EVALUATION METHOD OF NON-TASK-ORIENTED DIALOGUE SYSTEM BY HMM

Naoki ISOMURA, Fujio TORIUMI, Kenichiro ISHII

Graduate School of Information Science, Nagoya University

ABSTRACT

Recently, computerized dialogue systems have been actively investigated and used in various fields. In order to realize a practical system, the performance of the system should be evaluated quantitatively. An objective and quantitative evaluation method for task-oriented dialogue systems, such as reservation services, has already been established; however, non-task-oriented dialogue systems have been evaluated only by subjective methods like questionnaires. In this paper, we propose a new criterion that can evaluate non-task-oriented dialogue systems objectively and quantitatively. We assume that a human-human dialogue is an ideal dialogue. We design an HMM (Hidden Markov Model) by learning a sequence of human-human dialogue utterance tags that are automatically assigned. We apply n-gram for auto-tagging and evaluate the humanness of dialogues using HMM. In this simulation, the rate of correct auto-tagging is 54%. If we consider partly correct tags as completely correct tags, the correct rate becomes 82%. Furthermore, it was clarified that the proposed method based on HMM can evaluate the humanness of a dialogue.

1. INTRODUCTION

Recently, computerized dialogue systems have been actively investigated and used in various fields. Not only task-oriented dialogue systems like reservation services but also non-task-oriented dialogue systems like chattering are highly anticipated. The purpose of this study is to design a dialogue computer that simulates a human interviewer. A professional interviewer always tries to please an interviewee. Before starting a dialogue, he/she fully investigates the interviewee and prepares questions accordingly. We call this type of dialogue 'interview type' dialogue. Therefore, our aim is to design a dialogue system that can realize an interview type dialogue. The system prepares a mass utterance collection (script collection) before starting the dialogue in order to better please the human interviewee.

Toward this aim, we must solve two problems:

- Making a large number of scripts automatically, and
- Selecting an appropriate utterance from the scripts automatically.

In order to solve these problems and to realize a practical system, the performance of the system should be evaluated quantitatively. Objective and quantitative evaluation methods for task-oriented dialogue systems have already been established[1]. For example, task achievement rate or the time required for task accomplishment can be used as criteria. On the other hand, non-task-oriented dialog systems have been evaluated only by subjective methods like questionnaires.

Since non-task-oriented dialogue sytems have no clear purpose, it is difficult to evaluate a dialogue objectively and quantitatively. Sogabe and Ishii proposed a new criterion that can evaluate non-task-oriented dialogue systems objectively and quantitatively [2]. The criterion utilizes the similarity between a human-human dialogue and a human-machine dialogue. Their evaluation method requires a considerable amount of time and effort because utterance tags must be assigned manually to both human-human dialogues and human-machine dialogues.

We take only written dialogues as the object of our study rather than spoken dialogues. In this paper, we apply n-gram for auto-tagging and build an HMM (Hidden Markov Model) by learning a sequence of human-human dialogue utterance tags. Then we evaluate the humanness of a dialogue using HMM.

2. DIALOGUE EVALUATION

In order to design a dialogue computer that can provide pleasant conversation to humans, we examined how a professional interviewer please an interviewee.

A professional interviewer fully investigates an interviewee before conversation and then tries to please the interviewee during the conversation.

For example, an interviewer obtains information about an interviewee from web pages or from books before the interview. Moreover, he/she tries not to interrupt the interviewee's utterances, to show understanding, and to avoid using YES/NO type questions in favor of 5W1H type questions.

In order to verify the effectiveness of an interview type dialogue, we checked the differences in pleasantnesss of dialogues between the convasations of an interviewer and a non-interviewer. The data discussed below were collected in the following way.

Table 1. Pleasantness of dialogues

	average dialogue pleasantness
`general` type	3.50
`interview` type	4.00

1. Before interview, an interviewee writes his/her profile
2. Interviewer reads the interviewee's profile
3. Interviewer converses with interviewee for 30 minutes through text chat
4. After interview, interviewee gives scores reflecting the dialogue's pleasantness

Table 1 presents the averages of dialogue plesantness, showing that interview type dialogue was evaluated more highly than general dialogue.

We could evaluate plesantness using thorough questionnaires, but we needed an objective and quantitative evaluation method for dialogue systems.

3. PROPOSED METHOD

3.1. Outline of method

We propose an evaluation method for dialogue humanness. We assume that human-human dialogue is an ideal dialogue. If a dialogue has high similarity to human-human dialogue, we consider the dialogue to have high humanness. We used dialogue to build an utterance collection during a conversation. Then, we formed an HMM of human-human dialogue to calculate the humanness of a dialogue. If the probability calculated by the obtained HMM is high, we regard the dialogue as close to a human-human dialogue.

In using an HMM, it is important to determine what kind of output symbols we should use. In this simulation, we used special tags that correspond to utterances for these symbols. The SWBD-DAMSL (Switchboard Discourse Annotation and Markup System of Labeling) tag is available to describe a type of utterance. This tag is used for utterance type descriptions, for example, Yes-No-Question, Statement-opinion, and so forth. It is possible to express a utterance meaning by tagging one or more SWBD-DAMSL tags. We used a simplified DAMSL tag, but it is difficult to assign simplified DAMSL tags to many utterances. Thus, an automatic tagging technique was employed using a small amount of standard data that was tagged manually.

We used the following procedure in the proposed evaluation method. First, we designed an HMM by learning human-human-dialogue (steps 1, 2, 3). Next, we evaluated a dialogue using HMM (steps 4, 5).

Table 2. Simplified DAMSL Tags

(1) Uninterpretable, (2) Self-talk,
(3) 3rd-party-talk, (4) Statement,
(5) Question, (6) Directive,
(7) Influencing-addressee-fut-actn,
(8) Committing-speaker-future-action,
(9) Other-forward-function,
(10) Thanking, (11) Apology, (12) Agreement,
(13) Understanding, (14) Other

1. Assign simplified DAMSL tags to utterances manually (standard data)
2. Assign simplified DAMSL tags to learning dialogues using standard data
3. Determine HMM parameters by using leaning dialogues
4. Calculate HMM output probability of a dialogue to be evaluated
5. Evaluate the dialogue by HMM output

3.2. Assigning simplified DAMSL tags

In this research, we assigned SWBD-DAMSL tags to 48 dialogues (3590 utterances). The number of tag combinations was 473. This means that there were 7.6 utterances per tag combination, so the data were not large enough to use as standard data. Consequently, we used simplified DAMSL tags for expressing utterance meanings. Simplified DAMSL tags are obtained by classifying SWBD-DAMSL tags into a smaller number of tags. Table 2 shows the simplified DAMSL tags, which include 14 tags in all. All utterances have one or more simplified DAMSL tags. There are tags such as "Answer + Statement" , "Agreement + Statement", and so on.

Human-tagging has a high accuracy, but it is difficult to assign tags to a large number of utterances. For this reason, we use auto-tagging of simplified DAMSL tags.

First, we assign simplified DAMSL tags to utterances manually, and we call this standard data. Second, we calculate the similarity between a given utterance and every utterance in the standard data and then assign the same tag of the utterance that has the highest similarity. We use the 2-gram of words for calculating similarity.

When we define the collections of 2-grams included in two dialogues as A and B respectively, the similarity S of two dialogues can be calculated by the following procedure.

$$S = \frac{2 \times |A \bigwedge B|}{|A| + |B|} \quad (| \ | \ numberof2-gram)$$

Table 3. Conditions on an experiment

the number of dialogue	48
the type of tag	simplified DAMSL tag
the number of tag combination	89
evaluation method	leave-one-out cross-validation

4. EXPERIMENT OF AUTO TAGGING

4.1. Experimental methodology

The purpose of this experiment is to show the effectiveness of n-gram auto-tagging. Table 3 shows the conditions of the experiment, in which we used 48 dialogues (3590 utterances). These dialogues are not speech dialogues but text dialogues, and they included human-human dialogues and human-machine dialogues. We used three dialogue computers: Arisa [3], Sixamo [4], and our system, KELDIC (Ken's Laboratory Dialogue Computer) .The dialogue computer KELDIC is based on ELIZA [5], which uses predesigned scripts to reply. Arisa also uses scripts, while Sixamo uses human-human dialogue data and Markov chains to reply.

We used the leave-one-out cross-validation method to evaluate the auto-tagging. In other words, we divided 48 dialogues tagged by hand into 47 standard data and 1 test data. We assumed test data as unknown and performed auto-tagging, and then compared the results to human-tagging. We modified the dialogue that became test data, repeated 48 times, and evaluated the auto-tagging by the number of correct taggings.

4.2. Results

Since the correct tags of all test data are determined, we compared auto-tagging tags with the correct tags. As a result, 54% of utterances were assigned completely correct tags and 82% were assigned at least partly the correct tags. "Partly correct" means that an utterance was assigned a part of correct tags. For example, if the "Question" tag is assigned to an utterance whose correct tag is "Question + Statement", the tagging is partly correct.

The 2-gram is used in this method, so in the case that the standard data have no 2-gram utterances, auto-tagging fails. If the test data contain a small amount of words, the same utterance may have various tags; accordingly, additional methods, not only 2-gram, would be needed.

Errors occurred due to the following reasons:

- Standard data has no 2-gram utterances
- The same utterances in the standard data have different tags

The first problem will be solved by increasing the standard data. The second problem will be solved by a technique that uses the context of dialogue.

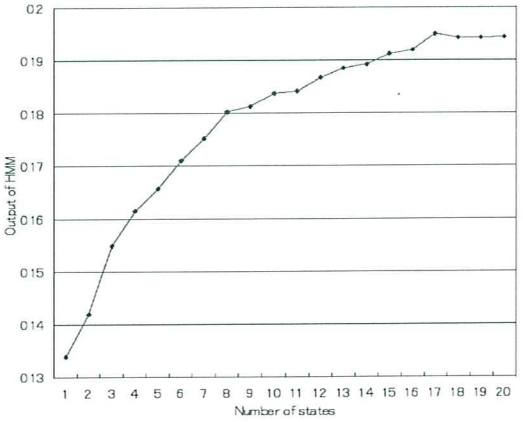

Fig. 1. Output and the number of states of HMM

Table 4. Conditions of HMM

number of states	20
output symbol	simplified DAMSL tags
dialogue for learning	197 human-human dialogues
input	tag sequence

In this experiment, the number of tagged human-human dialogue data is 6. We applied an approach based on adaptive networks [6] to take the context information into consideration. Using the leave-one-out method, 62% of utterances were assigned completely correct tags and 91% were assigned at least partly correct tags. This experiment demonstrates that we can improve tagging performance by using context information.

5. EXPERIMENT OF DIALOGUE EVALUATION BY HMM

5.1. Experimental methodology

We designed the HMM by learning human-human dialogue and then investigated the difference in HMM output between a human-human dialogue and a human-machine dialogue. Table 4 shows the condition of the HMM. We see from Figure 1 that the appropriate number of states is 20. We used the 2-gram method for auto-tagging of dialogues for HMM learning. We designed an HMM that is adapted to human-human dialogue using these dialogues as a multiple-observation sequence [7].

In this experiment, we used three dialogue computers: KELDIC, Arisa, and Sixamo. These are not natural compared with humans.

We examined the HMM output to discover whether it was possible to evaluate the humanness of dialogues.

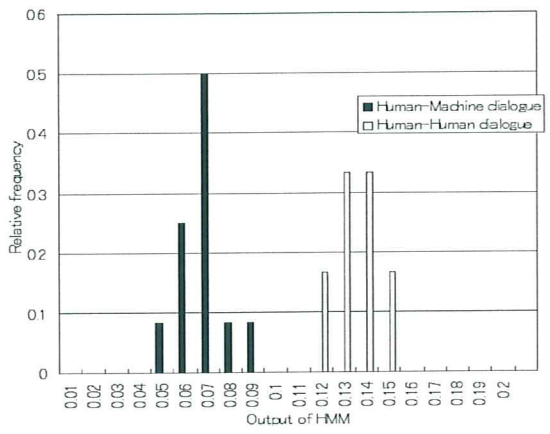

Fig. 2. Output by HMM

5.2. Results

Figure 2 shows the outputs of human-human dialogues and human-machine dialogues by HMM. The input dialogues are different from the dialogues used for learning.

Figure 2 indicates that the outputs of human-machine dialogues are lower than 0.1, while human-human dialogues are higher than 0.1. Consequently, the HMM that is adapted to human-human dialogue represents natural dialogue. Since the outputs of human-machine dialogues are low, they are unnatural. In the case of ordinary human-human dialogue, an utterance tagged "Question" is probably followed by the utterance "Statement". In the case of human-machine dialogues, however, a human utterance is followed by an unnatural machine utterance, and an unnatural machine utterance is followed by a meaningless human utterance. We concluded that dialogues can be evaluated by using an HMM that is optimized for human-human dialogues.

In this method, we interpret a dialogue as a simplified DAMSL tag sequence. Thus, the difference between the tag sequence of HMM learning and the tag sequence of the input dialogue is essential to the output. The output of HMM will be low when

- Tagging is not sucessfuls or
- The sequence of tags is much different from that of human-human dialogues.

6. CONCLUSION

In this paper, we propose a new criterion to evaluate non-task-oriented dialogue systems objectively and quantitatively. We assumed a human-human dialogue is an ideal dialogue. We designed an HMM by learning a sequence of human-human dialogue utterance tags that were automatically assigned. We applied n-gram to auto-tagging and evaluated the humanness of dialogues using HMM. In this simulation, 54% of utterances were assigned completely correct tags and 82% were assigned at least partly correct tags. Therefore, if we relaxed semantic conditions, the rate of correct tagging would be 82%. Furthermore, it was clarified that an HMM method is effective for evaluating the humanness of a dialogue.

There are two open issues that need to be further explored in the future. First, the auto-tagging by 2-gram has a problem: Although the context information of a dialogue is important, the 2-gram method did not take it into consideration. Thus, also using the previous utterance for the auto-tagging would be effective.

Second, in this paper, the output symbols of HMM were simplified DAMSL tags; however, there may be other useful symbols. Therefore, in addition to Using simplified DAMSL tags to evaluate the humanness of a dialogue, using other symbols may help to evaluate other features.

When we design dialogue systems to be pleasant for humans, the humanness of a dialogue is an inportant subject. In addition, the pleasantness of a dialogue is also important. We expect that it is possible to evaluate pleasantness by using the appropriate observation sequences for HMM.

7. REFERENCES

[1] M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, pp. 271–280, 1997.

[2] Masayoshi Sogabe, Fujio Toriumi, and Kenichiro Ishii, "An evaluation method of non-task-oriented dialog system," *IPSJ SIG Technical Report*, vol. 2005-NL-170, pp. 105–110, 2005.

[3] http://www.nagisanet.com/cgi/index.htm

[4] http://yowaken.dip.jp/sixamo/sixamo.rb.html

[5] J. Weizenbaum, "ELIZA-a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[6] A.L. Gorin, S.E. Levinson, L.G. Miller, A.N. Gertner, A. Ljolje, and E.R. Goldman, "On adaptive acquisition of language," *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 601–604, 1990.

[7] S.E. LEVINSON, L.R. RABINER, and M.M. SONDHI, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.