Model and Phonetic Decision Tree. ([2], [3]) The Hidden Markov Model represents states of speech and transitional relationships between these states, and the Phonetic Decision Trees are implemented according to each state. In addition, TransTac Program which DARPA sponsored uses the Gaussian Mixture Model consisting of 1300 acoustic states and one hundred thousands of Gaussians as the acoustic model. Some studies implemented the Deep Neural Network for the acoustic model. A few programs of speech recognition such as Apple's SIRI provide small numbers of models and data to enable the user's device to participate in computing. However most of the commercialized speech recognition programs employ lots of data and complex models so they do not compute results in user-end products such as mobile phones but transmit recorded data to server computers and receive computed results. ([4])

Such systems like SIRI not only have a speech recognition model, but also a language generation model. Once these systems convert voice to text, they analyze its meaning and do something for a good response. Since the system responds to the users, many people feel comfortable and perceive it as a smart system.

Studies on language generation can be classified into 2 groups: a knowledge-based type and a data-driven type. The knowledge-based speech recognizing system has difficulty in applications to different situations; otherwise, the system of the data-driven class requires a large amount of data to be implemented. Furthermore, general-purpose language generation systems of corporations need to communicate with server computers so devices which are not able to connect to the internet are constrained to use these systems.

Therefore, we suggest the language generation model which requires a relatively small amount of data and situational knowledge in this research. To learn the temporal correlation of dialogues, we used the structure of the Hidden Markov Model. Speech recognizing models apprehend intentions of dialogue and generate appropriate sentences according to the prediction of the conversation-behavior class. The results of this study is meaningful as there is the possibility of considering approximately long term intentions of previously articulated speeches due to the structure of the Hidden Markov Model and implementations of the language generator employing correlations of dialogues among the training data is possible.

This research used recorded dialogues between clerks and customers at a coffee shop for training and testing our speech recognizing model. If additional information such as the menu and data about conversations between shop assistants and customers are provided, the speech recognition model can be utilized in a diverse range of sale businesses as well as cafes or restaurants.

## 2 Related works

Cuayáhuitl et al. (2005) assumed that user's responses are related to the sequence of responses. They made Natural Language Generation (NLG) Model using three kinds of the Hidden Markov Model: HMM, Input-HMM (IHMM), and Input-Output-HMM (IOHMM) with 2-gram. This model generates sentence using user's intention predicted by this model and state which is classified as system. According to the evaluation of three different HMM model using DARPA Communicator data, all of them shows similar performance. ([5])

Jung(2009) indicated problems of existing data-driven-based language generation model: the limit of user patterns and poor user type controllability. He suggested logistic regression based Markov framework model. He regarded that user's intentions given discourse context are consist of domain knowledge and discourse knowledge. This model regards intention as a combination of frames. It makes overcome the limit of user patterns and can simulate various user discourses. ([6])

Chu-Carroll (1998) made intention recognizing model using HMM and n-gram. This model uses discourse history as well as current utterance. She assumed that if a model uses discourse structure, it will be better performance than before. She made two models which difference is whether use discourse structure or not. However, the gap of these two models is just less than 2%p. She also used n-best results. In her result, as n is larger, the performance is better. ([7])

Most of commercial NLG systems are based on the finite-state model and it is hard to expand the model reflecting new situations because of fixed numbers of states. On the other hands, corpus-based model is relatively easy to expand, so Kang and Ko (2013) used MDP and partially observed MDP (POMDP) and POS tagging. They grouped adjacent state to large state and the task completion rate was 91%, 90% (F-score) for each of two kinds of corpora according to results of their experiments. ([8])

## 3 Discourse System

We assumed that person's each utterance implies an intention of previous dialogues and these two factors, utterance and intention, are highly correlated. Therefore, as Fig. 1 displays, we suggested a new intention classifier/ utterance generation model based on

HMM. Since our assumption about people's discourse is that series of intention affect next utterance, the HMM was utilized to build our model and Fig. 2 depicts exact scheme of the HMM.

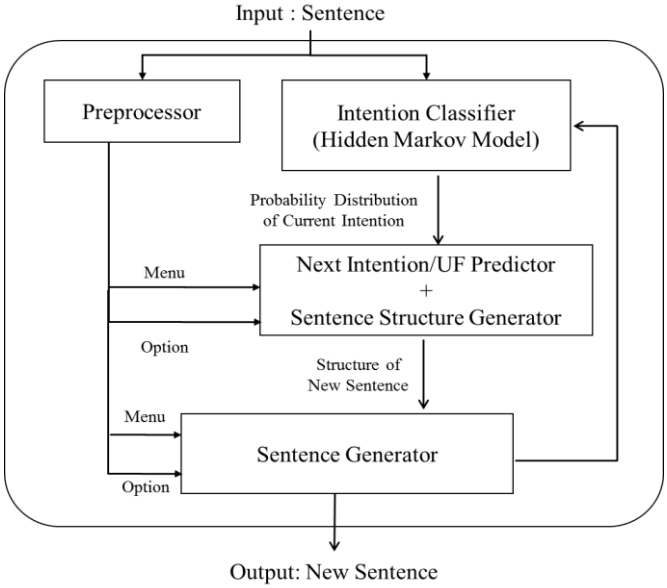


**Fig. 1.** Structure of our suggesting model which analyzes purposes of input sentences and makes replies to them

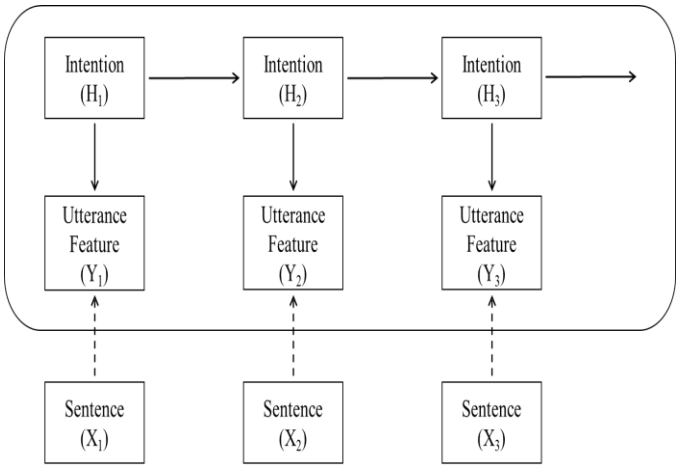


**Fig. 2.** Structure of the Hidden Markov Model and variables in Eq. (1)-(4)

When person's speech is entered into the model, preprocessor and intention classifier treat the input sentence. The role of preprocessor is extracting information about menu or option if they exist in the input. It makes a list about menu and option which appear in the input and this extracted information will be utilized by sentence structure generator and sentence generator. Intention classifier is the HMM-based intention predictor which produces probability distribution of input's intention over subdivision of intention. In the case of generating next sentence using predicted intention, sentence structure generator chooses appropriate structure of sentence from previously collected corpus considering probability of next intention and the number of menu and option in the sentence structure. And then, sentence generator merges the chosen structure of sentence and information of menu and option of previous input speech to make output which will be a reply to the input utterance. In addition to sentence generator model on the basis of intention prediction, we constructed and tested the performance of sentence generator which chooses sentence structure by considering next utterance feature (UF). Detailed explanation will be provided in following sections.

### 3.1 Intention Classifier

As mentioned above, intention classifier uses HMM to infer intention of an entered sentence and predict intention of probable following utterance in the dialogue. This concept of intention classifier was represented in the paper [9], some authors of this thesis's previous work. Similar to conventional modeling technique using HMM, the suggested model takes a hidden layer representing the intention of utterances and a visible layer for the utterances. As the HMM can only take discretized nodes and the utterance set is too sparse to deal with small dimensional data, sentences are divided into some groups (utterance features) by a few rules to represent properties of utterances such as price, menu, and actions.

### 3.2 Intention-based Sentence Structure Generator

Since the model is built on the HMM, the system is able to anticipate an intention of user's utterance by equation (1). Meaning of each variable of equation (1) is as in the following: $H_t$ means intention of t-th utterance, $X_t$ stands for t-th speech and $Y_t$ is the utterance feature which t-th statement belongs to.

$$\begin{aligned} P(H_t|X_t,H_{t-1}) &= P(H_t|Y_t,H_{t-1})P(Y_t|X_t) \\ &= P(Y_t|H_t)P(H_t|H_{t-1})P(Y_t|X_t)/P(Y_t|H_{t-1}) \end{aligned} \tag{1}$$

Although most of other models choose just one intention which has maximum probability, our model does not use it, but all of its probability. To reflect possibility of transferring from one intention to many different intentions, choosing one intention randomly according to probability of transition between intentions is more appropriate than just choosing maximum probable intention. Therefore, the system predicts next intention by considering transitional probability of HMM and picks proper structured utterance among collected structures of predicted intention by its weight, w of equation (2).

$$\begin{aligned} w &= 1 + N_s && (N_s < N_u) \\ &= 10(1 + N_s) && (N_s = N_u) \\ &= 0 && (N_s > N_u) \end{aligned} \tag{2}$$

$N_s$ : the number of information in system's utterance

$N_u$ : the number of information in user's utterance

System changes wildcard (i.e. '<menu 1>') in this structure to information what user's utterance has. We assumed generated sentence is well-fitted in dialogue, so we use its intention to predict next input as equation (3).

$$\begin{aligned} P(X_{t+1}|X_t, H_{t-1}) &= P(X_{t+1}|H_{t+1})P(H_{t+1}|X_t, H_{t-1}) \\ &= P(X_{t+1}|H_{t+1})P(H_{t+1}|H_t)P(H_t|X_t, H_{t-1}) \end{aligned} \tag{3}$$

### 3.3 Utterance group-based Sentence Structure Generator

To compare performance of models which have slightly different method to generate next sentence, we implemented other type of sentence structure generator which calculates probability of utterance features of next utterance. Thus, this model uses transitional probability between hidden variables, intention, and estimation probability from hidden variables to visible variables, utterance, of HMM structure. The equation for next utterance feature generating is similar to one for next intention generating model. The system generates a structured utterance accorded by same rules, and changes this utterance's wildcard to user's information.

$$\begin{aligned} P(Y_{t+1}|X_t, H_{t-1}) &= P(Y_{t+1}|H_{t+1})P(H_{t+1}|X_t, H_{t-1}) \\ &= P(Y_{t+1}|H_{t+1})P(H_{t+1}|H_t)P(H_t|X_t, H_{t-1}) \end{aligned} \tag{4}$$

## 4 Experimental results

### 4.1 Data

To evaluate performance of our model, we set a goal of experiment as apprehending intention of each sentence in conversations between customers and café clerks and made a program which is able to replace café clerks. Therefore we used conversation data collected from a coffee shop by using a recorder. The data are composed of 130 dialogues about 20 hours in length if one dialogue is defined as a consecutive sequence of conversation between a customer and a shop assistant.

We set 23 intention categories and added intention labels to all sentences of collected data. The 23 kinds of intention and example sentences for each type of intentions are in table 1. Also, the concept of utterance feature is introduced to discretize visible variable of HMM which represents each speech of speakers. We made several standards to assort sentences into utterance features and assessed their efficiency with cross-validation test of intention classifier. Table 2 shows groups of criteria for classifying utterance features. 'Menu', 'price' and 'option' in table 2 are conditions which categorize sentences whether they include some of menu, information about price or optional choices. The word 'Action' in table 2 signifies activities like paying with the credit card or signing. Furthermore, 'complete/incomplete' of conditions' group2 in table 2 reflects one characteristic of Korean language which has structural components of sentence notifying ending of sentences. These standards and information about speakers are used to make utterances into utterance feature which has discrete and finite value. Example sentences for each utterance feature are displayed in table 3.

In addition to assorting intention and utterance feature, the pool of sentence structures was built to be used by sentence structure generator. For each sentence of the data set, words of menu and option are replaced by specially pre-assigned words such as '<menu>' or '<option>' and this converted sentence is assembled into sentence structure groups according to its intention. Table 4 shows an epitome of the converting process. 'Converted Sentence' is the result of changing original sentence's information about menu and option into wildcards and 'Information' contains the information about menu and option.

**Table 1.** Subgroups of intentions and example sentences of them

| Intention of Sentence | Example Sentence |
|---|---|
| Beginning/End of Conversation | Dummy state to distinguish each dialogue |
| Greeting | "Hello" |
| Inducement to Order | "What do you want?" |
| Announcement to Order | "Hello, I want to order now." |
| Ordering(Incomplete) | "One Snyder's and" |
| Ordering(Complete) | "A bowl of adzuki-bean ice dessert with cereals and soybean milk and a cup of Chamomile, please." |
| Confirmation of Order | "A cup of cocoa and a cup of hot chocolate." |
| Question about Menu | "What kinds of adzuki-bean ice dessert are available?" |
| Change or Cancellation of Order | "Then just give me a cup of puer tea." |
| Inquiry about Option | "Do you want hot or iced?" |
| Request about Option | "No, I want a hot one." |
| Notification of Price | "It is 3800won." |
| Question about Payment | "Do I have to pay it now?" |
| Reconfirmation of Price | Words that a customer shadows. "3800won" |
| Question about Packaging | "Is it take-out?" |
| Explanation about Menu | "The topping for nuts and fruits has four options: black tea, green tea, chocolate and adzuki-bean ice dessert." |
| Payment | "Here" while a customer gives a credit card to a clerk. |
| Request to Sign | "Please sign here" |
| Signing | Action of signing. |
| Simple answer(Yes/No) | "Yes" for the question of a customer |
| Backchannel | "Yes" which a shop assistant says during a customer's ordering. |
| Gratitude Greeting | "Thank you." |
| Others | "Can I use a laptop computer and an iPad which are set on the table?" |

**Table 2.** Conditions for grouping sentences into utterance features

| Group of Conditions | Conditions for Assortment |
|---|---|
| Group 1 | Menu, Price, Others |
| Group 2 | Menu(Complete/Incomplete), Price, Others |
| Group 3 | Conditions of Group 2 + Option |
| Group 4 | Conditions of Group 3 + Action |

**Table 3.** Utterance features and example sentences of each UF

| Group of Sentences | Example Sentences |
|---|---|
| Customer – Menu(Complete) | "Give me a cup of iced latte with soybean milk." |
| Customer – Menu(Incomplete) | "Mango and strawberry yogurt and" |
| Customer – Option | "No, double please." |
| Customer – Price | "Did you say 5800won?" |
| Customer – Action | Activity of paying or signing. |
| Customer – Others | "Can I use a laptop computer and an iPad which are set on the table?" |
| Clerk – Menu(Complete) | "A cup of iced latte with soybean milk." |
| Clerk – Menu(Incomplete) | "Yeah, a cup of tonic water with plum and" |
| Clerk – Option | "Do you want iced café mocha?" |
| Clerk – Price | "It is 15000won." |
| Clerk - Action | Dummy but counter group of 'Customer – Action' |
| Clerk – Others | "Would you sign it?" |
| NULL | Dummy group to represent a measurement of a 'Beginning/End of Conversation' state |

**Table 4.** An example sentence and conversion result

| | |
|---|---|
| Original Sentence | Strawberry and citron adzuki-bean ice dessert, fruit and milk adzuki bean ice dessert and hot green plum tea please. |
| Converted Sentence | <menu 1>, <menu 2>, and <option 1> <menu 2> please. |
| Information | <menu 1> : strawberry and citron adzuki-bean ice dessert<br><menu 2> : fruit and milk adzuki-bean ice dessert<br><menu 3> : green plum tea<br><option 1> : hot |

### 4.2 Results of Intention Classifier

For assessment of intention classifier's performance, three specific results of classifier were considered: comparison of ability to estimate intention between classifiers using different conditions to categorize utterance feature, difference of classifier's accuracy related to the amount of data to train the model, and dialogue structure which can be found from transition probability of HMM.

First of all, Fig. 3 expresses accuracy of predicting intention of several intention classifier models. Accuracy of each model was obtained from 10-fold cross-validation test. First model in Fig. 3 is intention classifier using decision tree model and this classifier was suggested by Junhyuk Oh. ([10]) Other four classifiers are HMM-based intention classifiers which are explained in this paper and cond. 1 to 4 means standards for utterance feature which were shown in table 2. As depicted in Fig. 3, the increase of the number of utterance features improves performance of intention classifiers. This is because intention of utterance has 23 categories so using small kinds of utterance group results in one utterance feature-to-multiple intention correspondence. For example, when intention classifier uses condition group 1 of table 2, 10 intentions are related to sentences including menu, 3 intentions are associated with utterances involving price, and 10 intentions are connected to 'others' group of sentences. However, as conditions for UF increases, development of accuracy contracts so making UF much finely does not guarantee huge growth of performance.

In addition to comparing performance of intention classifiers which use different model or utterance feature, the intention classifier gives meaningful correlation between the amount of training data and accuracy of the classifier. Fig. 4 expresses results of cross validation test while using 10% to 90% of collected data for training. For each case of percentage, training data were chosen randomly and cross validation test was repeated 10 times so results in Fig. 4 are accuracy's mean and standard deviation. The graph shows approximate convergence of performance when using 80% or more data so HMM-based intention classifier requires relatively small amount of data to reason intention of sentences 90% accuracy.

Lastly, probability of transition between hidden variables of HMM contains information about structure of discourse. Fig. 5 is the graphical representation of transition probability of HMM. Intention of previous sentence is on the horizontal axis and intention of next utterance is on the vertical axis. Intention of each row and column follows intention ordered in table 1: a square at left lower corner signifies probability of intention shift from 'Beginning/End of Conversation' to 'Beginning/End of Conversation' and a square at right upper corner stands for probability that 'Others' intention changes to 'Others' intention. As the specific case of intention transition more probably occurs, color of a related square becomes darker. Five most probable cases of intention shift are described in table 5. These cases seem obvious but they also show that intention classifier learns structural characteristics of dialogues in specific situation.

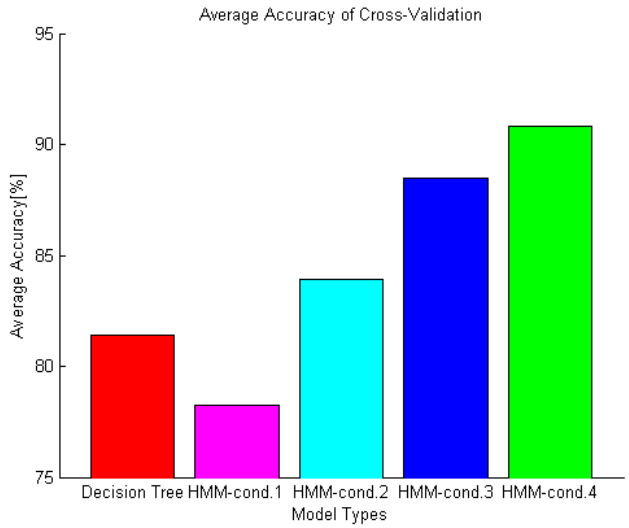


**Fig. 3.** Average accuracy of several models from 10 fold cross validation test (Decision Tree[10] and HMM model with UF condition in table 2)

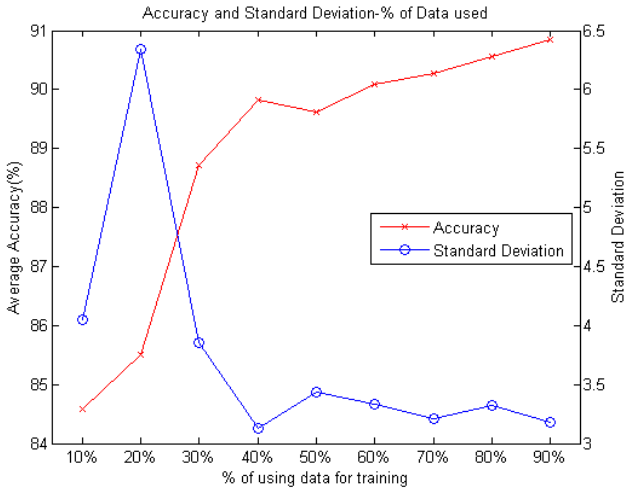


**Fig. 4.** Average accuracy and standard deviation of cross validation test according to different amount of data to train the classifier

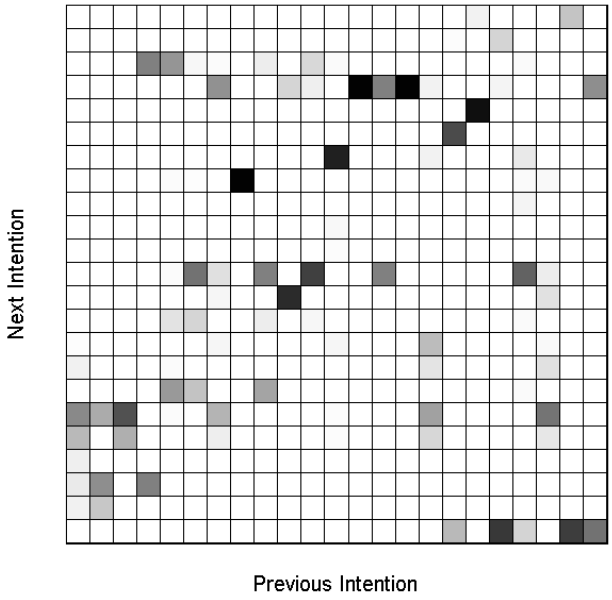


**Fig. 5.** Transition probability of HMM (previous intention on horizontal axis and next intention on vertical axis)

**Table 5.** 5 most probable cases of transitions and their probability

| Transition between Intentions | | Transition Probability |
|---|---|---|
| Previous Intention | Next Intention | |
| Question about Menu | Explanation about Menu | 0.99 |
| Question about Payment | Yes/No | 0.99 |
| Question about Packaging | Yes/No | 0.99 |
| Request to Sign | Signing | 0.94 |
| Notification of Price | Payment | 0.87 |

### 4.3 Results of Sentence Generator

We used survey to evaluate the performance of utterance generator. For the process of making the system and testing it, collected data are divided into two groups: 80% of 130 dialogues for training and rest of them for testing. And then, HMM for the intention

classifier and the utterance generator learns the training data and the system completes sequence of conversation when guests' parts of whole dialogues and the system's own outputs are used as input. For example, guest's ordering utterance which is the first sentence of the conversation is entered into the system so first output is obtained. Then, a sequence of three utterances, guest's first sentence- system's first output- guest's second utterance, becomes input of the system and related output is generated. By this method, result conversations whose relevance will be tested are generated from customers' utterances of 20 percent of data.

Therefore, the system generates 52 user-system dialogues (26 intention based generation, and 26 UF based generation). The questionnaire is made of 52 generated dialogues and 13 real dialogues (control group) in random order and surveyed people cannot know each dialogue's generating way. The subjects give marks each generated utterance and each dialogue how plausible it is by score from one to seven (seven is very plausible).

Fig. 6 expresses average score of naturalness of utterances, and Fig. 7 expresses average score of naturalness of dialogue. These scores were obtained from nine people's replies. According to these figures, control group gets the highest score in both category, and intention based generation follows control group. UF based generation gets the lowest score. The score of control group in Fig.6 and Fig.7 is less than perfect score, and it means that original conversations were a little unsatisfactory to subjects in terms of unaffectedness.

As depicted in Fig. 6 and Fig. 7, intention based generation gets higher score than UF based generation. Since UF based generation uses the weight of each utterance in same UF as equation (2) and sentences in same UF have various intentions, UF based generation shows poor performance in terms of context. Also intention can represent better than UF because it has more detailed conditions: there are 23 intentions and 13 UFs. Thus, we suggest UF based generation model as well as intention based generation model because the system's output is much more related to UF, the visible variable of HMM, rather than intention, the hidden variable of HMM but intention based generation works better than UF based generation. To be specific, UF based generation easily fails to predict next sentences and causes generation error which occurs in the case that all of utterance structure of UFs cannot represent user's information. In our survey, 27 generation errors have occurred because of UF based generation, and these errors comprise about 26% of speeches of UF based generation.

### 4.4 Result of Application in other situration

To evaluate applicability of our model, we tested the model in other situations while using existing café data as training data. We collected 12 dialogues of other café and a ticket office of bus terminal each, and make menu and option table respectively. As the method of test in above section, the system generates utterances from user's intentions.

Fig. 8 and Fig. 9 shows average score of different situation. Although training data is gathered in café, the score of terminal is higher than that of another café. It rests in difference of dialogic structure. Structure of most terminal dialogues is 'Ordering (complete)' -> 'Notification of payment' -> 'Payment' -> 'Request to sign' -> 'Sign'. It is the most frequent pattern in both training data and terminal data, so our system can generate utterances correctly. On the contrary to this, dialogic structure of another café is very different from one of the first café. Clerk always asked to save points, and every menu have options in the test cafe while training café data never appear concept of point, and most of menu do not have options. This structural difference made poor performance of utterance generator in café test data but the generator produced appropriate replies in the case of the bus terminal's ticket office.

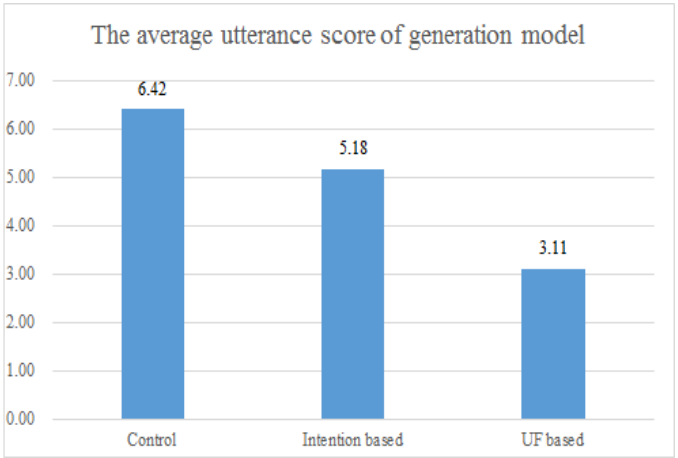


**Fig. 6.** Average utterance score of generation model

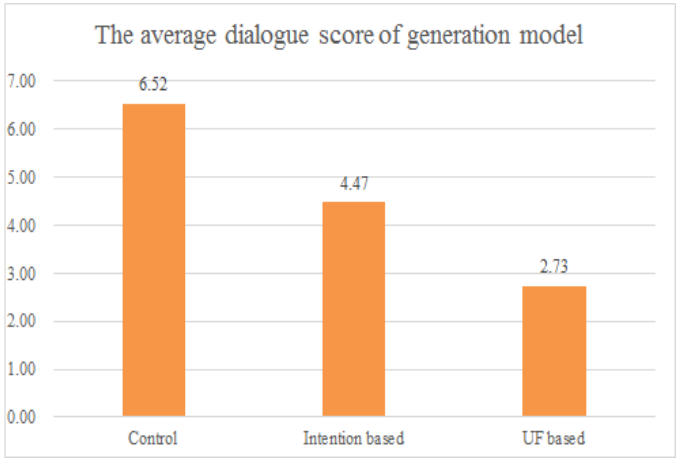


**Fig. 7.** Average dialogue score of generation model

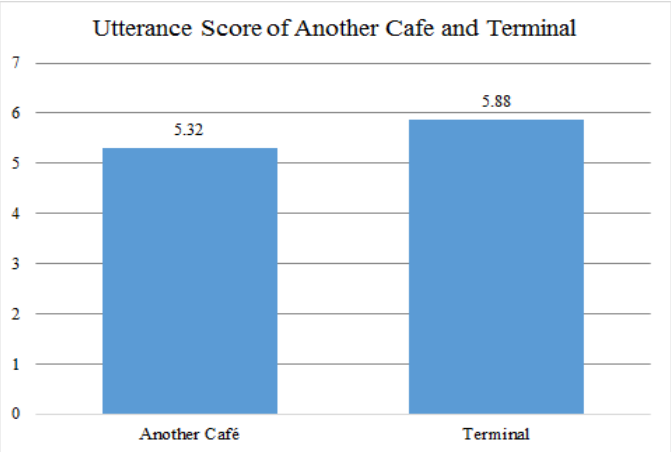


**Fig. 8.** Average utterance score of other situations

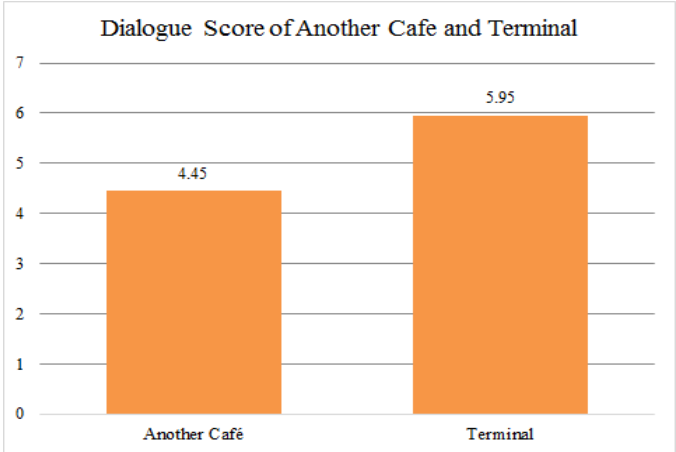


**Fig. 9.** Average dialogue score of other situations

## 5 Conclusion

In this paper, we assumed that dialogues between two people can be modeled as HMM so both the intention classifier which estimates speaker's purpose in each utterance and the sentence generator which creates responses to input sentences are implemented on the basis of HMM structure.

The intention classifier shows 90% accuracy of estimating intention of sentences using only 130 dialogues though the performance depends on given situations of conversation and criteria for utterance features. Also, the sentence generator responds to people as human does if 80 percentage of similarity is satisfactory. The system and pre-obtained sentences data occupy only 20kb in memory and average computing time for answering a sentence is about 9ms so this suggesting system has some advantages over corporations' speech recognizing programs in terms of simple structure of model, and less amount of storage consumption and computation time.

The HMM-based intention classifier and sentence generator can be applied not only as substitute of clerks in diverse shops and restaurants but to more general case of conversations focusing on the dialogic structure.

## References

1. Rybach, David, and Michael Riley. "Direct construction of compact context-dependency transducers from data." INTERSPEECH. 2010.
2. Schuster, Mike. "Speech recognition for mobile devices at Google." PRICAI 2010: Trends in Artificial Intelligence. Springer Berlin Heidelberg, 2010. 8-10.
3. Lei, Xin, Andrew Senior, Alexander Gruenstein, and Jeffrey Sorensen. "Accurate and compact large vocabulary speech recognition on mobile devices." INTERSPEECH. 2013.
4. Singh, Piyush Pratap. "Survey of Most Powerful Language Software's." International Journal of Innovative Research & Studies Vol 3, Issue 3, March 2014.
5. Cuayáhuitl, Heriberto, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. "Human-computer dialogue simulation using hidden markov models." Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on. IEEE, 2005.
6. Jung, Sangkeun, Cheongjae Lee, Kyungduk Kim, and Gary Geunbae Lee. "Hybrid approach to user intention modeling for dialog simulation." Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics, 2009.
7. Chu-Carroll, Jennifer. "A statistical model for discourse act recognition in dialogue interactions." Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-01. 1998.
8. Kang, Sangwoo, Youngjoong Ko, and Jungyun Seo. "A dialogue management system using a corpus-based framework and a dynamic dialogue transition model." AI Communications 26.2 (2013): 145-159
9. Lee, Seungwon, Eunsol Kim, and Byoungtak Zhang. "Modeling of Speech Intention using the Hierarchical Model." KIISE 2014 Winter Conference, November 2014.
10. Oh, Junhyuk, Hyosun Chun, and Byoungtak Zhang. "Generating Cafeteria Conversation with a Hypernetwork Dialogue Model." In *Proceedings of the 14th International Symposium on Advanced Intelligent Systems (ISIS 2013)*, pp. 1424-1435, 2013.