Debate Dialog for News Question Answering System 'NetTv'
-Debate Based on Claim and Reason Estimation-

Rikito Marumoto, Katsuyuki Tanaka, Tetsuya Takiguchi, Yasuo Ariki

Abstract NetTv is a question answering system for net news videos with a voice interface. When watching any news, a question answering system can answer a wide range of questions from users for deeper understanding of news. In this study, in order to help users to understand news deeply, when using NetTv, we propose a debate system which can debate with users about the watching news. In the debate system, we estimate the user's claim (approval / disapproval / neither) and reason (presence / absence) using CNN, and use the information to select an opinion stored in debate database and debate by presenting it to the user. In the experiment, eight subjects actually use the system to evaluate naturalness (whether the system response to the user's utterance was natural), deepness (whether the discussion was established), and user satisfaction (whether new knowledge was obtained) in 5 stages. From the result of subjective evaluation, we show that the debate system gives new knowledge and clarify the present problem of the debate system.

1 Introduction

In recent years, a system using a voice interface such as Siri developed by Apple Computer, Inc. attracts wide attention[1, 2]. NetTv developed by us is a question answering system using video indexing and voice interface in net news videos. That is, it aims at alleviating user's search burden and improving news knowledge by constructing a system that enables users to comfortably watch videos and solve the questions that occurred during watching news videos.

In this study, in order to help users to understand news deeply, we are developing a dialog system which can debate with users about any news. Several studies related to the debate have been conducted[4, 5], but in the debate system we develop,

Rikito Marumoto, Katsuyuki Tanaka, Tetsuya Takiguchi, Yasuo Ariki

Department of Computer Science and Systems Engineering, Graduate School of Engineering, Kobe University, e-mail: takigu@kobe-u.ac.jp
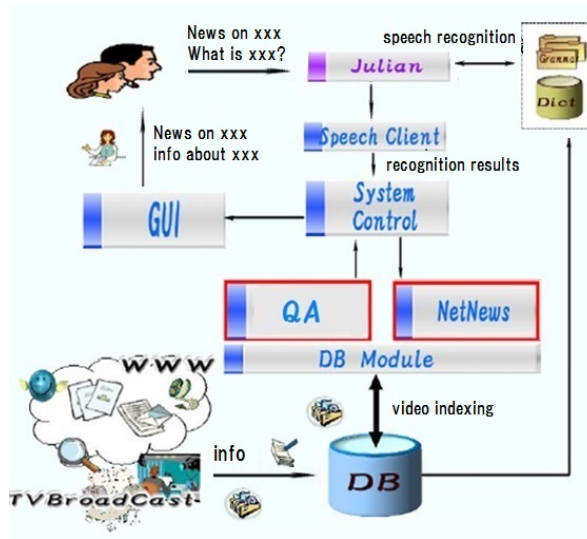
Fig. 1 NetTv architecture

the user's claim (approval/disapproval/neither) and reason (presence/absence) are estimated through the dialog. Thereafter, our system presents its opinion against the user's claim by retrieving approval or disapproval opinions from the debate database for the discussing agenda. In the system, CNN model[3] is used for claim and reason estimation.

2 Overview of NetTv

NetTv driven by users aims to retrieve information and answer questions from users via speech recognition in a dynamically changing environment day by day. As shown in Fig. 1, the NetNews module and the QA (Question Answering) module play a key role.

In the NetNews module, text information on the website is used as meta-information for the news being broadcast on the net, and by indexing the metadata to the news, voice search is enabled. In order to keep up with rapid changes of news content, the links and detailed articles are automatically collected from news sites. Then the index tables of URLs, movies, headers and words of detailed articles are created. Using these four index tables, voice search is achieved by using keywords.

In the NetTv system, to answer a wide range of questions from users, a question answering module is constructed. In the QA module, after question type is estimated to the question from the user, corresponding processing for each question type is performed and the answer is returned to the user. We prepared four question types; *factoid type*, *definition type*, *why type* and *how type*. For the *factoid type* question,
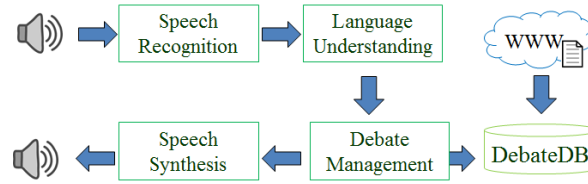
Fig. 2 Debate system architecture

answers are retrieved using Knowledge Q & A API published by NTT docomo. For the *definition type* question, the summary sentence described in Wikipedia is extracted and returned to the user as an answer. For the *why type* question, the answer is extracted based on existing study[6] and returned to the user. NetTv does not implement *how type* question answering at present.

In this paper, we extended the NetTv to have a debate function which enabled it to debate with the user on an agenda chosen by the user. In this paper, we focus on this debate system.

3 Debate System

Debate is contention in argument where opinions are in conflict with each other. Therefore, in this study, as shown in Fig. 2, the LU (Language Understanding) module estimates the user's claim (approval/disapproval/neither) and reason (presence/absence) on an agenda that the user selected before starting the debate. Thereafter, the DM (Debate Management) module decides the opinion that is against the user's claim (disapproval if the user's claim is approval) and debates with the user by presenting it. In order to present the opinion to the user, approval and disapproval opinions on any agenda are automatically collected from WWW in advance, and the debate database is constructed.

3.1 Language Understanding

The LU module estimates the user's claim (approval / disapproval / neither) and reason (presence/absence). In this study, it is estimated based on the CNN model proposed by Shi[3]. Fig. 3 shows the CNN model architecture. The input data are news Tweets of the Sankei newspaper official account and the paired Replies to them acquired by Twitter API. In the CNN model, each word in Tweet and Reply is vectorized by word2vec learned with Wikipedia. Then, the sentences of Tweet and Reply are represented as embedding matrix respectively, where the embedded word vectors are arranged in the appearing order in the sentence as shown in Fig. 3. On these two matrices, CNN is applied. CNN consists of a convolution layer, a
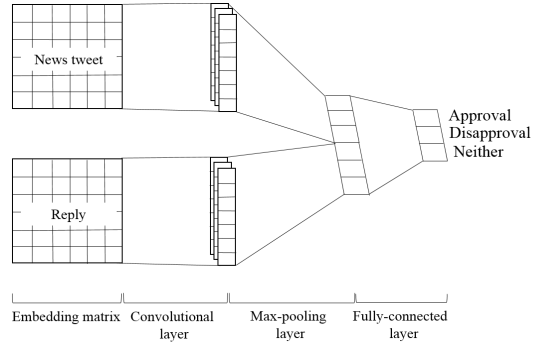
Fig. 3 CNN model architecture for claim estimation

max pooling layer, and a fully connected layer for each tweet and reply matrix, and concatenates reply and tweet information in max pooling layer. Thereafter, fully connected layer computes posterior probabilities of three claim classes (approval / disapproval / neither) or two reason classes (presence/absence) through softmax function.

3.2 Debate Database

Debate database is constructed through a two-stage search method; initial search and re-search. In the initial search, when X is an agenda, a sentence "I agree (disagree) with X" is fed to the search API and web pages containing the expression "I agree (disagree) with X" are collected. Then texts are extracted from the collected web pages, and the query sentence as well as the proceeding and succeeding sentence are grouped as an opinion candidate. Here, as a query sentence used in the initial search, a number of synonymous expressions are used in addition to the expressions mentioned in the example.

In the re-search, frequent words are extracted as related words from the set of opinion candidates obtained in the initial search. The extracted set of related words is added to the end of "X agree (disagree)" and the web pages are collected by the extended query given to the search API. The text is extracted from the collected web pages, and a passage including three or more related words is obtained as an opinion candidate. The initial search and re-search as described above are implemented for approval and disapproval respectively. In this way, opinion candidates are collected.

However, since the opinion candidates obtained in the re-search are poor in quality, by filtering it, only high-quality opinions can be selected. As such a filter, we constructed SVM classifier that classifies approval and disapproval, using high-quality opinion candidates obtained by initial search as SVM training data. Each opinion candidate obtained by initial search and re-search is fed to the constructed classifier, and the opinion candidates with score near the middle between the ap-

proval and disapproval are omitted, and only high-quality opinion candidates are stored in the database. In addition, the obtained approval and disapproval opinion candidates are individually clustered using LDA (Latent Dirichlet Allocation), and each opinion class is stored in a database.

3.3 Debate Management

The debate management module generates the system utterance. In LU module described in Sect. 3.1, when the user's claim is estimated to be "neither", or when the reason is estimated to be "absence", the system generates an utterance that asks the user about a claim.

When the system estimates the user's claim and reason, the system selects an opinion from the debate database, which is against the user's claim. For example, if the user's claim is estimated to be "approval", the system utterance is selected from the disapproval opinion stored in the debate database. As for the selection method, except for the opinions that were selected in the past, a disapproval opinion whose cos similarity is high (not low for the sake of weak disapproval) with the user's utterance is selected. In addition, by presenting the opinion in a topic (class) different from the topic (class) presented to the user in the past, it is possible to present a wide range of opinions to the user.

4 Evaluation

First, classification experiments of claim(approval / disapproval / neither) and reason(presence / absence) are described. The training data are news Tweets of the Sankei newspaper official account and the paired Replies to them acquired by Twitter API. The number of training data in the claim classification is 1486 sentences (approval: 458, disapproval: 831, neither: 197), the number of training data in the reason classification is 1301 sentences (absence: 1083, presence: 218). In both classification experiments, 50 data were used for each class. The results are shown in Table 1. Accuracy was 80% in both the claim and the reason classification.

Table 1 Experimental result (accuracy(%)) in the claim(approval / disapproval / neither) and reason(presence / absence)

	Claim	Reason
CNN	80.0	80.0

Then, a subjective evaluation experiment was conducted on the agenda items of "Casino bill", "Capital punishment" and "Resuming nuclear power plant operations". In the subjective evaluation experiment, eight subjects used the debate system and evaluated naturalness (whether the system response to the user's utterance

was natural), deepness (whether the discussion was established), and user satisfaction (whether new knowledge was obtained) in 5 stages. Fig. 4 shows the results of subjective evaluation experiment. The results of naturalness and deepness are 2.24 and 2.38 points, respectively. The result of the user satisfaction shows three points, which is higher than the other evaluation indices, indicating that the user gains a certain degree of knowledge. In order to further improve user satisfaction, it is necessary to increase the number of opinions of DB and improve the quality of opinion.
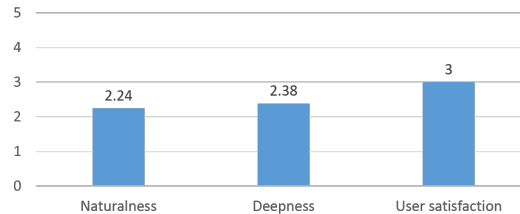
Fig. 4 Experimental result of subjective evaluation

5 Dialog example

Below is an example of a dialog with the system in which the user approves of a casino bill that is on the agenda (Japanese sentences have been translated into English).

System : What do you think about the casino bill?
User   : I approve of the bill.
System : Please tell me why.
User   : Since the bill would approve not just casinos, I think it would be good if it also approved the construction of recreational facilities, such as amusement parks.
System : Personally, I'm opposed to the casinos. I don't know what they will be built for. If we want to get more foreign currency, we should focus on sightseeing. However, if Japanese people aren't allowed to enter the casino, I think it would be all right to build them.
User   : In fact, there are cases overseas in which casinos have been successful, so the construction of casinos would lead to an increase in the number of foreign tourists and also promote the economy.
System : There are already many places that provide hospitality and meet the needs of tourists from overseas. Japan is a country that prohibits and criminalizes gambling. I disapprove of the casino bill, as I've stated above.
User   : Even if casinos are prohibited, I think it is necessary to increase the

                number of tourists and promote economic growth for a prosperous
                future.

System : "Atlantic Club Casino" went out of business there were too many
                casinos and because the number of tourists who don't gamble increased.
                As a result, many workers whose jobs were related to the casino were
                laid off, and eventually the regional economy collapsed.
                  (Several phrases are omitted due to the length of the sentences.)

User   : Is that so ... ?

System : There are also the following opinions. The government's purpose in
                allowing casinos is not the expected tax revenue, but the increase in the
                number of tourist and the resulting stimulus to the tourism industry and
                to regional development. Tax revenue from casinos is small, totaling 2-3
                trillion yen at most, and the return rate is 97- 98%, isn't it? If you take
                into account the tax incentives that will be given to the casinos, the tax
                revenue will be insignificant, I think.

As shown in the dialog example above, there may be a gap between the user's utterance and the system's response to it. One reason is that the number of opinions stored in the DB is small, and it is also necessary to devise how to collect high-quality opinion candidates. Since the opinion presented to the user is obtained from the web, the debate system has some possibility to present an opinion containing the wrong information to the user. This problem must be dealt with carefully in near future.

6 Conclusion

In this paper, we proposed a debate system in order to help users to understand news deeply in the news information retrieval and question answering system "NetTv". In the debate system, we first estimate the user's utterance claim (approval/disapproval/neither) and reason (presence/absence) using the CNN model. Thereafter, we developed a debate system which could present an opinion against the user's claim by searching from the debate database where opinions are stored by extracting from the Web for any agenda. From the result of subjective evaluation, we show that the debate system gives new knowledge and clarify the present problem of the debate system.

References

1. Lori Lamel *et al.*. User evaluation of the mask kiosk. *Speech Communication*, pp. 131-139, 2002D
2. Koichiro Yoshino and Tatsuya Kawahara. News Navigation System based on Proactive Dialogue Strategy. *Int'l Workshop Spoken Dialogue Systems*, 2015D

3. Hongjie Shi, Takashi Ushio, Mitsuru Endo, katsuyoshi Yamagami and Noriaki Horii. A Multichannel Convolutional Neural Network For Cross-language Dialog State Tracking. *IEEE Workshop on Spoken Language Technology*, 2016D

4. Misa Sato, Kohsuke Yanai, Toshihiko Yanase, Toshinori Miyoshi, Makoto Iwayama, Qinghua Sun, Yoshiki Niwa. End-to-End Argument Generation System in Debating. ACL-IJCNLP, 2015.

5. Peter Potash and Anna Rumshisky. Towards Debate Automation: a Recurrent Model for Predicting Debate Winners. *Empirical Methods on Natural Language Processing(EMNLP)*C 2017D

6. Ushio Shibusawa, Takahiro Hayashi and Rikio Onai. Development and Evaluation of a System for Extracting Answers of a "'Why' Type Question from the WEB. *Information Processing Society*Cpp.1512-1523, 2007(in japanese)D