases. We collect about 200 phrases to form a task-specific dictionary.
2. Phrase clustering: This step clusters the phrases into word segment category. The clustering method is based on the bidirectional word bigram probabilities. The main criterion is to cluster the phrases with similar syntactic and semantic structure.
3. Training of HMM : The training corpus is tagged with five intentions. Each HMM is trained by the subcorpus belonging to its corresponding intention.

3.3. Intention Identification

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

$$S(PS_k \mid U) = \max_{1 \le h \le H} [\log P_h(WS_k \mid U)] + \alpha \sum_l \log P(ph_l^k \mid ph_{l-1}^k) \quad (5)$$

where PS_k is the $k-th$ phrase sequence. H is the number of the intention HMMs. ph_l^k is the $l-th$ phrase in the sequence of the $k-th$ phrase sequence. $P(ph_l^k \mid ph_{l-1}^k)$ is the phrase bigram probability. For an input speech U, $P_h(WS_k \mid U)$ expresses the probability corresponding to the $k-th$ WSC sequence WS_k via the $h-th$ intention HMM. It can be denoted by the following equation:

$$P_h(WS_k \mid U) = \max_{1 \le i \le N} \delta_L^{k,h}(i) \quad (6)$$

$$\delta_l^{k,h}(i) = \max_{S_1 S_2 \ldots S_{l-1}} P[S_1 S_2 \ldots S_l = i, o_1 o_2 \ldots o_l \mid \lambda_h] \quad (7)$$

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 in state *i*. For example, the phrases in the sentence "我要訂下午二點的飛機"(I want to book the flight departing at two o'clock this afternoon) can be segmented into "我要訂"(I want to book),"下午"(this afternoon),"二點"(two o'clock),"的"(de), and "飛機"(flight). The corresponding word segment categories are "Action", "Time", "Time ", "Filler", and "Flight," respectively. The intention HMM with the highest probability in all the intention HMMs should be the "Booking" HMM.

4. DIALOGUE MANAGER

The dialogue manager processes the user's intention and fill out the semantic slots for a specific intention. According to the current status of the semantic slots, the dialogue manager gives an appropriate response. For example, when the semantic slots for booking are fully filled out, the user completes a booking process. There are five semantic slots in the booking process. They are date, time, departure, destination, and number of people. Due to the speech recognition error and the requirement of information, it is necessary for the system to interact with the user. The block diagram of the dialogue manager is shown as in Figure 2.

4.1. Dialogue Strategies

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 make 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 and response properly according to the new content in the semantic slot.

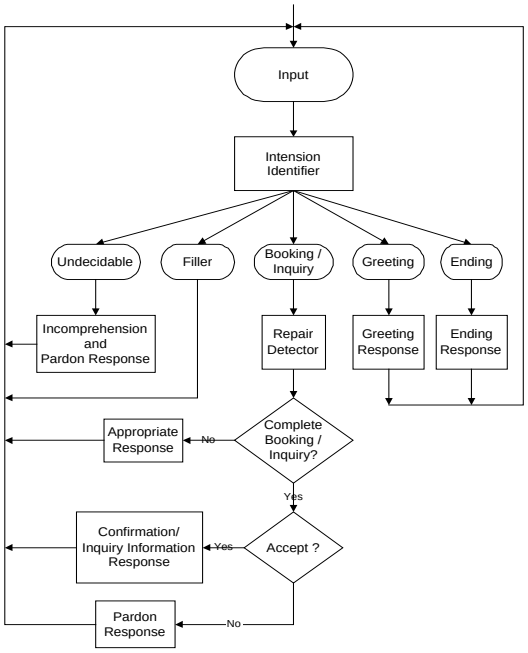

Figure 2. The block diagram of the dialog manager.

4.2. Dialogue Response Generation

In an active system, the system communicates with the users iteratively in order to obtain all the necessary data to process a complete booking. The system must give the user an appropriate response at the same time. The interaction of the dialogue depends on the intention identifier which measures the user's intention. The responses are classified into three types.

- Greeting response: In the beginning of a dialog, the system prompts a greeting message "This is Far Easten airline. How may I help you?"

- Inquiry response: When the semantic slots are not completely filled out, the system will ask the user to provide the necessary information.
- Multiple-answer response: If the inquiry is ambiguous, that is, there are more than one answer for the user's inquiry, our system lists all the possible choices and asks the user to select what he/she wants. For example,

  User : "Is there any flight departing at about three p.m. "

  System: "The closest flights are at 2:40 and 3:30. Which one do you want?"
- Failure response: If the system cannot respond properly for three times, the system will transfer the booking process to a real operator.
- Ending response: It responds when the system completes a dialogue. A successful booking process implies that the five semantic slots have been filled out.

5. EXPERIMENT

In order to evaluate the proposed method, a spoken dialogue model 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. .

5.1. Experiment on Intention 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 intention HMM's. They were also used as the close-test database. The second speech database was first transcribed into text and used as the open-test database to evaluate the system performance. This database contains 725 sentences. Table 1 shows the results of close test and open test with text input. From this table, the "ending" intention HMM gives the highest intention identification rate because the phrases in the "ending" intention are different from that in other intentions. The "filler" intention HMM has lower performance because it contains many out-of-vocabulary phrases.

Intention	Booking	Inquiry	Greeting	Ending	Filler
Close Test(%)	93.6	92.5	90.9	97.3	82.2
Open Test(%)	90.6	84.0	82.6	91.2	80.0

Table 1. The identification results for open and close test.

5.2. 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. The experimental results shown in Table 2 list that about 80.3% of the sentences can be responded correctly with intention HMM at a speech recognition rate of 78%. Using intention HMM, the correct response rate can be improved by 8%. These results show that intention HMM is useful to identify the intention from a sentence. The intention will dominate the semantic meaning of a sentence and affect the success of a dialogue.

Total speech sentences	725
Correct response rate without Intention HMM	72.6%
Correct response rate with Intention HMM	80.3%

Table 2. Correct response rate at a speech recognition rate of 78%.

6. CONCLUSION

In this paper, we have proposed a corpus-based HMM to identify intention from a sentence. This model combines not only the bigram of the phrases, but also the syntactic and semantic structure of a sentence. Experimental results show that the system can achieve the correct response rate of 80.3% using intention HMM. In other words, it shows that using intention HMM is capable of identifying the intention of a sentence and achieves encouraging improvement in dialogue processing.

7. 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