Cafebot:

A Conversational Cashier Robot in Korean Cafes

Cheolho Han, Kyoung-Woon On, Eun-Sol Kim, and Byoung-Tak Zhang
Department of Computer Science and Engineering
Seoul National University
South Korea
{chhan, kwon, eskim, btzhang}@bi.snu.ac.kr

***Abstract*—This paper discusses developments on Cafebot, a conversational cashier robot in Korean cafes. Cafebot has conversations with customers, which requires modeling the intention and response of both speakers, Cafebot and customers. The human voice is not directly understandable by robots, so it must be converted to a text. The text is also not adequate for training robots; we need to change it to digital numbers. Then Cafebot's conversational model learns the intention and response represented in those numbers. Through these processes, Cafebot gets to talk to customers in cafes. This paper introduces several methods that are available in some primary processes.**

Keywords—conversational agents; cognitive robots; bag of words; doc2vec; nearest neighbors; hidden Markov models; support vector machines; hypernetworks

I. Introduction

Cafebot [1-4] is a conversational cashier robot in Korean cafes. Cafebot has conversations with customers, which requires modeling the intention and response of both speakers, Cafebot and customers. It can be achieved by applying several methods in preprocessing, which converts the speech to the vector, and in modeling, by which the robot learns how to talk. This paper illustrates those methods.

II. Preprocessing

A. Hand-crafted Features

The utterance feature (UF) [2, 3] is introduced to discretize a variable which represents each speech of speakers. We made several standards to assort sentences into utterance features. Table 1 shows groups of criteria for classifying utterance features. 'Menu', 'Price' and 'Option' are conditions which categorize sentences whether they include some of menu, information about price or optional choices. The word 'Action' signifies activities like paying with the credit card or signing. Furthermore, 'Complete/Incomplete' of conditions' Group 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.

This work was partly supported by the Institute for Information & Communications Technology Promotion (R0126-16-1072-SW.StarLab), Korea Evaluation Institute of Industrial Technology (10044009-HRI.MESSI, 10060086-RISF), and Agency for Defense Development (UD130070ID-BMRR) grant funded by the Korea government (MSIP, DAPA).

Table 1. 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

B. Bag of Words

The Bag-of-words (BoW) model [1] is used for feature extraction of spoken sentences. In the BoW model, a dictionary T is constructed from all of the distinct words in the corpus. Then a sentence is converted to the feature vector where each component is the number of occurrences of each word in T.

C. doc2vec

A doc2vec [4] is a method to convert a document to a vector. One is given by genism (genism.models.doc2vec) in Python. The doc2vec separates words by spaces, which leads to differentiating the same words due to different postpositions in Korean. Therefore, each sentence was converted to morphemes before applying the doc2vec.

III. Model

Cafebot's conversational model learns the intention and response represented in vectors. The model has several candidates for itself. Some of them can be applied to both the intention and response, but others only to the intention.

A. Nearest Neighbors

Nearest neighbors (NNs) [4] are one of primitive types of classifiers. NNs can be used to classify the speaker's intention. The k-NNs algorithm determines the class of the input instance, a sentence in conversation, by the majority of the classes of the k previously seen data nearest to the input. The data should be memorized, and they need to be compared to each input, so it may have high memory and time complexity.

B. Hidden Markov Models

Hidden Markov models (HMMs) [2, 3] are used to model sequential data, where some types of data are observable, but

others are not. HMMs can be used for both the intention classification and the response generation. Training or inference algorithms for HMMs are well known.

C. Support Vector Machines

Support vector machines (SVMs) [1] are a very powerful classifier, and they were actively used in the early 2000s. The most basic form of them is a linear SVM, which maximizes the margin between two classes that are linearly separable. With the kernel trick, SVMs could be applied to broad ranges of data.

D. Hypernetworks

Hypernetworks (HNs) [1] are an extension of graphs. A graph has edges that are a (ordered or unordered) pair of vertices, whereas an HN has edges that are an n-tuple of vertices. Therefore, a graph is a special case of the HN when $n = 2$. HNs are a kind of n-gram models and probabilistic graphical models. They can be used for both the classification and generation.

IV. Results

Through the above processes, we could achieve high accuracy of intention classification and generate responses fairly well. To quantify the generation results, we estimated its accuracy and had people evaluate the naturalness.

A. Intention

The HMM based on hand-crafted features achieved over 90% accuracy in the intention classification (Fig. 1, [3]). The 1-NN based on the doc2vec and the SVM based on the BoW showed 77.4% and 80.14%, respectively.

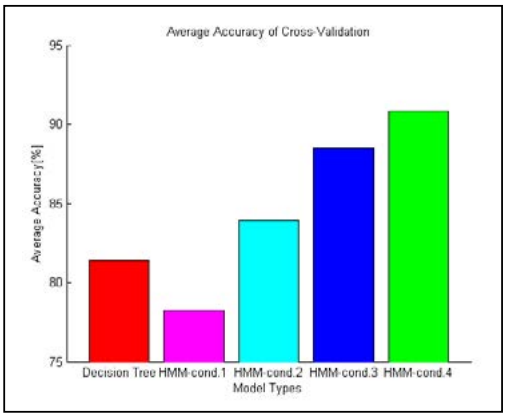

Fig. 1. The accuracy of the intention classification by the decision tree and the HMM with various types of hand-crafted features [3].

B. Response

The accuracy of response by treating the generation as the classification was over 50% when the training data were 110 episodes among 130 episodes in total (Fig. 2, [1]). Intention-based HMM showed naturalness score of 4.47 quite close to 6.52 of the real dialogue when the training data were 80% of the whole data (Fig. 3, [3]).

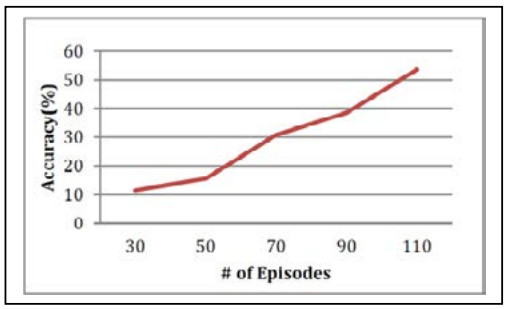

Fig. 2. The average naturalness of dialogues evaluated by 9 people. Control is the real dialogue. Intention-based HMM reflects the intention in its generation and uses the sentence as observation. UF-based HMM uses the hand-crafted utterance feature as observation [1].

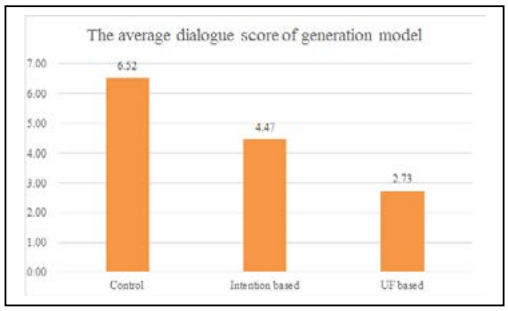

Fig. 3. The average naturalness of dialogues evaluated by 9 people. Control is the real dialogue. Intention-based HMM reflects the intention in its generation and uses the sentence as observation. UF-based HMM uses the hand-crafted utterance feature as observation [3].

V. Conclusions

Cafebot has conversations with customers in cafes. It was accomplished by several methods to model the intention and the response.

References

[1] J.-H. Oh, H.-S. Chun, and B.-T. Zhang, "Generating cafeteria conversations with a hypernetwork dialogue model," in *Proc. the 14th International Symposium on Advanced Intelligent Systems (ISIS 2013)*, pp. 1424-1435, 2013.

[2] S. Lee, E.-S. Kim, and B.-T. Zhang, "Modeling of Speech Intention using the Hidden Markov Model." *Korean Institute of Information Scientists and Engineers 2014 Winter Conference*, pp. 1417-1419, Nov. 2014.

[3] S. Lee, J. Hwang, E.-S. Kim, and B.-T. Zhang, "An adaptive computational discourse system based on data-driven learning algorithm," *The 16th International Symposium on Advanced Intelligent Systems (ISIS 2015)*, 2015.

[4] S.-H. Choi, E.-S. Kim, and B.-T. Zhang, "An Intention Prediction Method for Dialogue using Paragraph Vector," *Korea Computer Congress 2016*, pp. 977-979, Jun. 2016.