NAIST-IS-DD1461201

Doctoral Dissertation

Building Open-domain Conversational Agent by Statistical Learning with Various Large-scale Corpora

Hiroaki Sugiyama

August 8, 2016

Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Hiroaki Sugiyama

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Associate Professor Masahiro Araki	(Kyoto Institute of Technology)
Assistant Professor Koichiro Yoshino	(Co-supervisor)
Assistant Professor Sakriani Sakti	(Co-supervisor)

Building Open-domain Conversational Agent by Statistical Learning with Various Large-scale Corpora*

Hiroaki Sugiyama

Abstract

We present our work on developing open-domain conversational dialogue agents that make rapport with users through conversation. The wide variety of dialogue topics and action types of user utterances make it difficult to develop such agent. Conversational agents are required to control the dialogue flows and to respond to open-domain user utterances and questions about agent's personality. Besides, it is desired to evaluate the developed agents without manual annotations that require huge cost. This thesis discusses following four components to realize such conversational agent. First, we propose dialogue control methods that automatically estimates the appropriateness of agent actions on the basis of real dialogues between users. Second, we propose a novel utterance generation method that simultaneously realizes both the suppression of irrelevant agent utterances and automatic expansion of conversation topics. Third, we develop a question-answering system for specific personality questions about the agent on the basis of corpus-based approach with large-scale personality database. Finally, we proposed automated and replicable evaluation method for conversational agents using large-scale multi-references.

Keywords:

Conversational systems, open-domain, personality questions, automatic evaluation, inverse reinforcement learning

*Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1461201, August 8, 2016.

大規模多種コーパスの統計学習に基づく雑談エージェントの開発*

杉山 弘晃

内容梗概

本研究は，ユーザと自然な雑談を行うことでユーザとの関係性を構築する，雑談エージェントに関する研究である．雑談機能は，エンタテインメントやカウンセリング目的のみならず，スムーズにタスクを達成したり，使い勝手の良い対話エージェントを実現する上でも重要である．雑談において，ユーザは対話相手の嗜好や経験（パーソナリティ）を質問したり，それに基づく非常に幅広いトピックの自己開示発話を行う．雑談エージェントがこれらのユーザ発話に適切に応答するには，挨拶や質問といった対話行為の流れを適切に制御しつつ，ユーザのパーソナリティ質問に回答し，かつユーザ発話の話題に関連した発話を生成する必要がある．さらに，構築した雑談エージェントを評価する上では，高コストな人手の主観評価に依らず，自動的に低コストで評価できることが望ましい．本研究では，上記の要求に対し，一問一答形式の応答生成に対象を絞り，以下の 4 つの要素技術について論じる．1 つ目は，ユーザの対話行為に対してエージェントが出力すべき対話行為を適切に推定する，対話制御に関する研究である．2 つ目は，任意の話題を持つユーザ発話に対して関連する発話を生成するオープンドメイン発話生成に関する研究である．3 つ目は，エージェント自身のパーソナリティを問う質問に対する応答生成に関する研究である．4 つ目は雑談エージェントを自動的に評価する枠組みに関する研究である．これらの技術により，人と自然に対話できる，雑談エージェントの実現を目指す．

キーワード

雑談エージェント，オープンドメイン，パーソナリティ質問，自動評価，逆強化学習

*奈良先端科学技術大学院大学 情報科学研究科 博士論文, NAIST-IS-DD1461201, 2016 年 8 月 8 日.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1. Background

Dialogue is an important and natural activity for human-beings. Human-beings are social animals and naturally talk to each other for the purposes not only of achieving actual benefit such as information exchange or consensus building, but also of establishing and maintaining social ties with the dialogue partners [Scott Thornbury, 2006, Laver, 1975, Cheepen, 1988, Eggins and Slade, 1997]. When we talk with dialogue partners who have social ties, we can feel the connection and comfort with them; this condition or phenomenon is called as *rapport* in social psychology [Huang et al., 2011]. In this thesis, we call a dialogue that focuses on the former purpose (achieving actual benefits) as *task-oriented dialogue*, and the latter (establishing social ties) as *conversational dialogue*, or *conversation.* Figure 1.1 represents an example of conversation between humans. This illustrates that they exchange some information about themselves to establish social ties, instead of achieving actual benefits from the information. In this way, conversation plays an important role in forming the solidarity of our society [Eggins and Slade, 1997] and is crucial for establishment and maintenance of such rapport [Bickmore and Cassell, 2000]. Here, like a small talk before business negotiation, both types of dialogues can appear in a single dialogue and they are switched according to the dialogue procedure.

Also for dialogue agents that talk with people, conversation is important to establish rapport. It is reported that users frequently try to have conversation for establishing *rapport* even with a dialogue agent that is developed for the purpose of achieving tasks

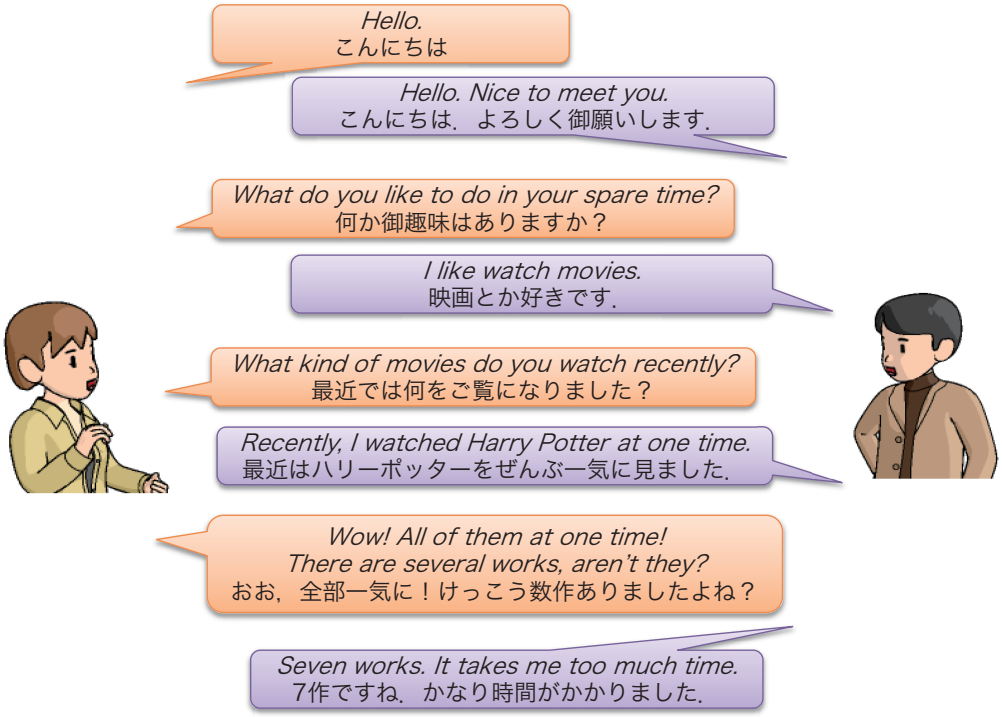


Figure 1.1. Example of conversation between humans

such as navigation or the presentation of regional information [Takeuchi et al., 2007]. Such rapport is effective to both provide users with comfortable feelings and improve task-achievement performance. Examples of the advantages are as follows.

- Users have comfortable and relief feelings through conversation when they are attentively listened [Bickmore et al., 2005]. This not only makes users want to talk with such agents [Meguro et al., 2010], but also provide relief from psychological disorders such as PTSD [DeVault et al., 2014].

- Conversation is effective for dialogue agents to be trusted by users. For example, a real estate agent robot can make rapport with users, which is effective to improve task performance [Bickmore and Cassell, 2001].

- Conversation makes users to have attachment to dialogue agents. Users tend to use such attached dialogue agents longer than non-attached one [Vardoulakis et al., 2012].

- Agents can acquire users' unconscious preferences through conversation. Such preferences enables the agents to provide much more appropriate recommendation for users [Carberry et al., 1999, Wärnestål, 2007].

To develop a conversational agent that makes rapport with users, it is necessary to appropriately respond user utterances. If the agent cannot respond user utterances, users think that the agents do not want, nor have enough capability, to talk with them; this makes users disappointed and stop talking with the agent. To generate reasonable agent responses, it is required to model user's state, such as user's belief, desire and intentions, derived from the user utterances. Previous work that develop task-oriented dialogue agents in limited domain assume that a user state can be defined manually; i.e., all or the most of user utterances can be mapped to the pre-defined user state [Misu et al., 2011, Williams, 2007, Nakano et al., 2000]. For example, in a flight ticket reservation task, we can define a user state on the basis of the limited number of conditions that are required to achieve the task, such as departure and destination places, schedules and budget. Besides, with this assumption, since the contents or requirements in user utterances are limited in pre-defined area, the range of required agent responses is also limited. This limitation enables agent developers to manually create agent utterances associated with each user state, which are expected to be reasonable and consistent through a dialogue.

On the contrary to such limited domain dialogue, user states in open-domain conversation are difficult to be modeled manually. User utterances in conversation inherently have wide variety of topics and action types (greetings, questions, self-disclosure, etc) [Robinson et al., 2008, Takeuchi et al., 2007, Meguro et al., 2010]. This variety makes it difficult to model the whole space of a user state, especially when we consider discourse relations such as contradiction or entailment between user and agent utterances through the dialogue.

To be feasible the automatic modeling of a user state and generation of agent utterances in open-domain conversation, we focus on the generation of one-turn responses and ignore the history of user utterances. With this limitation, while our agent possibly says utterances irrelevant or inconsistent to the dialogue contexts, we can discuss the definition of a user state and open-domain response generation without regard to the complexity caused from their discourse relations.

In addition, we focus on linguistic information of the dialogues and discusses the

User utterance type	Problems	Fundamental approach
Typical patterns	Various wordings	Estimate user dialogue-acts from various wordings of user utterances and write rules for response generation
Self-disclosure	Wide variety of topics and dialogue-acts (subset of self-disclosure such as *preference* and *habit*)	Create utterances that meet Grice's Maxims (contain new information relevant to user utterances without any irrelevant information). Predict agent dialogue-acts to realize whole sentence of an agent utterance.
Questions	More specific relations are required than self-disclosure	Question-answering with large-scale manually created corpus.

Table 1.1. Problems for each user utterance dialogue-acts

development of text-chat based conversational agents, since we consider that the exchange of linguistic information is necessary to build relationship with users. We believe that our improvement becomes a fundamental clue for the further development of conversational agents that can handle the dialogue contexts and generate consistent responses for open-domain user utterances.

2. Problems and fundamental approaches

Although we focus on the one-turn response generation and ignore the history of utterances, utterance generation in open-domain conversation remains difficult problem because of the too wide range of topics of user utterances. In this thesis, to analyze such utterances, we categorize user utterances on the basis of *dialogue-acts* that represent the functions of the utterances like greetings or questions. We adopt Meguro's categorization of dialogue-acts that are proposed for the analysis of listening-oriented dialogue, which aims to make rapport with users by attentive listening [Meguro et al., 2010]. They defined 32 dialogue-acts shown in Table 2.1, and these are roughly categorized into three types of dialogue-acts shown in Table 1.1: *self-disclosure* and information provision, *question*, and other *typical patterns* such as greetings or acknowledgment. We discuss approaches to respond each type of

Maxim type	Definition
Maxim of Quantity	*Make your contribution as informative as is required for the current purposes of the exchange.*
	Do not make your contribution more informative than is required.
Maxim of Quality	*Do not say what you believe to be false.*
	Do not say that for which you lack adequate evidence.
Maxim of Relation	*Be relevant.*
Maxim of Manner	*Avoid obscurity of expression.*
	Avoid ambiguity.
	Be brief (avoid unnecessary prolixity).
	Be orderly.

Table 1.2. Definition of Grice's Maxims [Grice, 1975]

utterances.

First we consider the response generation for typical pattern type of user utterances. Although this type of utterances don't have contents, the variety of wordings of user utterances is too large to manually describe responses for all the utterances. For suppressing such variety of wordings, dialogue-act estimation is an effective approach. There are much existing work for the estimation of dialogue-acts [Stolcke et al., 2000, Ritter et al., 2010]. If user dialogue-acts are estimated, the agent developer can manually create responses associated with each dialogue-act [Meguro et al., 2011].

Second, we discuss about the agent utterance generation for the self-disclosure user utterances. Since self-disclosure utterances contain the wide variety of topics, it is difficult to generate agent responses with the same manual development approach as the typical pattern type of utterances. In this study, we leverage Grice's Maxims shown in Table 1.2 to define the conditions that are required in the utterance generation [Grice, 1975]. From the *Maxim of Relations*, it is important that generated utterances contain only the information relevant to user utterances. From the *Maxim of Quantity*, considering that the dialogue participants are required to proceed the dialogue, it is also important to contain new information relevant to user utterances to avoid parroting user utterances. From the Maxim of Manner, the generated utterance are required to be brief, so the agent utterances cannot contain information irrelevant or prolixity to the user utterances. On the contrary to the above Maxims, since we focus on the one-turn responses, Maxim of Quality that requires consistency is not important in our objective;

i.e., generated utterances are not required to have evidences of the utterances as long as they are not believed to be false or lie. Taken together, agent response utterances should contain new information relevant to user utterances, with suppressing irrelevant information. In this approach, the appropriateness of generated utterances is limited with the complexity of information that the agent extract from user utterances to be used as a source of agent utterances. If an agent extracts only a word from a user utterance, the relevancy that the agent can calculate is limited with topic-word level, which is not enough to capture *events*, or what the user did or felt. On the other hand, if the agent can deal with predicate-argument structures that represent events rather than topic words, the agent can generate utterances with event-level relevance. This means that the more complex information the agent can deal with, the more appropriately relevant utterances the agent can generate. To generate agent utterances, in addition to the topic information, it is necessary to decide a dialogue-act of the agent utterance. We can utilize the agent dialogue-acts to filter inappropriate candidates of agent utterances, or modify candidate utterances to express the predicted agent dialogue-act with hand-coded rules [Higashinaka et al., 2014]. Therefore, we believe that the prediction of agent dialogue-acts from user utterances is also important problem. In this thesis, we call this problem as *dialogue control*.

Finally, we discuss about the answering for user questions. Questions are roughly categorized into two types: factoid questions consist of *question-information* and *question-fact*, and personality questions consist of the other question types, which ask speaker's habits, experiences or preferences. Since factoid question-answering (QA) has been actively developed and it requires information that are not contained one-turn user utterance, this study focus on the answering for latter personality questions. Such personality questions are frequently asked in the beginning of dialogues that initiates a dialogue topic. This is so natural activity for humans, task-oriented dialogue agents are also asked with personality questions [Takeuchi et al., 2007]. Since personality questions are used as a trigger of dialogues, if agents cannot answer such questions, the dialogue easily stops; therefore, answering for personality questions is a crucial function for dialogue agents. When an agent answers questions, the acceptable range of responses is narrower than responses to self-disclosure utterances; therefore, it requires another approach to capture the meaning of such personality questions and answer them. Since it is difficult to capture the meaning of complex structures of

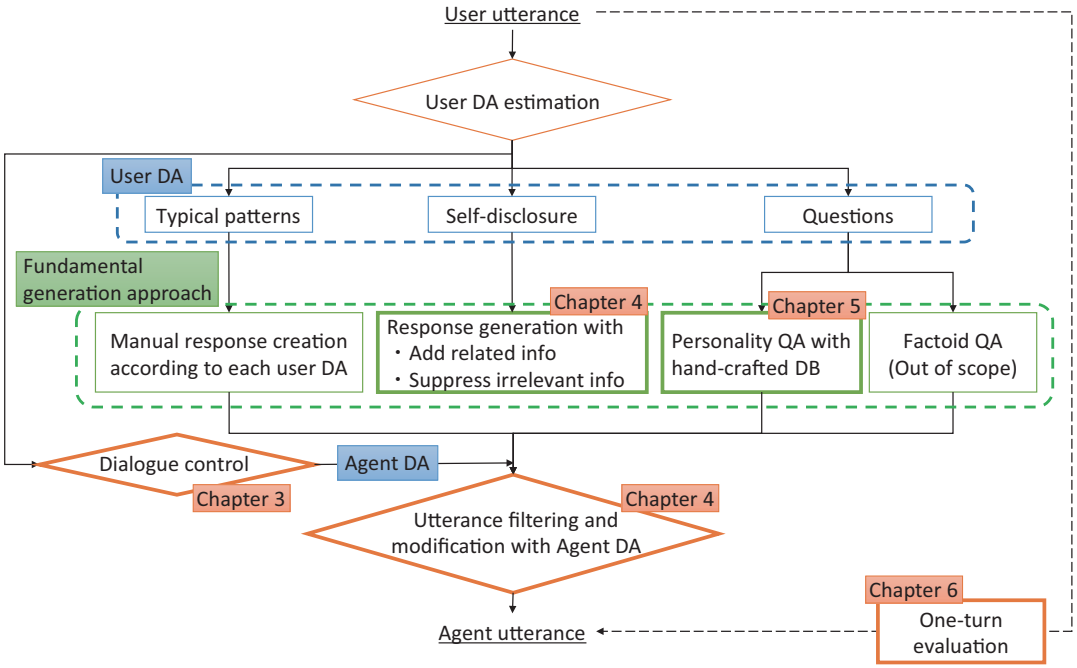


Figure 1.2. Information flow in our approaches. DA means a dialogue-act. Bold lines represent our focuses in this thesis.

user utterances, we adopt the retrieval based question-answering approach previously proposed by Batacharia [Batacharia et al., 1999]. They manually defined question categories that seem to be frequently asked in conversation such as name or present address. This approach makes agents appropriately answer for pre-assumed range of questions; however, the coverage of manually defined questions are not examined. In this thesis, we leverage large-scale personality question-answer corpus to both examine the coverage of personality questions defined with such approach and develop and examine the effectiveness of QA system that covers the wide range of user questions.

From the above analysis, we focus on the following three problems for the agent utterance generation: dialogue-control, response generation for open-domain utterances, and answering for personality questions, which correspond to Chapter 3, 4 and 5. With solving these problems, we can obtain a response utterance for a user utterance in open-domain conversation. Figure 1.2 shows the flow of the response generation. At first, input user utterances are classified into the three types of dialogue-

acts. When the user dialogue act is typical patterns, responses can be generated using rules created by hand. For self-disclosure type of utterance, responses are generated so as to add new information related to the user utterance with suppressing the contamination of irrelevant information. When the user utterance is question, we categorize it into personality questions and factoid questions. If it is a factoid question that is out of our focus, we leverage existing factoid question-answering systems [Kolomiyets and Moens, 2011, Wang, 2006]. Dialogue control predicts agent dialogue-acts, with which we can filter inappropriate candidates of utterances or modify the wordings of the candidates, especially in response generation for self-disclosure type of user utterances.

In addition to these problems, we believe that the evaluation of agent utterances is also an important problem. If we can evaluate the utterances automatically, it is easier to develop and improve the utterance generation methods. Therefore, we also tackle to the development of automatic evaluation methods for conversational agent utterances, which correspond to Chapter 6.

3. Existing works

3.1 Dialogue control

Conversational agents have to determine their appropriate actions for specific user states. This is called *dialogue control*, which is a major topic in dialogue agent research area. Recent studies [Singh et al., 1999, Williams and Young, 2007, Williams and Young, 2005, Williams, 2007, Misu et al., 2012] adopt reinforcement learning (RL) to realize the dialogue control. RL automatically decides the appropriate agent actions for user states, which is called as policy function, in order to maximize the total expected rewards received by a reward function; therefore, if a reward function meets objectives of a dialogue, we can automatically decide the optimal agent actions for the objectives.

Most task-oriented dialogue agents adopt task-completion with fewer dialogue turns as the objective of the dialogues. In this case, the reward function is easily defined with two types of rewards; one is that the agent obtain large positive reward if an agent completes the task, and another is that the agent obtain small negative reward

each time it takes actions. With this reward function, the agent comes to complete the task with minimum dialogue turns.

Contrary to this, when an agent that aims to make rapport with users, we don't know how to formulate the objective into reward functions. Despite of this difficulty, we can find adequate dialogues even when their goals are not obvious. To realize dialogue control for such less-goal oriented agents on the basis of reinforcement-learning, some studies annotate ratings like Likert scales to dialogue corpus and utilize them as rewards [Meguro et al., 2010, Williams and Young, 2005]. However, since RL calculates the policies based on the weighted average of the rewards, such ordinal scale ratings are inadequate to be used for rewards of RL.

Some work adopt inverse reinforcement learning (IRL) to automatically define the reward function such that the agent can reproduce the dialogue sequences performed between humans. This approach supposes the sequences are equally appropriate for a given task; on the contrary, as we explained above, the adequacy of the dialogue sequences are not identical since each person's dialogue strategy in conversation varies among people. This variety possibly makes some sequences inadequate. When a reward function is estimated from the sequences containing inadequate ones, an inadequate reward function should be obtained. To obtain adequate reward function, it is necessary to be able to utilize evaluations of the sequences for the estimation of the reward function.

3.2 Response generation of conversational dialogue agents

Much work has been proposed to address utterance generation for dialogue agents. Slot-filling is one of the major approach that an agent uses template sentences with blank slots designed in advance, and fills the slots using information acquired from user utterances or other resources [Cassell, 2000, Oh and Rudnicky, 2000]. Although this approach can treat dialogue contexts and works well when we develop task-oriented dialogue agents whose slots are easily defined, it is difficult to define slots for conversational agents whose goals are not obvious.

For dialogues where such slots cannot be defined, rule-based utterance generation is widely used. First we construct dialogue example database that consist of the pairs of matching *patterns* and *responses* associated with the patterns, which are called as *rules*. The rules are created manually or gathered from real dialogues. Then, an

agent with this approach retrieves patterns that match a user utterance and outputs responses associated with the retrieved patterns [Weizenbaum, 1966, Wallace, 2004, Lee et al., 2006]. This rule-based approach works well when the range of dialogue topics are narrow; however, to generate utterances of conversational agents, since the variety of topics in conversations is huge, an enormous amount of resources are required to build enough rules to cover all topics and to maintain the developed the rules without contradiction. To make it feasible to generate reasonable utterances, one approach is to avoid generating responses that contain new information relevant to user utterances. For instance, ELIZA, which is a famous rule-based conversational agent for counseling, tends to respond to user utterances with repetition or general questions like *why do you think so?* [Weizenbaum, 1966]. Although recent rule-based agents have repeatedly won the Loebner Prize[1] (a competition for chatter bots), most users noticed this simple behavior and disappointed them.

Another approach to automatically generate agent utterances that are relevant to such wide-variety topics of user utterances is retrieval-based approach. This approach retrieved sentences from the web or microblogs as agent utterances by shallow sentence matching with user utterances [Shibata et al., 2009, Ritter et al., 2011]. This approach can generate responses relevant to user utterances by leveraging wide-variety of topics of the web articles. However, since the retrieved sentences include the inherent contexts of the document in which the sentences originally appeared, the retrieved sentences have the possibility of containing information that is irrelevant to user utterances.

As an improvement of the retrieval-based approach, machine-translation (MT) based approach are proposed. This approach solves the response generation as a kind of machine-translation problems from user utterances to agent utterances. Recent advances in neural networks such as Long Short Term Memory networks (LSTM) have been introduced to machine-translation, and they have been applied also to the utterance generation for dialogue agents [Sutskever et al., 2014]. These approaches can handle typical response patterns, which have a possibility to reduce the development cost of rule-based agents; however, since the obtained relations are limited to such trivial, frequently appeared patterns as “my birthday‘” to “happy birthday” and co-occurrences like “Potter” to “Harry”; thus, expansion of conversation

[1] http://www.loebner.net/Prizef/loebner-prize.html

topics is also limited. Besides, the generated utterances are sometimes syntactically inappropriate sentences.

As described above, in open-domain conversation, it remains a difficult problem to generate dialogue agent utterances that contain non-trivial new information with suppressing the contamination of irrelevant information to user utterances. In this thesis, we focus on automatic expansion of conversational topics that are not trivial and are relevant to user utterances.

3.3 Answering to agent's personality questions

In conversations, people often ask questions related to the specific personality of the person with whom they are talking, such as favorite foods and experience playing sports [Tidwell and Walther, 2002]. Such question-answering interaction is important to establish *rapport*, because such questions express that the speaker wants to know about and make rapport with the dialogue partner, and the answers (self-disclosure utterances) are effective to build relationship with them. Nishimura et al. showed that such personality questions also appeared in conversations with conversational agents [Nisimura et al., 2003].

Besides, these questions are conversation triggers that are used to begin a conversation. If an agent avoids answering personality questions like ELIZA [Weizenbaum, 1966] which repeats almost the same questions or asks the talker back the question why the talker asks such questions, users will be disappointed with the agent because such behavior denotes that the agent do not want to make *rapport* with users; therefore, the capability to answer personality questions is an important function in the development of conversational agents.

Most previous research on the personality of conversational agents has investigated the agent's personality using roughly-grained categories, such as the Big-Five [Caspi et al., 2005, John and Srivastava, 1999, Mairesse and Walker, 2007]. All of these studies parametrized the personalities, but they did not deal with specific subjects of the personalities, which are required to answer personal questions.

To answer specific questions about agents' personalities, Batacharia et al. developed the Person DataBase (PDB), which consists of question-answer pairs (*QA pairs*) evoked by a pre-defined persona named *Catherine*, a 26-year-old female living in New York [Batacharia et al., 1999]. Their approach retrieves a question sen-

tence that resembles the user's question utterance from the PDB and returns an answer sentence associated with the retrieved question. Although they developed such QA system, they did not evaluate the effectiveness for improving user satisfaction of the system. Traum et al. recently proposed time-offset interaction [Traum et al., 2015, Leuski and Traum, 2011], but they only evaluated the adequacy of each generated response and did not examine the effectiveness of the personality QA.

We focus on the development of a question-answering system for such specific personality questions by corpus-based analysis and development with large-scale personality database.

3.4 Evaluation of conversational agents

Evaluation is important to determine how to improve the agents. Most studies manually evaluated them with subjective user satisfaction by users [Meguro et al., 2010, Sugiyama et al., 2013] or with objective interaction qualities by trained expert annotators [Schmitt et al., 2011, Sugiyama et al., 2014]. However, we consider such approach is not desirable since not only it requires a huge amount of cost but also the annotated evaluation scores are not replicable; i.e., it is difficult to compare between scores of a proposed approach and previously reported one.

To automatically evaluate dialogue agents, several works are proposed that analyze the dialogues performed between users and the agent. The most famous example of this approach is PARADISE proposed by Walker et al., which leverages task-dependent metrics and human responses related to the subjective impressions [Walker et al., 1997]. For less goal-oriented dialogue agents such as conversational agents that do not have task-dependent metrics, Schmitt et al. proposed Support Vector Machine (SVM) based evaluation method that predicts user satisfaction leveraging with several features such as Automatic Speech Recognition (ASR) results and their confidence scores, emotion-tags and dialogue-act tags [Schmitt et al., 2011]. Hidden Markov Models (HMM) based evaluation methods are proposed that model human-human dialogues using HMM and predict detailed user satisfaction transitions (smoothness, closeness and willingness to continue) of human-computer dialogues [Higashinaka et al., 2010, Engelbrech and Hartard, 2009]. Although these metrics are able to be calculated automatically, these require actual dialogues between humans and agents that takes huge cost and are difficult to be reproduced.

With the aim of elimination of manual interaction, some work adopt user simulator to develop dialogues. User simulator, which mimics real users' behaviors, is widely used for evaluation of task-oriented dialogue agents. Since users in task-oriented dialogues like troubleshooting are assumed to have clear objective and the domain knowledge related to the task is able to be covered, users' utterances are predictable and it is feasible to develop such simulator [Williams, 2007, Misu et al., 2011].

On the contrary, the user utterances in conversation are difficult to be predicted, since the objective of the users varies. This makes it difficult to design user simulators; therefore, some work have begun to adopt reference-based approach that evaluates agent utterances for input utterances without actual dialogues, according to the distances scores like BLEU scores [Papineni et al., 2002] with pre-defined appropriate reference sentences for the input utterances [Ritter et al., 2011, Nio et al., 2014].

While such a reference-based evaluation methodology shows high correlations with human annotators in machine-translation, Ritter et al. reported that the reference-based approach does not show high correlation with human annotations in chat-oriented dialogues. In machine-translation, since agents are required to generate sentences that have exactly the same meaning as the original input sentences, only one or just a few reference sentences could be enough to cover the appropriate range of target sentences. On the other hand, in conversations, since the appropriate range is much larger than machine-translation, appropriately evaluating the responses is difficult. Galley et al. proposed Discriminative BLEU (Δ BLEU), which leverages 15 references with manually annotated evaluation scores to estimate the evaluation of chat-oriented dialogue agents' responses [Galley et al., 2015]. This method calculates utterances scores as the average of BLEU values weighted with the manual evaluation scores; i.e., if an utterance is close to negative references, this utterance will obtain a low score. This method shows medium system-wise correlations that are calculated between the average of 100 responses of annotated and estimated scores; however, they reported low sentence-wise correlations (Pearson's $r \leq 0.1$).

We assume that the reason of the low sentence-wise correlations is insufficient reference size for covering the range of agent utterances and their naive handling for negative references. The 15 references they used seem insufficient to capture the differences between the utterances with word-sensitive metrics like BLEU, which cannot consider synonyms or negation terms like *not*. Besides, although references should be

gathered intensively around positive utterances to increase the sensitivity of the differences, most of their references seem to be far from the positive areas. Naive handling of the negative references causes another problem. Since the range of negative utterances is too large to be captured, there should be some negative utterances that are far from all the references. Although such utterances should be rated with negative score of -1, the method evaluates such utterances with medium value of 0 instead of negative of -1. These problems caused such low sentence-wise correlations.

4. Approaches in this thesis

Even if we focus on one-turn response generation, there remains these problems to generate agent utterances in open-domain conversation. These problems arise with the difficulties: *considerable variations* of available state-action pairs, dialogue topics, personality questions and the appropriate range of them. We believe that the key for solving these problems is the amount of data. In this section, we describes our approaches how to leverage respectively developed large-scale corpus to solve the problems.

4.1 Dialogue control

Previous IRL methods find reward functions on the assumption that target sequences are equally appropriate for a given task; however, this assumption is not suitable for the dialogue sequences in conversation. Our idea for this problem is to estimate a reward function such that total rewards obtained through dialogue sequences match their human-annotated ratings. This is simple and reasonable approach to utilize the ratings in IRL framework; however, another problem arises that absolute values of the ratings are ordinary scales that are hardly consistent between annotators. Instead of using the absolute values of ordinary scales that are not adequate to be used as the target values of the reward function, we utilize the pairwise-preferences of the ratings to estimate the reward function, which is called as preference-learning based inverse reinforcement learning (PIRL). Our proposed method leverage large-scale test chat dialogues with ratings to estimate the reward function such that the pairwise preferences (orders) of the annotated ratings among the dialogue sequences match those of total expected

rewards through dialogues.

4.2 Open-domain utterance generation

To generate appropriate response utterances that have non-trivial information that relates to open-domain user utterances, we propose a novel utterance generation method that synthesizes a new sentence with consists of both a primary topic of a user utterance and a new topic relevant to the user utterance topic. This way, we can generate agent utterances that contain new information relevant to user utterances; i.e., we can suppress the generation of parrot utterances.

To automatically define the relevancy between topics, we utilize dependency relations that express more specific relationship than normal co-occurrence. We propose a utterance generation method that combines two strongly related *semantic units* (phrase pairs with dependency relations that represents the topics of utterances) to create an agent utterance; here, the first semantic unit is the one found in the user utterance and the second semantic unit is the one that has a dependency relation with the first one in a large text corpus.

When we combine the two semantic units into a sentence, it is difficult to select syntactically and semantically appropriate post-positional particles or conjunctions that are used to complete between the semantic units. To avoid this difficulty, we utilize example sentences in large-scale utterance corpus to know how to select such post-positional particles or conjunctions. This way, we can generate syntactically and semantically appropriate agent utterances that contain new information relevant to open-domain user utterances.

4.3 Answering for personality questions

Since previous QA systems for personality questions are developed with Person DataBase (PDB) that are designed by a few developers, it is difficult to know the coverage of the PDB or the QA system for personality questions in real conversational dialogues.

This study gathers a large number of question-answer pairs from many question-creators and a few answer-creators, and manually categorizes the pairs so that each *question category* represents identical meaning. Using this question-answer pairs, we

investigate the types of questions that are frequently asked and the coverage of questions in human-human conversations. Besides, leveraged with the QA pairs that contain several question sentences for each question category, we can develop a QA system that are expected to be robust for word fluctuation.

4.4 Automatic evaluation of conversational agents

There exists two major problems in previous Δ BLEU approach: sparseness of the references and inappropriate handling of negative references. To increase the coverage of utterances and sensitivity of small differences especially around positive utterance area, we intensively create many references that close to positive areas. To handle negative references appropriately, we propose a regression-based automatic evaluation method that evaluates utterances based on the similarities or distances to many reference sentences and their annotated evaluation values. Since we can leverage the distances from positive references to estimate utterance scores, this method can evaluate an utterance that are far from all the references with a negative score.

5. Contribution of the Thesis

This study discusses and gives a clue to the development of conversational agents that make rapport with users. Since user utterances in conversation have wide variety of topics and dialogue-acts, it is difficult to model the user utterances or user states manually. Moreover, if we consider discourse relations between the utterances such as contradiction or entailment, input information for the agent becomes too complex and sparse to develop the agent. To suppress the complexity of input information to feasible level, we focus on one-turn response generation for open-domain user utterances. This limitation enables us to develop utterance generation methods that can respond all the type of user utterances.

In this thesis, we try to give primary solutions to the problems on the basis of leveraging large-scale corpora that are specially designed to solve each of the problems. The contributions of this study are as follows, and their relations for the one-turn utterance generation are shown in Figure 1.2.

- We proposed a preference-learning based inverse reinforcement learning (PIRL)

that calculates the appropriateness of agent actions on the basis of real human-human dialogues with ratings. Different from previous IRL that estimates a policy function with the assumption that training sequences are equally appropriate, our PIRL leverages the ratings of training sequences to estimate the policy function. This advantage is necessary to realize appropriate dialogue control for conversational agents, since even human-human dialogues possible contain inadequate sequences in conversation.

- We proposed a novel utterance generation method for conversational agents that introduces new information related to open-domain user utterances and suppresses the contamination by irrelevant information. Previously proposed rule-based utterance generation methods cannot cover the wide range of topics, and retrieval- and MT (machine translation)-based methods sometimes generate parrots or irrelevant utterances, with which users feel difficulty to continue talking. To expand the dialogue topics, our method combines two topics; one topic is extracted from the user utterances, and another topic is retrieved from large-scale corpus so that the two topics have dependency relations. Our method generates utterances that have new information relevant to the current topics, with which users are easier to continue talking than conventional methods.

- We developed a question-answering system for questions that ask agent's specific personalities, using manually created large-scale question-answer pairs. Such questions are used as a conversation trigger, which should be answered otherwise the dialogue will easily break. Besides, such questions and their answers are effective to establish rapport between users and such agents. However, previous work about such QA systems did not examine the coverage of frequently asked questions in real dialogues by their systems, nor the effectiveness for the improvement of interaction quality. We first developed Person DataBase (PDB) with large-scale personality question-answer pairs for six personas gathered from many questioners and a few answerers and categorize the questions manually. This hybrid method of question gathering with crowd sourcing approach and careful categorization of gathered questions by trained experts is a key to develop the PDB with wide variety of topics, which enables us to investigate frequently asked questions and the coverage of developed questions in

real conversations. Through the investigation, we revealed that some frequently asked questions are overlooked by previous QA systems. We also developed a personality QA system and examined the effectiveness of the system through subjective evaluations.

- We proposed an automatic evaluation system for the one-turn responses of conversational agents. Our proposed method leverages large-scale multi-references with ratings to estimate the agents' evaluations. Most previous work subjectively evaluated such agents, but it requires huge amount of cost; besides, the scores are not replicable and this makes it difficult to compare newly proposed approach and previous ones. Although some auto-evaluation methods are proposed, most of them focus on the evaluation of developed dialogues, which also take huge amount of cost to be created and are not replicable. A few methods that do not require the actual dialogues are proposed; however, these methods show insufficient estimation performance because of the naive handling of negative samples. Unlike these methods, our regression-based method with large-scale references intensively gathered around positive utterances estimates scores that are replicable and show high correlations with subjectively annotated scores.

6. Outline of the thesis

Outline of this thesis is as follows. First we introduce general architecture of conversational dialogue agents and the technologies of its components in Chapter 2. In Chapter 3, we describe our dialogue control module leveraged with preference-learning based inverse reinforcement learning and examined effectiveness of our approach to learn appropriate dialogue control from dialogue sequences with ratings. In Chapter 4, we describe our open-domain utterance generation method that retrieves topics relevant to user utterances dependency relations. In Chapter 5, we describe the analysis of personality questions gathered by many questioners, and the development of our personality question-answering system. In Chapter 6, we describe our automatic evaluation framework for conversational agents leveraged with large-scale multi-references. We conclude this thesis and show the future directions in Chapter 7.

Chapter 2

Conversational agents

1. Introduction

This chapter introduces general architecture of conversational dialogue agents and their evaluations. Figure 2.1 illustrates the architecture, which consists of the following three modules: utterance understanding module that extracts the topics of the utterances and the action types of utterances (e.g., greetings or questions), dialogue control module that decides agent actions, and utterance generation module that generates agent utterances according to the contents of user utterances and the agent actions. In addition, developed dialogue agents are used to be subjectively evaluated though dialogues between agents and human users. This chapter describes the detailed technologies of these modules.

2. Modules of the agent architecture

2.1 Utterance understanding

Utterance understanding is the first part of the modules which accepts user utterances and extracts information from them. The extracted information is roughly classified into the aspects of utterance contents and user states.

Extraction of contents One of the important functions of understanding of contents is to find dialogue topics that are attracted by the user. *Centering*, proposed in a theory

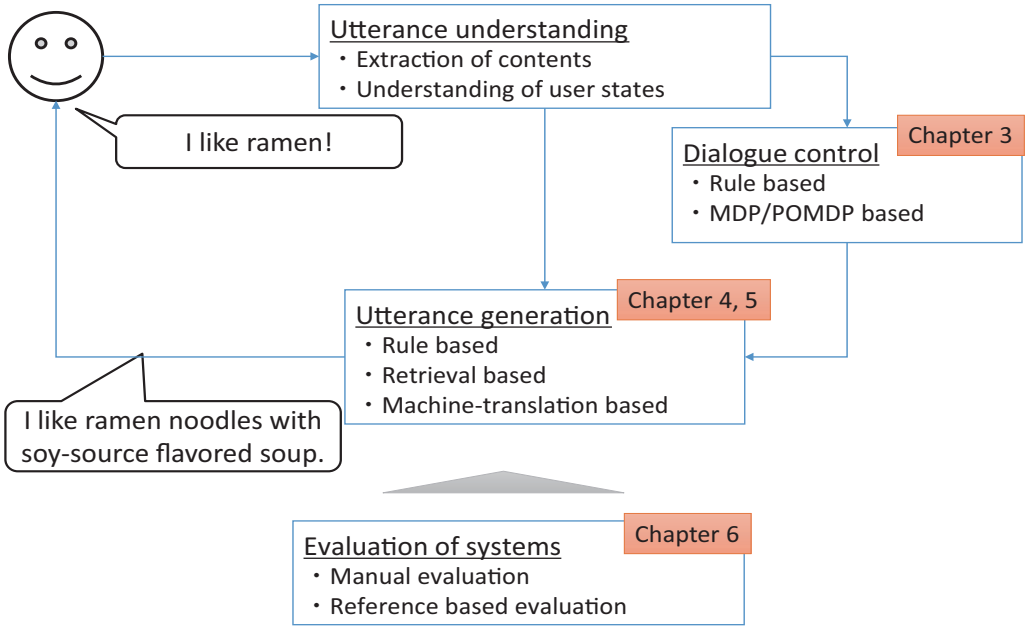


Figure 2.1. System architecture

of discourse, is a widely-used computational model in conversational agents to coordinate such attention [Walker et al., 1998]. This model calculates the salient noun words (NPs) from previous user and agent utterances as *centers* using ranking methodology.

In addition to the center-words, predicate-argument structures (PAS) are also important to understand the contents. Center-words that consist of only NPs sometimes are not enough to capture the dialogue topics. For example, we cannot distinguish the dialogue topics either *read a book* or *burn a book* only with the center *book*.

Anaphora resolution is an important function to extract these information accurately. If we do not resolve anaphora expressions, agents cannot capture the contents of user utterances, which are necessary to generate agent utterances that appropriately relate to user utterances. This problem is critical particularly for Japanese, which omits many arguments including *I* and *you* as zero-anaphora. To overcome this problem, Imamura et al. proposed a predicate-argument structure analyzer with zero-anaphora resolution for dialogue agents [Imamura et al., 2014]. They reported that their PAS analyzer trained with both dialogue and newspaper corpus can resolve zero-anaphora like *I* and *you*, which are not resolved by conventional PAS analyzer trained only with newspaper corpus.

Understanding of user and dialogue states To generate appropriate agent utterances, it is important to understand user's internal states or dialogue states. In task-oriented dialogue, such states are usually defined with the condition of slots that are designed to check the progress or completion of tasks. Much work on the estimation of the conditions of slots are performed; for example, a shared task for the dialogue state tracking on task-oriented dialogues called Dialogue State Tracking Challenge has been held several times.

Contrary to this, it is difficult to define such slots for conversational agents, where their domains or goals are not obviously defined. For the modeling and automatic detection of the structures of such dialogues, dialogue-act (DA), which represents the abstract meaning of an utterance in terms of actions such as *questions* or *greetings*, is widely used [Stolcke et al., 2000, Higashinaka et al., 2014, Meguro et al., 2010, Sugiyama et al., 2013]. Table 2.1 shows an example of dialogue acts defined by Meguro et al. for the development of listening-oriented dialogues agents.

2.2 Dialogue management

This section focuses on the explanation of dialogue-control, which is a major topic in dialogue agent research area that decides appropriate agent actions for user states. Previous studies realize the dialogue control with rule-based approaches that a human defines agent actions for each user state [Wallace, 2004]; however, if the number of rules increases, it is difficult to define the rules consistently.

To avoid this difficulty, recent studies adopt Partially Observable Markov Decision Process (POMDP) to model the states and actions [Williams, 2007, Meguro et al., 2010, Williams and Young, 2005, Williams and Young, 2007]. This model assumes that Markov Process, where each step of state s_t is modeled with previous state s_{t-1} and action a_{t-1} by transition probability $p(s_t|s_{t-1}, a_{t-1})$. In POMDP environment, states can be obtained with noisy observation o_t that is modeled by observation function $p(o_t|s_t)$. To obtain optimal policy function (i.e., mapping function from (partially observable) states to actions), this model utilizes reinforcement learning (RL). RL automatically decides the appropriate agent actions for user states in order to maximize the total expected rewards received from a reward function designed by a human; therefore, if a reward function meets the objective of a dialog, RL automatically decides the optimal agent actions for the objective. This condition is

Greeting	Greeting and confirmation of dialogue theme.
	e.g. Hello. Let's talk about today's lunch.
Information	Delivery of objective information.
	e.g. My friend said that restaurant is good.
Self-disclosure	Disclosure of one's preferences and feelings.
sub:fact	e.g. I live in Tokyo.
sub:experience	e.g. I ate a hamburger for lunch.
sub:habit	e.g. I always go out for dinner.
sub:preference	e.g. I like a hamburger.(positive)
sub:preference	e.g. I don't like a hamburger.(negative)
sub:preference	e.g. Its taste is near my homemade taste.(neutral)
sub:desire	e.g. I want to try it.
sub:plan	e.g. I will go there next week.
sub:other	
Acknowledgment	Encourages the conversational partner to speak. e.g. Well. Aha.
Question	Utterances that expect answers.
sub:information	e.g. Please tell me how to cook.
sub:fact	e.g. What kind of curry?
sub:experience	e.g. What did you eat for dinner?
sub:habit	e.g. Did you cook yourself?
sub:preference	e.g. Do you like it?
sub:desire	e.g. Don't you want to eat rice?
sub:plan	e.g. What are you going to eat for dinner?
sub:other	
Sympathy	Sympathetic utterances and praises. e.g. Me, too.
Non-sympathy	Negative utterances. e.g. Not really.
Confirmation	Confirm what the conversation partner said. e.g. Really?
Proposal	Encourage the partner to do. e.g. Try it.
Repeat	Repeat adjacent the partner's utterance.
Paraphrase	Paraphrase adjacent the partner's utterance.
Approval	Bring up or show goodwill toward the partner. e.g. Absolutely!
Thanks	Express one's thanks e.g. Thank you.
Apology	Express one's regret e.g. I'm sorry.
Filler	Filler between utterances. e.g. Uh. Let me see.
Admiration	Express one's affection. e.g. A-ha-ha.
Other	other utterances.

Table 2.1. Definition and example of dialogue act tags[Meguro et al., 2010]

easy to meet when the objective is represented with an obvious goal (i.e., goal-oriented dialog) such as troubleshooting [Williams, 2007]. However, when an obvious goal does not exist (i.e., less-goal oriented dialog) the reward function is difficult to design to meet the objectives. For example, considering building a dialogue agent that aims to provide counseling treatment, we don't know where the obvious goal is. For this kind of task, some studies annotate ratings to dialogue corpus and utilize them as reward function [Meguro et al., 2010, Williams and Young, 2005]. However, since the ratings are ordinary scales, they are not adequate to be used for the rewards that are summarized through dialogue sequences to define appropriate policies. Besides, the ratings have an ambiguity that evaluators annotate individual ratings even if they intend to the same appropriateness for the objectives. These characteristics make it difficult for rating-based reward functions to meet the objectives.

To set an appropriate reward function automatically, inverse reinforcement learning (IRL) has been proposed [Ng and Russell, 2000, Abbeel and Ng, 2004] and is adopted for dialogue control [Chandramohan et al., 2011, Boularias et al., 2010]. IRL finds a reward function, with which an agent generates similar sequences as the training ones in corpora. This indicates that all the sequences are assumed as equally optimal in the current IRL studies. Therefore, if the training sequences contain admissible (but not optimal) ones, we have to discard them in advance; otherwise (i.e., including the admissible sequences), non-optimal sequences will be generated with the estimated reward function. However, if the state-action space is large, such discard increases the sparseness of the data which causes harmful effects to the reward estimation. To learn the optimal policy for conversational agents whose state-action spaces are large and the objectives are not clear, it is required a novel IRL model that can train with dialogues with ratings.

2.3 Utterance generation

Rule-based approach To respond to such utterances, some conversational agents adopt a rule-based approach. These agents, whose rules are composed of many pattern-response pairs that are built by hand, find patterns that match the phrases contained in the user utterances and generate response sentences associated with the patterns. Since

rule-based conversational agents have repeatedly won the Loebner Prize[1] (a competition for chatter bots), we consider their approach effective for developing conversational agents. However, since the variety of topics in conversations is huge, an enormous amount of resources are required to build enough rules to cover all topics.

Retrieval-based approach To respond to user utterances on various topics, Shibata et al. proposed a retrieval-based approach that extracts sentences from a corpus whose contained sentences are collected from the web with keyword-search [Shibata et al., 2009]. They reported that their agent with a domain-specific corpus can respond to fine-grained topics; however, each retrieved sentence is too long to be used as an utterance and might contain irrelevant information.

Unlike web documents, Twitter contains many short conversational sentences. On Twitter, since users often post sentences related to their daily lives and chat using its in-reply-to function, these sentences are written about daily topics in light, breezy styles, making them very suitable for conversational agent's utterances. Focusing on these features, Ritter et al. proposed IR-status and IR-response approaches [Ritter et al., 2011]. IR-status retrieves reply posts whose associated source posts most resemble user utterances. This approach is reasonable to leverage the in-reply-to function; however, when it cannot find similar sentences or the relation between source and reply posts depends on unobserved contexts, it generates irrelevant, incomprehensible sentences as agent utterances. The IR-response approach resembles IR-status, but it retrieves reply posts that most closely resemble user utterances. Even though this approach avoids generating irrelevant utterances, IR-response has difficulty expanding the conversation topics; if the same sentence as the user utterance is contained in the corpus, it parrots the user utterance. Ritter et al. compared the approaches and reported that IR-response obtained better user evaluations than IR-status. The reason for this is that IR-status generated many unreasonable responses, which arise from many unreasonable tweet-reply pairs because of hidden contexts of the pairs that exist and are understandable only between the users. On the other hand, IR-response sometimes parrots the user utterances, but hardly generates responses irrelevant to user utterances. The result shows that irrelevant utterances are worse than parrots that have no new information.

[1] http://www.loebner.net/Prizef/loebner-prize.html

Machine-translation based approach As an improvement of the retrieval-based approach, some work proposed machine-translation (MT) based approach that solve the response generation as a kind of machine-translation problems from user utterances to agent utterances. Ritter el al. also proposed MT-chat that generates agent utterances with a machine-translation method that utilizes source-reply pairs as a parallel corpus. They compared MT-chat and the retrieval-based approaches and reported that MT-chat obtained better user evaluations than the other methods since it introduces phrases related to user utterances to agent utterances by fragmenting sentences into more fine-grained units. However, since the obtained relations are limited to such fixed patterns as "my birthday'" to "happy birthday" and co-occurrences like "Potter" to "Harry"; thus, expansion of conversation topics is also limited.

Recent advances in neural networks such as Long Short Term Memory networks (LSTM) have been introduced to machine-translation, and applied to the utterance generation of dialogue agents. Especially, encoder-decoder model, which can automatically generate a sentence associated with input sentence [Sutskever et al., 2014], gathered much attentions because it only requires dialogue sentences without any annotations nor external knowledge to develop dialogue agents. This approach can handle typical response patterns, so it has a possibility to reduce the development cost of rule-based agents; however, it has the same difficulty as the MT-chat, and is more difficult to control the generation of utterances than non neural network based MT systems.

2.4 Evaluation of the agents

To make improvements in such chat-oriented dialogue agents, evaluation is important. Previous work has evaluated their agents by hand [Higashinaka et al., 2014, Sugiyama et al., 2014], which is common practice in dialogue research. But such approach not only requires a huge cost but also is not replicable; i.e., it is difficult to compare a proposed agent's scores with previously reported other agents' scores.

Some previous work that exists on chat-oriented dialogue agents evaluates their agents on the basis of the appropriateness of the responses for input sentences [Nio et al., 2014, Sugiyama et al., 2014]. Although most studies manually evaluated the responses, some automatically evaluated the responses by a reference-based approach, which calculates the distance scores like BLEU scores

[Papineni et al., 2002] between agent responses and the references for each input sentence [Ritter et al., 2011]. While such a reference-based evaluation methodology shows high correlations with human annotators in machine-translation, Ritter et al. reported that the reference-based approach does not show high correlation with human annotations in chat-oriented dialogues. In machine-translation, since systems are required to generate sentences that have exactly the same meaning as the original input sentences, only one or just a few reference sentences could be enough. On the other hand, in chat-oriented dialogues, since the appropriate range is much larger than machine-translation, appropriately evaluating the responses is difficult.

To overcome this problem, Galley et al. proposed Discriminative BLEU (Δ BLEU), which leverages 15 references with manually annotated evaluation scores to estimate the evaluation of chat-oriented dialogue agents' responses [Galley et al., 2015]. Their approach showed 0.484 of corpus-wise Pearson's r; however, this is not enough to substitute manual evaluations. Besides, they reported low correlations (sentence-wise Pearson's $r \leq 0.1$). We assume that this is because of lack of the references considering the wide-variety of utterances in conversations.

Chapter 3

Dialogue control based on inverse reinforcement learning

1. Introduction

In this chapter, we propose a preference-learning based inverse reinforcement learning (PIRL) that estimates a reward function from sequences with ratings. Preference-learning is a subfield of supervised learning that learns a preference (order relations) model from observed preference information. To evaluate the subjective appropriateness of conversations, ordinal scales such as Likert scales are widely used; therefore, we use only the preference of the ratings instead of the absolute values of the ratings.

2. Background: Inverse Reinforcement Learning

Reinforcement learning and inverse reinforcement learning is generally represented using the Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, R)$, where $\mathcal{S}$ is a finite set of states; $\mathcal{A}$ is a finite set of actions; $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is a transition function; $\gamma \in (0, 1]$ is a discount factor of future rewards; and $\theta : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$ is the reward function representing a target task. In an usual dialogue policy learning problem, $\mathcal{S}$ means user's actions and $\mathcal{A}$ means system's actions.

RL aims to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ in order to maximize total expected rewards. The policy is generally defined as $\pi(s) = \arg\max_a Q^{(\theta)}(s, a)$, where $Q^{(\theta)}(s, a)$ is an

action-value function that represents expected cumulative rewards of all paths that begin with an action a at a state s under the reward function θ. The quality of action-value function $Q^{(\theta)}(s, a)$ depends on the reward function θ; however, designing appropriate reward function is difficult when the criteria of a given task is not obvious.

To avoid this problem, inverse reinforcement learning (IRL) is used to find a reward function with which the learning agent generates trajectories or paths ζ, consisting of states s_t and actions a_t, that induce a feature count close to that of the experts. Maximum entropy IRL, which is state-of-the-art IRL algorithm, estimates a reward function θ^* that maximizes the likelihood of the observed paths $\tilde{\boldsymbol{\zeta}}$ with

$$\begin{aligned} \theta^* &= \arg\max_{\theta} L(\theta) && (3.1) \\ &= \arg\max_{\theta} \sum_{examples} \log P(\tilde{\zeta}|\theta, \mathcal{T}). && (3.2) \end{aligned}$$

In the standard problem setting of IRL [Abbeel and Ng, 2004, Ziebart et al., 2008], a reward function is defined as a linear combination, $\mathrm{reward}(\boldsymbol{f}_s) = \theta^\top \boldsymbol{f}_s$, where $\theta(||\theta||_1 \leq 1)$ is a reward weight parameter, and $\boldsymbol{f}_s : \mathcal{S} \rightarrow [0, 1]^K$ is a K-dimensional feature vector of a state s. By using this definition, a trajectory's reward is calculated with

$$\begin{aligned} \mathrm{reward}(\boldsymbol{f}_\zeta) &= \theta^\top \boldsymbol{f}_\zeta && (3.3) \\ &= \theta^\top \sum_{s_t \in \zeta} \boldsymbol{f}_{s_t}, && (3.4) \end{aligned}$$

and the probability of generating the trajectory is defined with

$$\begin{aligned} P(\zeta|\theta, \mathcal{T}) &= \sum_{o \in \mathcal{O}} P_{\mathcal{T}}(o) \frac{e^{\theta^\top \boldsymbol{f}_\zeta}}{Z(\theta, o)} I_{\zeta \in o} && (3.5) \\ &\approx \frac{e^{\theta^\top \boldsymbol{f}_\zeta}}{Z(\theta, \mathcal{T})} \prod_{s_{t+1}, a_t, s_t \in \zeta} P_{\mathcal{T}}(s_{t+1}|a_t, s_t). && (3.6) \end{aligned}$$

Here, $\mathcal{O}$ is action outcomes and o in an outcome sample that specifies the next state for every action. The indicator $I_{\zeta \in o}$ is 1 when ζ is compatible with o and 0 otherwise. Computing equation 3.5 is generally intractable, so they approximate it with equation 3.6 under the assumption that transition randomness has a limited effect on agent's behavior and that the partition function is constant for all $o \in \mathcal{T}$.

In this definition, policies $\pi : \mathcal{S} \rightarrow \mathcal{A}$ is stochastically defined with

$$\pi = P(\text{action } a|\theta, \mathcal{T}) \tag{3.7}$$

$$\propto \sum_{\zeta: a \in \zeta_{t=0}} P(\zeta|\theta, \mathcal{T}) \tag{3.8}$$

3. Preference-learning based Inverse Reinforcement Learning

In most less-goal oriented dialogue corpora, each dialogue sequence has a manually rated score that represents the appropriateness of the sequence [Meguro et al., 2010, Williams and Young, 2005]. The conventional IRL studies [Abbeel and Ng, 2004, Chandramohan et al., 2011, Boularias et al., 2010] assume that each training sequence is equally appropriate for a given task; thus, conventional IRL are not supposed to adopt to sequences with ratings. To estimate a reward function using sequences with ratings, we propose a preference-learning based inverse reinforcement learning (PIRL). Our algorithm estimates a reward function, with which it ranks training sequences $\boldsymbol{\zeta}$ with the same preferences as the training one $\boldsymbol{o}^*$. In this study, we adopt pairwise preference [Herbrich et al., 2000] to represent the preference model for the sake of simplicity of implementation. We define the pairwise preference using the magnitude relations of the training ratings $\boldsymbol{e}^*$ as $o^*_{i,j} = \frac{e^*_i - e^*_j}{|e^*_i - e^*_j|} = \{-1, 0, 1\}$. We model the pairwise preference learning as a binary classification using only the pairs that have a relation $o^*_{i,j} = 1$ (i.e., $e^*_i > e^*_j$) as

$$\begin{aligned} L(\boldsymbol{\theta}) &= P(\boldsymbol{o}^*|\boldsymbol{\zeta}, \boldsymbol{\theta}) \\ &= \sum_{i,j: e^*_i > e^*_j} \frac{(1 + o^{\boldsymbol{\theta}}_{i,j})^{\frac{1+o^*_{i,j}}{2}} \cdot (1 - o^{\boldsymbol{\theta}}_{i,j})^{\frac{1-o^*_{i,j}}{2}}}{2M}, \end{aligned} \tag{3.9}$$

where $o^{\boldsymbol{\theta}}_{i,j} = \frac{e^{\boldsymbol{\theta}}_i - e^{\boldsymbol{\theta}}_j}{|e^{\boldsymbol{\theta}}_i - e^{\boldsymbol{\theta}}_j|}$ is a simulated-preference of sequences ζ_i and ζ_j under an estimated reward function $\boldsymbol{\theta}$, and M is the number of training sequence pairs. Here, we define simulated-ratings under $\boldsymbol{\theta}$ as

$$e^{(\boldsymbol{\theta})}_n = \sum_{s_{n,t}, a_{n,t} \in \zeta_n} Q^{(\boldsymbol{\theta})}(s_{n,t}, a_{n,t}), \tag{3.10}$$

input : Training sequences $\boldsymbol{\zeta}$ with their preferences $\boldsymbol{o}^*$
output: Estimated reward function $\boldsymbol{\theta}$

0. Initialize reward function $\boldsymbol{\theta}^0$.

for n **to** N **do**
  1. Calculate action-value function $Q^{\boldsymbol{\theta}^n}(s,a)$ with current reward function $\boldsymbol{\theta}^n$ (3.11).
  2. Evaluate training sequences $\boldsymbol{\zeta}$ with $Q^{\boldsymbol{\theta}^n}(s,a)$ (3.10) and calculate simulated-preferences $\boldsymbol{o}^{\boldsymbol{\theta}^n}$.
  foreach $\{i,j|i<j, e_i^* \neq e_j^*\}$ **do**
    if $o_{i,j}^* \neq o_{i,j}^{\boldsymbol{\theta}^n}$ **then**
      3. Calculate $\frac{\partial L_{i,j}}{\partial \boldsymbol{\theta}^n}$ (3.12).
    end
  end
  4. Evaluate convergence with L.
  5. Update $\boldsymbol{\theta}^n$ using the L-BFGS algorithm with (3.15).
end

Algorithm 1: Preference-learning based Inverse Reinforcement Learning

where $Q^{\boldsymbol{\theta}}(s_{i,t}, a_{i,t})$ is a simulated action-value function under $\boldsymbol{\theta}$ explained as follows.

Basically, our PIRL iteratively calculates the simulated-preferences of each training sequence pairs with a current reward function $\boldsymbol{\theta}^n$ and updates it with derivation $\frac{\partial L_{i,j}}{\partial \boldsymbol{\theta}^n}$ calculated for each pair that has different preferences (i.e., $o_{i,j}^* \neq o_{i,j}^{\boldsymbol{\theta}}$). Its details are illustrated in algorithm 1.

The algorithm's input data are a set of sequences $\boldsymbol{\zeta}$ with their preferences $\boldsymbol{o}^*$. In Step 1, it calculates current action-value function $Q^{\boldsymbol{\theta}^n}(s,a)$ with current reward function $\boldsymbol{\theta}^n$. Since our PIRL requires the derivation of $Q^{\boldsymbol{\theta}^n}(s,a)$ for updating reward $\boldsymbol{\theta}^n$, we define the action-value function with an approximate version of value iteration algorithm as

$$Q^{(\boldsymbol{\theta})}(s,a) = \sum_{s'} \{\theta(s,s') P_T(s'|s,a) + \max_{a'} \sum_{s''} P_T(s'|s,a)\theta(s',s'')P_T(s''|s',a')\}, \tag{3.11}$$

where $\theta(s,s')$ is a reward value when the user state is transitioned from s to s' and $P_T(s'|s,a)$ is a transition probability from s to s' with agent action a. Next, the method

calculates simulated-ratings $\boldsymbol{e}^{\boldsymbol{\theta}^n}$ with (3.10) and simulated-preferences $o_{i,j}^{\boldsymbol{\theta}^n}$. In Step 3, if preferences $o_{i,j}^*$ and $o_{i,j}^{\boldsymbol{\theta}^n}$ are different, it calculates the derivation of the reward function for the pair ζ_i and ζ_j with

$$\frac{\partial L_{i,j}}{\partial \boldsymbol{\theta}^n} \propto o_{i,j}^* \{ \sum_{s_{i,t},a_{i,t} \in \zeta_i} \frac{\partial Q^{\boldsymbol{\theta}^n}(s_{i,t}, a_{i,t})}{\partial \boldsymbol{\theta}^n} - \sum_{s_{j,t},a_{j,t} \in \zeta_j} \frac{\partial Q^{\boldsymbol{\theta}^n}(s_{j,t}, a_{j,t})}{\partial \boldsymbol{\theta}^n} \}. \tag{3.12}$$

Each factor of $\frac{\partial Q(s,a)}{\partial \boldsymbol{\theta}}$ is formulated as

$$\begin{aligned} \frac{\partial Q(s,a)}{\partial \theta(s_1, s_2)} &= \delta_{s,s_1} P_T(s_2|s_1, a) \\ &\quad + P_T(s_1|s, a) P_T(s_2|s_1, a'_{s_1}), \end{aligned} \tag{3.13}$$

where

$$a'_s = \arg\max_a \sum_s \theta(s, s') P_T(s'|s, a) \tag{3.14}$$

and $\delta_{s,s1}$ is a Kronecker delta.

The algorithm sums up the derivation as

$$\frac{\partial L}{\partial \boldsymbol{\theta}^n} = \sum_{i,j:o_{i,j}^* \neq o_{i,j}^{\boldsymbol{\theta}^n}} \frac{\partial L_{i,j}}{\partial \boldsymbol{\theta}^n} \tag{3.15}$$

and iteratively updates the reward function with the L-BFGS algorithm [Liu and Nocedal, 1989].

4. Experiments

Our PIRL estimates an appropriate reward function from dialogue sequences with preferences calculated with the annotated ratings. In this section, we examine the effectiveness of our algorithm through the comparison between the following three algorithms: Maximum Entropy IRL, RL with profit sharing, and our PIRL.

Maximum Entropy IRL is a state-of-the-art IRL algorithm [Ziebart et al., 2008] that estimates reward function using only high-rated sequences. Through the comparison with this, we examine the influence of the data sparseness caused by discarding low-rated sequences.

RL with profit sharing is a popular approach to estimate action-value function using annotated ratings as reward function [Grefenstette, 1988]. Through the comparison with this, we examine the influence of the ambiguity of rating-annotation, since this approach utilizes the absolute values of the ratings as a reward function.

4.1 Dialogue data

We used the dialogue data collected in our previous study [Meguro et al., 2010]. The study aims to build a listening-oriented dialogue agent that attentively listens to other dialogue participant. We collected 1259 listening-oriented dialogues using human subjects who consisted of ten listeners (five males and five females) and 37 speakers (18 males and 19 females) and labeled each sentence of the collected data using 32 dialogue-act tags (totally, 67801 dialogue-act tags are contained in the corpus). The collected dialogues were evaluated using two third-party participants (annotators), who were neither listeners nor speakers in our dialogue data collection. The annotators evaluated each dialogue sequence in terms of how they would have felt "being heard" after the dialogue if they had been the speaker of the dialogue in question. They provided ratings on a 7-point Likert scale for each dialogue.

We used the speaker's dialogue-act tags as user state space $\mathcal{S}$, and the listener's dialogue-act tags as agent action space $\mathcal{A}$. When one utterance contains several sentences, the algorithms blend action-value functions of plural user states as $Q'(\boldsymbol{s}, a) = \frac{1}{|\boldsymbol{s}|}\sum_{s'\in\boldsymbol{s}} Q(s', a)$. On the other hand, the algorithms can generate only one agent action for each utterance.

4.2 Evaluation criteria

To examine whether each algorithm can generate the optimal sequences, an agreement rate of sequences between the testing and the generated by the agent seems a straightforward criterion. However, since several agent actions appear for each user state in the training sequences, and several agent actions are suitable for each user state, it is infeasible to generate the agent actions in the testing sequences accurately. Therefore, we argue that the agreement rate of the sequences is not adequate criterion to evaluate the algorithms.

To compare the algorithms, we defined three criteria: Agreement rates of the preferences, correlation coefficients of the preference, and "expected-ratings". The agreement rates and the correlation coefficients of the preferences are consistent criteria between algorithms since they are robust to the ambiguity of rating-annotation. If an agreement rate and a correlation coefficient of the preferences between the algorithm and an annotator are equivalent to those between annotators, we can utilize the algorithm as an evaluation simulator. In addition, as an intuitive criterion, we add "expected-ratings" that algorithms are expected to gain in real dialogue. While we would like the annotators to evaluate the sequences, the evaluation is difficult even for annotators since the sequences contain only user state and agent actions instead of sentences. Thus, we calculate the expected-ratings as the averaged ratings of the sequences that gain top-n highest/lowest simulated-ratings. We assume that the sequences with top-n highest simulated-ratings resemble those that algorithms will generate in real dialogue; thus, we believe that the expected-ratings from the sequences with the highest simulated-ratings (i.e., high expected-ratings) resemble ratings that algorithms will gain in real dialogue. On the other hand, expected-ratings from sequences with the lowest simulated-ratings (i.e., low expected-ratings) are also important to examine whether the algorithms can inhibit generating inappropriate agent actions.

4.3 Experiment settings

We divide data (sequences and their ratings) into training, development, testing sets with two settings: All-data and selected-data. In the all-data setting, we make training and development sets with one annotator's data, and testing set with another's data. Since the data consists of the same sequences between the annotators, the training, development and testing sets are divided so that their sequences have no overlaps one another.

In the selected-data setting, at first we select data that has similar ratings between the annotators (the difference is equal to or lower than 1); and then, we divide this data into 300 for training, 300 for development, and 111 for testing sets at random. We added 200 randomly generated sequences with the lowest-ratings to the training set for the sake of increasing variation of the data since our data were expected to contain few fatal sequences since the dialogues were performed by humans. See Table 3.1 for the

statistics.

	All-data	Selected-data
# train sequences	700 (ME:113)	500 (ME:88)
# train pairs	149515	75414
# dev. sequences	300 (ME:300)	300 (ME:150)
# dev. pairs	18177	16768
# test sequences	459	111
# test pairs	55310	2236

Table 3.1. Statistics of the data sets. The "ME:*" means the case of Maximum Entropy IRL that utilizes only data with high ratings (equal to or higher than 4).

4.4 Results and discussion

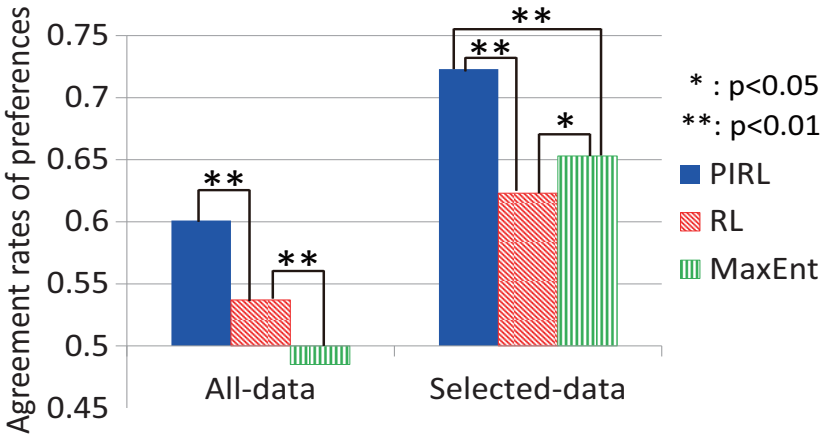


Figure 3.1. Comparison of the agreement rates of preferences. The agreement rate between the annotators was 0.632 in the all-data setting and 0.925 in the selected-data setting, and the random baseline was 0.5.

Figure 3.1 shows the agreement rates of the preferences between the algorithms and an annotator. This illustrates that our PIRL significantly outperforms the other algorithms in both settings. The reason why the agreement rates of the selected-data

setting are higher than that of the all-data setting is that we can remove the data that have opposite ratings between the annotators.

	All-data			Selected-data		
	Corr.	High	Low	Corr.	High	Low
PIRL	**0.200**	**4.66**	**3.46**	**0.363**	**5.06**	3.86
RL	0.069	4.46	4.13	0.218	4.46	**3.40**
MaxEnt	-0.001	3.80	3.60	0.285	4.66	3.93
Annotator	0.211	4.90	2.86	0.817	6.80	1.33

Table 3.2. Correlation coefficients and high/low expected-ratings. Annotator's coefficient is 0.211 in the all-data setting, and the average of the all-data ratings is 0.402 and the selected-data setting is 0.458. The higher/lower expected-ratings mean better performance in "High"/"Low" column.

Table 3.2 illustrates the correlation coefficients of the ratings and the high/low expected-ratings with the highest/lowest 15 sequences. Our PIRL shows higher correlation coefficients and high expected-ratings than the others, and lower expected-ratings in the all-data setting. This suggests that our PIRL estimates an appropriate reward function, with which it generates appropriate agent actions. Besides, the agreement rate of the preferences and the correlation coefficients of our PIRL are similar to the annotator's ones, our PIRL has a potential to be an evaluation simulator.

5. Conclusions

We proposed a preference-learning based inverse reinforcement learning (PIRL) that estimates a reward function for dialogue control from dialogue sequences with ratings. The contribution of our study is to extend the range of applications of inverse reinforcement learning (IRL) from sequences with single appropriateness to sequences with various appropriateness; thus, our PIRL can utilize non-optimal data, which is discarded in previous IRL, using pairwise preference information calculated with the ratings. Besides, our experiments show that our PIRL has a potential to be an evaluation simulator.

There are some studies that have the similar objective as our present study. Silva et al. proposed IRL with evaluation that utilizes pairwise preferences; however, this study force annotators to evaluate each pair of simulated sequences in each learning iterations [Freire da Silva et al., 2006]. While Cheng et al. proposed reinforcement learning based on preference-learning, this requires comparisons for each action pair [Cheng et al., 2011]. The evaluation costs taken in these studies are infeasible; on the other hand, our PIRL requires the rating evaluation only one time to each sequence pair; thus, it is easier to introduce than the other conventional algorithms.

Much work still remains. Since we examined the effectiveness of our PIRL with offline evaluation, we plan to evaluate our PIRL using online evaluation. Besides, it is very interesting topic to extend our discrete user state and agent action spaces to continuous distributions. A promising idea for this purpose is a topic model like Hidden Topic Markov Models [Boularias et al., 2010], which is used in dialogue control with IRL.

Chapter 4

Open-domain utterance generation using dependency relations

1. Introduction

To generate response utterances related to open-domain user utterances, it is important to suppress irrelevant information and expand conversation topics simultaneously. In this thesis, we combine two strongly related *semantic units* to create an agent utterance. A *semantic unit* is a phrase pair with a dependency relation, and *phrases* are phrasal units called *bunsetsu* segments in Japanese. Here, the first semantic unit is found in the user utterance, and the second semantic unit has a dependency relation with the first one in a large text corpus. Figure 4.1 shows the concept of our approach. For example, if user utterance "*Tokyo ni ikitai desu*" (I want to go to Tokyo) is given, first we extract an input semantic unit *Tokyo ni* (to Tokyo) $\rightarrow$ *ikitai desu* (I want to go) from it. Next, from a large text corpus, we retrieve semantic units that one of their phrases has a dependency relation with the phrases of the input semantic unit, such as "*Tokyo ni iketara Tokyo Tower wo mini ikitai*" (If I go to Tokyo, I want to visit Tokyo Tower). We extract such semantic units and combine the most frequently retrieved semantic units and the input one into a sentence like "*Tokyo ni ittara Tokyo Tower wo mini ikuno?*" (If you go to Tokyo, are you going to visit Tokyo Tower?).

By using automatically obtained phrase pairs, we can generate agent utterances for open-domain user utterances with simultaneously suppressing irrelevant information and expanding conversation topics. We examined the effectiveness of our approach by

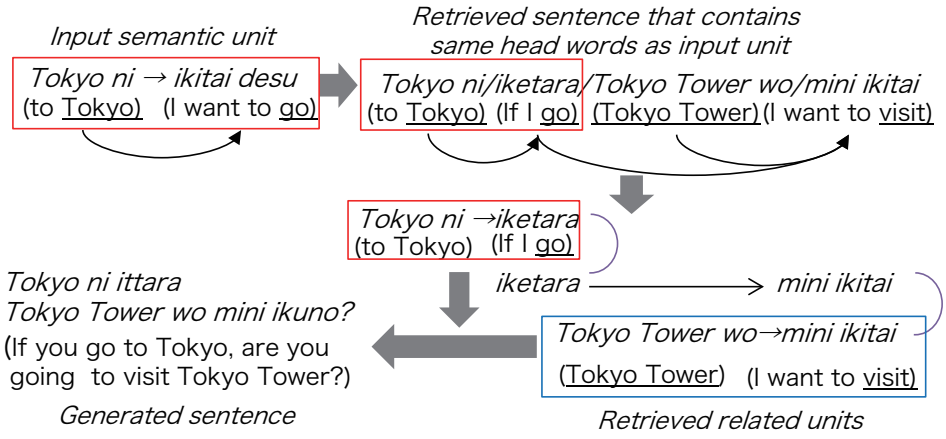


Figure 4.1. Concept of proposed approach explained with input user utterance "*Tokyo ni ikitai desu*" (I want to go to Tokyo)

comparing the appropriateness of the responses generated by conventional utterance generation approaches.

2. Proposed method

To generate response utterances related to open-domain user utterances, irrelevant information must be simultaneously suppressed and conversation topics expanded. Instead of retrieving sentences, one approach is to create them with information limited to that with a strong relation to user utterances. Unlike the retrieval, since the sentence creating approach explicitly suppresses irrelevant information, the created sentences will be more suitable for conversational agent utterances. To realize this, our method combines two strongly related *semantic units* (phrase pairs with a dependency relation; *phrases* are defined as *bunsetsu* in Japanese) to create an agent utterance. One semantic unit composes the user utterance, and the other unit has a dependency relation with the first one in a large text corpus.

Figure 4.2 outlines our method. For example, if user utterance "*Tokyo ni ikitai desu*" (I want to go to Tokyo) is given, first we extract an input semantic unit *Tokyo ni* (to Tokyo) → *ikitai desu* (I want to go) from it and retrieve the semantic units that are related to the input semantic unit like *Tokyo Tower wo* (Tokyo Tower) → mini ikitai (I want to visit) and *osushi wo* (some sushi) → *tabeyou* (let's eat). Next we combine the

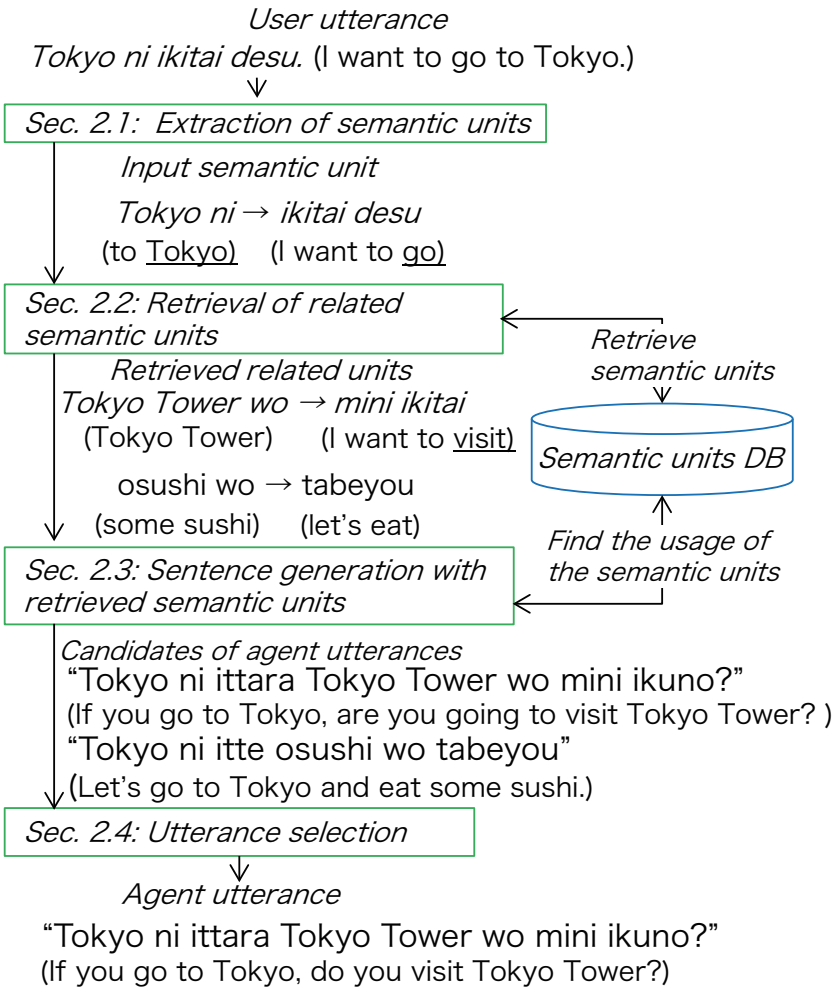


Figure 4.2. Outline of process of proposed method

input semantic unit and the related one to create a sentence like "*Tokyo ni ittara Tokyo Tower wo mini ikuno?*" (If you go to Tokyo, are you going to visit Tokyo Tower?) and "*Tokyo ni itte osusi wo tabeyou*" (Let's go to Tokyo and eat some sushi). Finally, we rerank the created sentences using some frequency-based statistics and choose the most appropriate one as an agent utterance. We explain the details of the processes below.

2.1 Extraction of semantic units from user utterances

From the input user utterances, we extract all the phrase pairs that have dependency relations as semantic unit candidates. We employ phrase pairs (not the phrases alone) as semantic units because only one phrase may lead to topics that have different or incorrect meanings to the user utterance topics. Following example shows the differences of phrases alone and phrase pairs. When *taikin wo* (much money) → *kaseida* (earned) is given as an input phrase pair, one possible response is "*eraine!*" (You've done it!); however, the same response is also generated to such topics as *taikin wo* (much money) → *rouhi shita* (wasted) or *taikin wo* (much money) → *surareta* (was picked) since only the phrase *taikin wo* (much money) cannot distinguish the original topics from them.

Some semantic units, however, contain stop words that are mainly used for grammatical purposes. Here, we define stop words with pronouns, auxiliary verbs like *suru, iu, naru, aru, iru* (do, seem, have, is), abstract nouns like *koto, no* (thing, that), and time-related words like *kinou, rainen* (yesterday, next year). Since these stop words are not expected to contribute to the expression of the topics of the user utterances, we remove such semantic units from the candidate units.

We also extract the following additional information from all the phrases: The standard form of head words, the part-of-speech tags, the phrase-ids that represent sentence order, the semantic attributes (one of the following, proper nouns, location, action, evaluative expressions defined in Kobayashi et al. [Kobayashi et al., 2005], demonstratives, questions, and none of the above), the literal strings of the whole phrase, the standard forms of phrases, and case-markers. Except for the evaluative expressions in semantic attributes, we obtain such information with a Japanese morphological analyzer called JTAG [Fuchi and Takagi, 1998] with a vocabulary extension using Wikipedia and Jdep [Imamura et al., 2007] as a dependency parser.

2.2 Retrieval of related semantic units

2.2.1 Semantic unit database

We construct a semantic unit database from which to retrieve semantic units with an arbitrary query of the information extracted in Section 2.1. We build this database as a collection of semantic units extracted from a large text corpus in the same way as

the semantic unit extraction from the user utterances described in Section 2.1, except for stop word filtering. We store the semantic units with sentence-ids, which represent sentences that the units originally belong to. We use Twitter as the corpus because it contains many subjective, conversational expressions and daily desires that are useful to generate conversational utterances.

2.2.2 Retrieval of related semantic units from database

Figure 4.1 illustrates the retrieval procedure from our developed database for an input semantic unit: *Tokyo ni* (to Tokyo) → *ikitai desu* (I want to go). We retrieve *seed* semantic units that have the identical head words to that of the input semantic unit, like *Tokyo ni* (to Tokyo) → *iketara* (If I go), from the semantic unit database. Here, the head words are *Tokyo* (Tokyo) and *iku* (go). For each seed unit, we retrieve semantic units that are originally belongs to the same sentence as the seed unit based on their sentence-ids. For example, since the seed unit originally belongs to sentence "*Tokyo ni iketara Tokyo Tower wo mini ikitai*" (If I go to Tokyo, I want to visit Tokyo Tower), we extract *iketara* (If I go) → *mini ikitai* (I want to visit) and "*Tokyo Tower wo* (Tokyo Tower) → *mini ikitai* (I want to visit). We filter the semantic units that share phrases with the seed semantic unit or those phrases without a dependency relation with the phrases of the seed unit. Here, the former unit is removed since *iketara* (If I go) is shared with the seed unit, and we adopt the latter as a candidate of related semantic units. We apply this procedure to each retrieved seed unit and obtain ones that appear more than once as related semantic units, like *Tokyo Tower wo* (Tokyo Tower) → *mini ikitai* (I want to visit) and *osushi wo* (sushi) → *tabeyou* (let's eat) (figure 4.2). When no related semantic unit is found, we substitute semantic units that have the phrases of the input semantic unit as seed semantic units and apply the procedure to each seed unit.

After retrieval, we aggregate the semantic unit pairs (seed and related units) by the contained head words. We remove the unit pairs that only appear once to avoid generating noisy sentences and use the top-N most frequent unit pairs for the following response generation. In this work, we set N with 10.

2.3 Response generation with retrieved semantic units

We combine the phrases contained in the aggregated unit pairs to generate response sentences. However, combining them with appropriate expressions and phrase orders is not easy. To avoid this difficulty, we adopt an example-based approach that utilizes the phrase orders and expressions to combine the units into sentences. We collect the literal strings of the phrases in each semantic unit pair and connect them with the appearance order in the sentence to which the units originally belonged; then, we aggregate the created strings and obtain the most frequent one as a candidate of response utterance.

In addition, since the string's tone is not always appropriate for conversational utterances, we slightly modify the end of each sentence to express predicted dialogue-act (details are shown in Table 2.1) by hand-coded rules. In this study, for simplification, we implemented a dialogue control module using linear-kernel Support Vector Machine, trained with the same data as we described in Chapter 3.

When the input semantic unit contains proper nouns and the number of retrieved units is less than infrequency threshold t, we use a template-based approach that drops the proper nouns into templates: "<proper noun>*tte amari shiranain desukedo douiu no desuka?*" (*I don't know <proper noun>. What is it?*). In this work, we set t to five.

2.4 Utterance selection with reranking

With the procedure explained in section 2.3, we obtain the candidates of agent utterances. In this study, we rerank the candidates and adopt the highest one with following three factors: *Topic saliency*, *relativeness between topics*, and *grammatical appropriateness*.

Topic saliency represents the saliency of semantic units extracted from the input user utterance, which is calculated with inverse frequency of each semantic units.

Second, *relativeness between topics* is the intensity of the relation between the units. We calculate this using the ratio of the frequency of a retrieved semantic unit from the corpus to that of a seed semantic unit from the input utterance.

The third term *grammatical appropriateness* evaluates the grammatical acceptability of the created candidates. We calculate this value with the number of appearance of the created sentence in the corpus, instead of perplexity of the sentence that is gen-

erally used to evaluate the grammatical appropriateness, since perplexity is affected with occurrence probability of each words contained in a target sentence; i.e., a sentence that is grammatically correct but has uncommon words sometimes obtains higher perplexity (lower grammatical appropriateness) than a sentence that is grammatically incorrect but contains only common words.

We design a objective function of the reranking with simply multiplying the factors as

$$\begin{aligned} score(\mathbf{s}_r|\mathbf{w}_i) &= saliency \cdot relativeness \cdot appropriateness \\ &= \frac{1}{\log(n(\mathbf{w}_i))} \frac{n(\mathbf{w}_r)}{n(\mathbf{w}_i)} \cdot n(\mathbf{s}_r), \end{aligned} \tag{4.1}$$

where $n(\mathbf{w}_i)$ and $n(\mathbf{w}_r)$ are the number of retrieved seed or related semantic units with head words $\mathbf{w}_i$ or $\mathbf{w}_r$ and $n(\mathbf{s}_r)$ is the number of identical strings $\mathbf{s}_r$ in the related semantic units.

To suppress the contamination of irrelevant information because of sparseness of data, in case all the candidates have small $n(\mathbf{w}_r)$ or $n(\mathbf{s}_r)$ (here, equal to or lower than 1), we use a word as a semantic unit instead of phrase-pairs.

3. Experiment

Humans naturally evaluate the effectiveness of utterance generation methods by examining them through chat experiments. However, since our proposed method generates topic-expanding utterances related to user utterances, it is not intended to respond to greetings or questions; thus, evaluating such a method through chat experiments is difficult. Therefore, we examine its effectiveness through module-based experiments that compare the appropriateness of utterances that are generated by several conventional approaches for input declarative sentences.

3.1 Evaluation procedure

Both the conventional methods (described in section 3.2) and ours assume that an input sentence is context-independent and contains one or more topics. To gather such sentences, first we collected many sentences from two types of corpora: Our conversational corpus and Twitter. The conversational corpus is composed of 3680 one-

Table 4.1. Evaluation criteria

Criterion	Description
(1) Response topics (saliency)	Appropriateness of topics included in agent utterances
(2) Consistency	Consistency with input utterances
(3) Topic relations (relativeness)	Relation of agent utterance topics to user utterances
(4) Comprehensibility (appropriateness)	Comprehensibility in Japanese (without strictly considering grammar correctness)
(5) Topic expansion	Inclusion of new information that expands dialogue topics
(6) Naturalness	Intuition of appropriateness as a response to input utterances
(7) Easy to continue talking	Ease of considering subsequent utterances

to-one text chats (130 K sentences) among people who talked without topic limitation [Higashinaka et al., 2014]. From it, we extracted sentences that were tagged with dialogue acts self-disclosure and information-provision, as defined by Meguro et al. [Meguro et al., 2010]. From the Twitter corpus, we gathered 150 M tweets and extracted sentences containing topical words defined by Google Trends 2012 in Japan[1]. Next, an annotator (not an author) evaluated the context-independency of each sentence on a 5-point Likert scale. We used sentences that obtained a maximum of five points and whose content was explicitly and context-independently written.

We randomly selected 140 sentences from each corpus and generated five responses for each sentence using the following five methods. If a method generated fewer than 5-best sentences, we substituted the worst ranked sentences for the missing ones until the number of generated sentences reached five. We shuffled the sentence orders for each input sentence, and three annotators (not the authors) subjectively labeled them on a 5-point Likert scale. Table 4.1 shows the evaluation criteria.

[1]https://www.google.co.jp/trends/topcharts#date=2012

3.2 Methods

We examined the effectiveness of our method by comparing the following five methods. Table 4.2 shows example utterances generated by each one.

Proposed We gathered one billion tweets (no overlap with the evaluation corpus) and filtered out those with such noisy expressions as *RTs* (retweets), @ (replies), URLs, brackets, and words repeated more than three times like *youyouyouyou*. We also removed sentences with fewer than ten characters and built a semantic unit database of 120 M tweets.

Word-driven We previously proposed a similar method that utilizes the dependency relations of words to generate sentences [Sugiyama et al., 2013]. The difference between these methods is how they defined semantic units; our previous method used a word as a semantic unit, not as a phrase-pair. Even though the topics of the user utterances are not completely expressed with only one word, this limitation sometimes effectively generates simple and comprehensible sentences. To develop this word-driven method, we used the same corpus as the proposed method.

IR-status This method is the same as IR-status [Ritter et al., 2011] that was explained in Section 2.3. We gathered 240 M source-reply pairs from tweets (no overlap with the evaluation and proposed method corpora), removed noisy tweets containing *RT*, *http*, and *#*, and obtained 45 M pairs. We used the Whoosh retrieval library[2], which retrieves sentences based on BM25F-weighted vector-space similarity [Pérez-Agüera et al., 2010].

IR-response This method is the same as IR-response [Ritter et al., 2011]. We used the same corpus as IR-status.

Rule-based Higashinaka et al. developed a conversational agent with a rule-based method that appropriately responds to 90% of the sentences in the conversational corpus described in Section 3.1 [Higashinaka et al., 2014]. This method consists of 149,300 rules (pattern-response pairs) written in AIML [Wallace, 2004] and retrieves

[2]https://pypi.python.org/pypi/Whoosh/

responses whose associated patterns match the left-hand side of the user utterances. Although the rule-based method cannot deal with wide range of topics, it has many predicate-driven rules such as *I like* * $\rightarrow$ *What do you like about it?*, which improve seeming coverage of topics. We expect this to be a formidable baseline against the statistical methods since agents with such rule-based methods have frequently won the Loebner Prize [Wallace, 2004].

Table 4.2. Example of generated utterances: English sentences were translated by authors

Input sentence	Generated sentences in Japanese (English)	Method
Twitter corpus *Machi jikan ga sukunaku narude arou typhoon* *no hi ni koso gesen ni mukau* (During typhoons, I go to arcades because there is almost no waiting time.)	*Gesen ni mukau oto gamer no kagami nan desune.* (You are a good music-video-game player who goes to video arcades.)	Proposed
	Machi jikan ha hima tte ii masune. (People complain that waiting is boring.)	Word
	Ittera. (Take care.)	IR-st.
	Tyhoon no hi ni ittara sukunakatta desuyo. (The wait was short because I went during a typhoon.)	IR-re.
	Arcade game ha naniga suki desuka? (What kind of games do you like?)	Rule
Conversational corpus *Sanma* [4] *ga suki desune.* (I like pacific saury[4].)	Sanma ga sakana no naka de sukida. (I like saury the most among all kinds of fish.)	Proposed
	Sanma tabetai desu. (I want to eat some saury.)	Word
	Suki desu. (I like it.)	IR-st.
	Waa! Sanma, suki. (Wow! I love saury.)	IR-re.
	Douiu tokoro ga toku ni suki desuka? (What do you like about it?)	Rule

3.3 Results and analysis

Table 4.3 shows the evaluation scores of the 1-best outputs for the conversational corpus. We employed the Wilcoxon signed-rank tests to calculate the significances. Since this corpus was used to develop a rule-based method, its score represents the performance for closed data; thus, it significantly outperformed the other methods. Among the statistical-based methods, the proposed and word-driven methods significantly outperformed the retrieval-based methods for (6) naturalness and (7) the easy to continue talking criteria, which are crucial for applying methods to develop conversational agents.

Table 4.4 shows the evaluation scores of the 1-best outputs for the Twitter corpus. In almost all the criteria, our proposed method significantly outperformed the other methods except for the word-driven method that resembles our proposal. This result means that our proposed method can improve the appropriateness of generated utterances in terms of the three factors: Topic saliency, relativeness between topics and grammatical appropriateness. Besides, our method appropriately introduce new information related to user utterances. These improvements raised overall naturalness and user satisfaction.

Contrary to the conversational corpus results, the rule-based method is inferior to our proposed method. There are two reasons for this result. First is that some of the input sentences of the Twitter corpus contain infrequent proper nouns for which the rule-based method lacks good rules. On the other hand, our proposed method generated sentences even for such infrequent proper nouns with the normal (phrase-pairs) procedure; the template-based utterance generation for infrequent proper nouns was used only once. Second, since the rule-based method was developed with the conversational corpus, the Twitter corpus is open data for the rule-based methods. Since retrieved rules by our rule-based method depend on the word ordering that inherently differs between corpus, the difference of corpus is so critical for the rule-based method to find appropriate rules. For example, if unimportant words appear in the more left-hand side of user utterances than important topic words, the method matches inadequate rules to the utterance. If we retrieve rules using inverse document frequency (IDF) weighted cosine similarity of bag-of-words expressions, this disadvantage can be relaxed; however, this method also abandon the carefully created patterns of word

[4]Sanma or saury is a popular fall fish in Japan

appearance. Further examination is required for the comparison between the retrieval methods from rules.

Table 4.3. Evaluation scores of 1-best outputs for conversational corpus (5 is the best. Bold is best score. Significances are calculated for comparison with proposed method. $*$: $p<.1$, $**$: $p<.05$)

	Proposed	**Word**	**IR-st.**	**IR-re.**	**Rule**
(1) Response topics	3.37	3.46	2.12^{**}	3.14	$\mathbf{3.86}^{**}$
(2) Comprehensibility	4.06	4.15	3.13^{**}	4.09	$\mathbf{4.52}^{**}$
(3) Consistency	3.36	3.48	2.28^{**}	3.30	$\mathbf{4.00}^{**}$
(4) Topic relations	3.49	3.54	2.29^{**}	3.38	$\mathbf{3.93}^{**}$
(5) Topic expansion	2.96	3.10	2.10^{**}	2.86	$\mathbf{3.41}^{**}$
(6) Naturalness	3.21	3.32	2.11^{*}	2.90^{*}	$\mathbf{3.90}^{**}$
(7) Easy to continue talking	3.12	3.25	2.10^{**}	2.83^{*}	$\mathbf{3.51}^{**}$

Table 4.4. Evaluation scores of 1-best outputs for Twitter (5 is the best, significances are calculated for comparison with proposed method. $*$: $p<.1$, $**$: $p<.05$)

	Proposed	**Word**	**IR-st.**	**IR-re.**	**Rule**
(1) Response topics	**3.42**	3.37	2.12^{**}	2.84^{**}	2.81^{**}
(2) Comprehensibility	**4.34**	4.23	2.92^{**}	3.84^{**}	4.17
(3) Consistency	**3.43**	3.37	2.12^{**}	2.97^{**}	2.78^{**}
(4) Topic relations	**3.53**	3.37	2.18^{**}	3.05^{**}	2.73^{**}
(5) Topic expansion	3.18	**3.23**	2.12^{**}	2.59^{**}	2.63^{**}
(6) Naturalness	**3.35**	3.30	2.11^{**}	2.73^{**}	2.83^{**}
(7) Easy to continue talking	**3.37**	3.36	2.06^{**}	2.49^{**}	2.69^{**}

Table 4.4 shows the differences between the proposed and word-driven methods are small. We believe that there are two reasons for this. One, most sentences in the conversational corpus have just one topic word, like "*violin*" (violin), with which the word-driven method easily generates such simple and comprehensible utterances as "*Violin hikitai desu*" (I want to play the violin) that only consist of two phrases *violin* (violin) $\rightarrow$ *hikitai desu* (want to play). In contrast, our proposed method tried to use phrase pairs even if the retrieved pairs were infrequent and possibly noisy; this prop-

Table 4.5. Evaluation scores of maximum of 5-best outputs for Twitter (5 is the best, significances are calculated for comparison with proposed method. $*$: $p<.1$, $**$: $p<.05$)

	Proposed	**Word**	**IR-st.**	**IR-re.**	**Rule**
(1) Related topics	**3.93**	3.76^{*}	3.40^{**}	3.76	3.04^{**}
(2) Comprehensibility	**4.59**	4.38^{**}	3.88^{**}	4.31^{**}	4.21^{**}
(3) Consistency	**4.07**	3.74^{**}	3.44^{**}	3.90	3.05^{**}
(4) Topic relations	**4.08**	3.79^{**}	3.51^{**}	3.94	2.99^{**}
(5) Topic expansion	**3.65**	**3.65**	3.05^{**}	3.36^{**}	2.74^{**}
(6) Naturalness	**4.12**	3.83^{**}	3.64^{**}	3.88^{**}	3.13^{**}
(7) Easy to continue talking	**3.87**	3.67^{*}	2.94^{**}	3.32^{**}	2.66^{**}

erty sometimes generated incomprehensible sentences. Second, to choose the 1-best outputs, we ordered the sentences using an ad-hoc criterion, which resulted in inadequate orderings. While the second reason seems to apply to the word-driven method, it is more critical for our proposed method because it has more sentence generation flexibility than the word-driven method. The second reason is supported by the result in Table 4.5, which shows the maximum scores of the 5-best outputs for the Twitter corpus. Table 4.5 shows that for most criteria, our proposed method significantly outperformed the other methods, including the word-driven method. This demonstrates that our proposed method has more expression flexibility as well as the potential to generate more suitable sentences than the word-driven method.

Comparing the two retrieval-based methods, IR-response outperformed IR-status in all the evaluation scores. This result is unexpected, since IR-status has the potential to expand dialogue topics and IR-response generates parrot utterances. Considering that the comprehensibility of IR-status is too low, despite its huge corpus, it failed to retrieve similar source sentences to the input one and generated irrelevant tweets as agent output.

4. Conclusion

We proposed a response generation method for open-domain user utterances that assembled phrase pairs into response sentences. This method generates sentences by leveraging the variety of a web-based corpus and suppressing the contamination of ir-

relevant information into input sentences. By experimentally examining the appropriateness of generated responses, our proposed method generated appropriate sentences as responses to open-domain input sentences.

Future work must improve the reranking of the orders of sentences to harvest the most preferable sentences from all generated sentences. Consistent utterance generation is another problem. Even though our experiments showed that our proposed method generated consistent sentences for the input sentences, we didn't examine the consistency in multi-turn conversations. We will tackle this problem with sentence relation analysis and modality recognition techniques. Another interesting idea is to generate sentences that affect the thinking of users, like evoking a specific emotion in them [Hasegawa et al., 2013]. This might help our method generate more appropriate utterances based on user emotions.

Chapter 5

Answering for personality questions

1. Introduction

In this chapter, we examine the effectiveness of personality QA in casual dialogues between users and conversational agents. To develop our QA system that can answer specific personality questions, such as favorite Asian foods or high school athletic participation, we adopt Person DataBase (PDB) approach that utilize question-answer pairs (*QA pairs*) evoked by a pre-defined persona. Since a wide range of topics appear in casual dialogues, large-scale QA pairs will be useful to cover a range of questions and generate appropriate answers.

In this chapter, first we explain the procedure of the development of our PDB, and analyze statistics of the PDB such as frequently asked questions or the coverage of the gathered questions for those appearing in real conversation. Second, we objectively examine the estimation accuracy of the question categories for personality inquiries that appeared in a human to human conversation corpus. Finally, we combine our personality QA system with a conversational agent [Sugiyama et al., 2014] and investigate the effectiveness of the personality QA system through chat experiments.

2. Development of a PDB

To create a number of question-answer pairs (*QA pairs*) related to an agent's personality, one approach lets people create such QA pairs for a pre-defined persona. However,

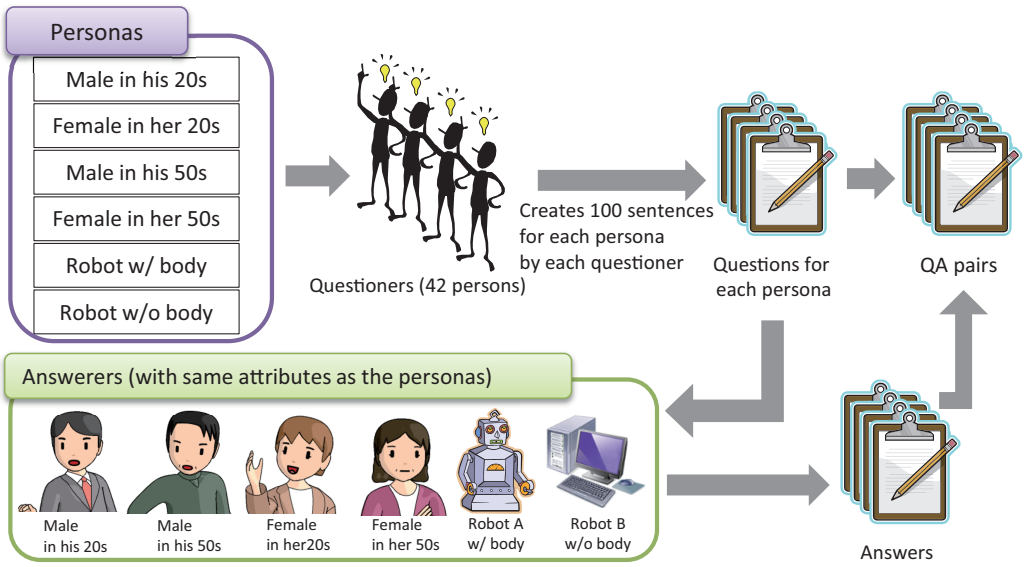


Figure 5.1. Overview of collection of QA pairs

when just a few people create them, it is difficult to know what kind of questions might be frequently asked in practice; accordingly they might overlook essential questions for conversational agents. To deal with this problem, we adopt an approach that allows many participants to create question sentences for pre-defined personae. Figure 5.1 illustrates the collection procedure of QA pairs.

First, we recruited 42 Japanese-speaking participants (*questioners*), balanced for gender and age, to create the question sentences. Each questioner created 100 or more question sentences for each of the following six personae listed in Table 5.1. The robot personae are expected to evoke different types of questions from the human personae. Each questioner created sentences under the following five rules: (A) create sentences about what he/she wants to ask naturally, (B) create sentences without omissions, (C) create one sentence for each question (no partition), (D) do not create duplicated questions for a persona (e.g., "*Where do you live?*" and "*Where is your current address?*"), and (E) do not copy questions from other sources like the web. Table 5.1 shows that we collected 26,595 question sentences.

Next, a participant called *an answerer* (not the questioner), who had the same attributes as one of the personae, created answers for the questions associated with the persona based on the following instructions: (a) create answers based on your own

Table 5.1. Persona attributes and statistics of collected PDB

Persona attributes	# of question sentences	# of question categories
(1) Human (male in his 20s)	4431	2537
(2) Human (female in her 20s)	4475	2263
(3) Human (male in his 50s)	4438	2732
(4) Human (female in her 50s)	4458	2279
(5) Robot (with body)	4426	2232
(6) Robot (without body)	4367	2665
Summation	26595	10082

experiences or favorite things, (b) create the same answers to the questions that represent identical subjects, and (c) create as many *Yes/No* answers as possible (called *Yes/No* restrictions). Instruction (a) suppresses inconsistency between answers, such as "*Yes*" for "*Do you have a dog?*" and "*No*" for "*Do you have a pet?*" However, for robot personae, the answerers create answers based on a robot character they imagined themselves. If various answers were created for identical-subject questions, it would be difficult to classify and analyze the answers. To suppress variation in the answers, we designed instructions (b) and (c). Instruction (b) directly suppresses variations, while instruction (c) is effective for question sentences that are answerable with "*Yes/No*" but are expected to be answered with specific subjects, such as "*Do you have a pet?*".

After the collection stage, the question-answer sentence pairs (*QA-pairs*) are classified to question categories, where each represents the identical subject, by another participant (not the author, not the questioner, and not the answerer) called *an information annotator*. This approach enables us to identify frequently asked question subjects based on the number of question sentences in each question category. Table 5.1 shows that the question sentences are classified to 10,082 question categories.

Finally, the information annotator annotated the collected QA pairs with the following information given in Table 5.2. We call the collected QA pairs with such information a *Person DataBase (PDB)*. Table 5.3 illustrates examples of the collected PDB.

Table 5.2. Information annotated in PDB (examples translated by the authors)

Information	Description	Examples
Question sentences	Created question sentences	*Who are you?*
Question categories	Categories that consist of question sentences which denote the same subjects	Names
Answer sentences	Created answer sentences	*I'm Taichi.*
Topic labels	Labels that consist of question categories which denote the similar subjects	Names
Answer types	Labels that represent types of answers	Name: Person names
Extended named entities	Labels that represent ENE of answers	Name: Person
Persona types	Attributes of persona	male in 20s

Table 5.3. Examples of PDB (all columns except Extended Named Entities (ENEs) translated by authors)

Question sentences	Question categories	Answer sentences	Topic labels	Answer types	ENEs	Persona
How accurately can you understand what people say?	How accurately you can understand what people say	98%	How accurately you can understand Japanese	Quantity: Other	Percent	Robot B
Do you want to hold a wedding ceremony overseas?	Whether you want to hold a wedding ceremony overseas	Yes	Whether you want to be married/Ideal marriage or family	Yes/No	No ENE	20s Male
When is your birthday?	Birthday	Sept. 10, 1986.	Birthday	Quantity: Date	Date	20s Male
Do you usually eat pancakes?	Whether you usually eat pancakes	No	Favorite sweets/Whether you like sweets	Yes/No	No ENE	20s Male
Do you buy groceries by yourself?	Whether you buy groceries by yourself	No	Preparation of dishes/grocery shopping	Yes/No	No ENE	50s Female
Do you have persimmon trees in your garden?	Whether you have persimmon trees in your garden	No	Arrangement of houses or rooms	Yes/No	No ENE	50s Female
Do you have any pets?	Whether you have some pets	No	What pets you have	Yes/No	No ENE	20s Female
Can you edit videos?	Whether you can edit videos	No	Whether you can edit videos	Yes/No	No ENE	20s Male

3. Analysis of the PDB

We analyzed the statistics of our obtained PDB, such as frequently asked questions, based on the annotated information and investigated the differences between the questions in PDB and those in real conversations.

3.1 Question categories

3.2 Statistics

First, we analyzed the number of question sentences in each question category to examine deviations in the sentences. Figure 5.2 shows their distribution, which is long-tailed (half of the question sentences belong to the top 11% (1110) categories), and 65.1% (6568) of the question categories have only one sentence.

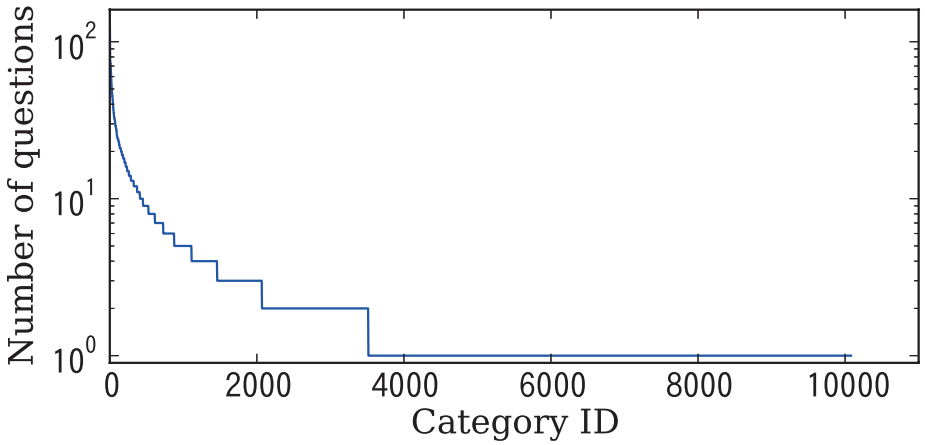


Figure 5.2. Distribution of question sentences in each question category

To reveal the frequently asked question categories and investigate the differences based on frequency, we sorted the question categories by their frequencies and divided them into four nearly equal-sized clusters. Figure 5.3 illustrates the Inverse Document Frequencies (IDFs) calculated using sentences in the conversation corpus, which we describe later (see Sec. 3.3.2). All of the differences were significant (Independent Student's t-test; $p < .01$). The difference between the top- and the high-ranked clusters is the largest. This indicates that the top-ranked cluster contains questions with fewer

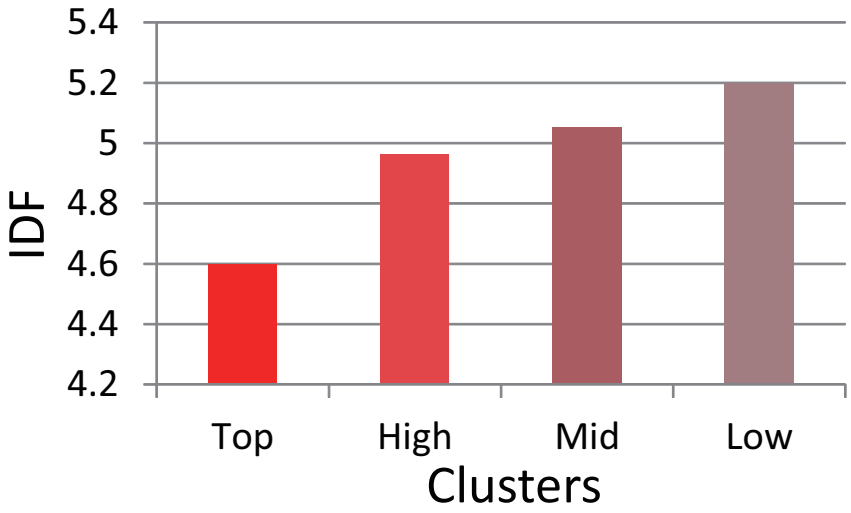


Figure 5.3. Averaged IDF values of questions in the clusters

subject-specific words (e.g., proper nouns) and the question sentences of the high-ranked cluster consist of many subject-specific words.

Table 5.4 shows the examples and the statistics of the top-, high-, medium-, and low-frequency-ranked clusters. The top-ranked question categories consist of two types of questions: Properties that all persons have, such as *Name* or *Living place*, and conversation triggers that we can easily expand a conversation, such as *Whether you like to cook* or *Whether you have a pet*. Common properties (the former type of questions) remain unchanged for each person for at least several years and have large variances among persons. These characteristics are essential factors in describing people; thus, such properties attract our interests and tend to be frequently asked. On the other hand, the latter type of questions are useful to fuel conversations, since each has a category word like *pet* or *cooking*, which leads to more fine-grained questions: "*What kind of pet do you have?*" or "*Have you ever baked a cake?*". By comparison, the questions described in a previous work [Batacharia et al., 1999] resemble our top-ranked questions. However, they did not describe some of the top 10 questions (e.g., *Whether you can drive a car*, which also ranks third in the ranking by females in their 20s). This indicates that some essential questions may be overlooked in question sets created by just a few people.

The high- and mid-ranked question categories shown in Tables 5.4(b) and 5.4(c)

Table 5.4. Examples of question categories

(a) Top-ranked categories (15 or more sentences, 1st-258th, 7403 sentences)

Question categories	#
Name	155
Birth place	111
Living place	98
Whether you can drive a car	97
Whether you have a pet	84
Whether you smoke	77
Whether you like to cook	75
Favorite color	75
Work	73
Whether you are married	73

(b) High-ranked categories (5 to 14 sentences, 259th-1110th, 6511 sentences)

Question categories	#
The number of siblings	14
Frequency of drinking alcohol	14
Time to make yourself up	14
Whether you go fishing	10
Favorite school meals	10
Whether you like museums of art	10
Color of your hair	5
Favorite rice ball ingredients	5
Favorite TV stations	5
Biggest regret	5

contain question categories that are related to the top-ranked question categories, such as *Favorite colors* and *Color of your hair*. While the high- and mid-ranked clusters are similar, mid-ranked questions have more specific subjects than do high-ranked ones because of limitations like *recently* in *The most memorable TV dramas recently*.

Some of the low-ranked question categories shown in Table 5.4(d) contain question categories subdivided from more frequent ones for the following reasons: Overly narrow subjects, limited conditions, grammatical tenses, and the *Yes/No* restriction. For example, *Whether you request life-prolonging treatment* asks about

(c) Mid-ranked categories (2 to 4 sentences, 1111th-3514th, 6113 sentences)

Question categories	#
The most memorable TV dramas recently	4
How many TV dramas you watch in a week	4
Whether you have disguised yourself as a woman	4
Anxiety about the future	3
Desirable travel companions	3
Whether you have corrected your teeth	3
Weak points (for robots)	3
Where you look at strangers	2
Whether you eat until you recover the cost in a buffet-type restaurant	2
Whether you go to public baths	2

(d) Low-ranked categories (1 sentence, 3515th-10,082th, 6568 sentences)

Question categories	#
Whether something is in fashion in your generation	1
Whether you request life-prolonging treatments	1
Whether you make accessories	1
Favorite tastes of snow cones	1
Whether you had sports heroes in childhood	1
Frequency of shaving in a day	1
Preferences of ties	1
Whether you become sensitive to earthquake after the 3.11 earthquake	1
Whether you like ball games	1
The way you search for new information	1

life-prolonging treatment that is subdivided from medical treatment. The question *Whether you had sports heroes in childhood* is subdivided from a more frequent question category *Whether you have sports heroes* because of the limitations of the term *childhood* . The question *Whether you go to public baths* is subdivided from *Whether you have been to public baths* based on different grammatical tenses. The question *Whether you go to a hot spring* is subdivided from *Your favorite hot springs* since the former should be answered with *Yes/No* but the latter should not because of the *Yes/No* restriction.

Figure 5.4 illustrates the variation in the number of cumulative question categories while the number of questioners increases. The top-, high- and mid-ranked clusters were saturated with only 5, 15 and 35 questioners, respectively. In contrast, even though we expected the increase in the low-ranked cluster to become slow with 40 questioners, the low-ranked cluster did not saturate and increased almost linearly. This indicates that the number of conceivable personal questions is huge.

Table 5.5 shows the ranking correlations of the orders of the top- and high-ranked question categories among the personae. This shows that the orders among human personae are not so different ($0.25 - 0.53$ of rank correlations). A male in his 50s showed lower correlations than did the other human personae; in particular, the correlation with a female in her 20s is the lowest (0.25). One characteristic difference between human personae is that questions of similar subjects are described differently depending on the individual's life stages, such as *Whether you have a boyfriend/girlfriend* and *Whether you are married*. In contrast, the rank correlations between the human and robot personae are negative. Table 5.6 illustrates the most frequently asked questions whose associated personae include Robot A or Robot B, and the questions are ranked in a ranking constructed using only questions associated with human personae. This shows that some of the questions do not appear with the human personae. For example, *Whether you can run* and *Whether you have emotions* are related to robot abilities or properties that obviously exist for humans. Therefore, to develop a PDB that can be used for conversational agents, it is necessary to apply robot personae in collecting personal questions.

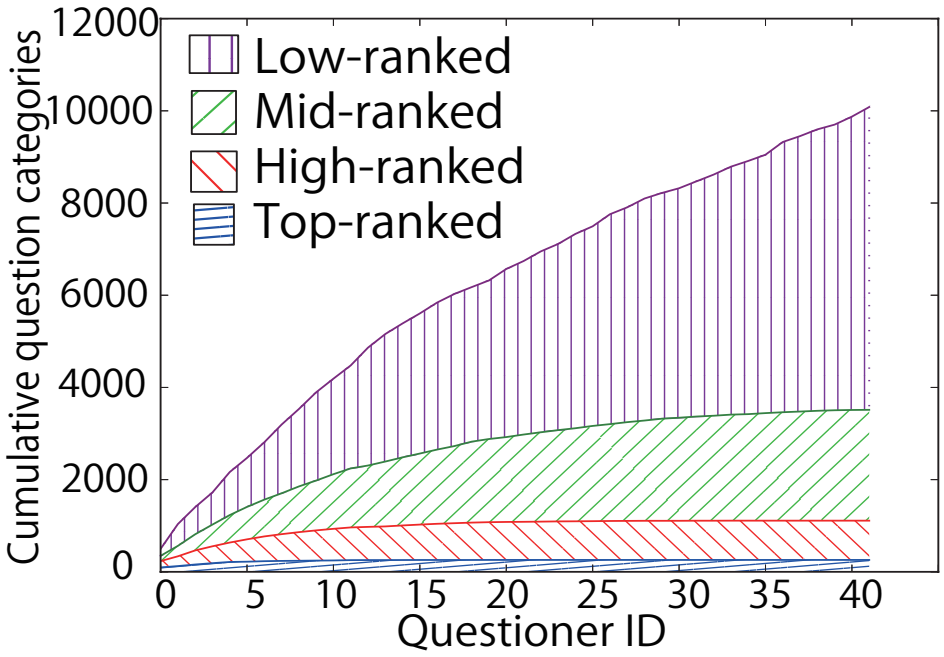


Figure 5.4. Variation in cumulative question categories of each cluster while increasing the number of questioners

3.3 Comparison with the conversation corpus

To compare the personality questions in PDB with those in real conversations, we extracted personality questions from the conversation corpus gathered by Higashinaka et al. [Higashinaka et al., 2014], which contains 3680 conversations based on text chat (with 134 K sentences). From 183 sampled conversations, we harvested 490 personality questions (7.8% of all sentences and 72.0% of all questions) with corresponding question categories labeled by two annotators (not the authors). The agreement rate of the labeled question categories between the annotators was 0.816.

The number of questions classified in each cluster is shown in Fig. 5.5: 85 (17.3%) top-ranked, 52 (10.6%) high-ranked, 29 (5.9%) middle-ranked, and 36 (7.3%) low-ranked. The other 288 questions (58.7%) were not included in our PDB. We assumed that these numbers were almost the same each other since the clusters have similar number of question sentences; however, the top-ranked cluster was associated with many questions in the conversation corpus. This is due to the short length of conversations in our corpus, which averaged 36.5 sentences for each conversation.

Table 5.5. Ranking correlations of the orders of top- and high-ranked question categories among personae. While correlations of human-human personae show high scores, that of human-robot personae shows negative correlations.

	20s M	20s F	50s M	50s F	Robot A	Robot B
20s M	1.00	0.53	0.37	0.37	-0.23	-0.06
20s F	0.53	1.00	0.25	0.47	-0.26	-0.05
50s M	0.37	0.25	1.00	0.41	-0.19	-0.09
50s F	0.37	0.47	0.41	1.00	-0.22	-0.10
Robot A	-0.23	-0.26	-0.19	-0.22	1.00	0.38
Robot B	-0.06	-0.05	-0.09	-0.10	0.38	1.00

Figure 5.6 shows the cumulative number of question sentences in the conversation corpus that were included in our PDB while the number of questioners is increased. This demonstrates that only one questioner could create most of the top-ranked questions; however, some of the top-ranked questions were overlooked, and few questions assigned to the other clusters were created. The coverage rate improved linearly until about 20 questioners were used, and with over 20 questioners, the improvement in each cluster except low-ranked became saturated. Consequently, we consider 20 questioners a reasonable number to collect a sufficient number of personal questions. Even though larger-sized PDB could possibly contain more questions in the corpus, since the low-ranked cluster improved steadily, this seems unsuitable when considering the slowdown of improvement at 20 questioners.

To answer the rest of the questions (*None* in Fig. 5.5), it is important to analyze the reason why they do not appear in our PDB. Table 5.7 shows the classification of such reasons and their examples, and Fig. 5.5 illustrates the proportion of each reason. In Fig. 5.5, 71.1% (205) of the reasons were (a) and (b). They contain specific words, such as *Gero-onsen spa* and *for Italy*, which limits the question subjects. For questions with an answer type of *Yes/No*, specific questions can be easily created with typical phrases and specific words such as "*Do you know Gero-onsen spa?*" or "*Do you like Gero-onsen spa?*". Even though (a) seems to be answerable since the answer type is *Yes/No*, it is difficult to maintain consistency with the other answers. The answer "*No*" will help to avoid such inconsistency, but agents that always say "*No*" would irritate users. The questions for reason (b) can also be answered using stochastic response-

Table 5.6. Examples of question categories of robot personae. *Rank* means the rank in a ranking, which is created with question sentences for human personae (N/A means that the category does not appear for human personae).

(a) Robot A (embodied)

Question categories	Rank
Name	1
Weight	42
Place of production	N/A
Whether you can cook	13
Height	29
Whether you can write a character	N/A
Whether you can run	N/A
Birthday	33
Whether you can sing a song	59
Whether you can drive a car	6

(b) Robot B (no body)

Question categories	Rank
Name	1
Whether you can sing a song	59
Whether you are male or female type	42
Birthplace	2
Manufactured purpose	N/A
Birthday	33
Whether you can change your voice	N/A
Whether you have emotion	N/A
Age	17
Manufacturer's name	N/A

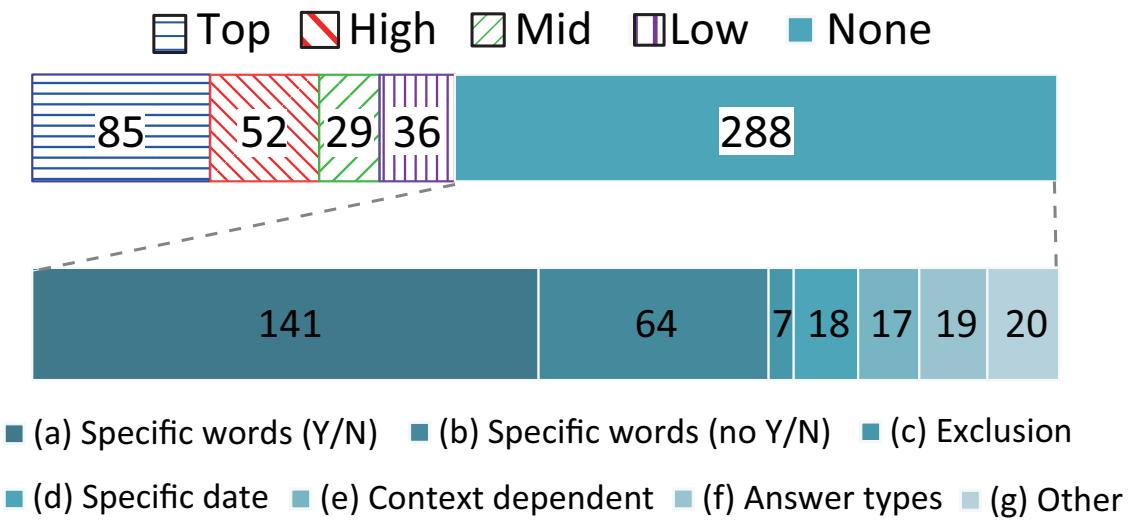


Figure 5.5. Cluster assignment of questions in the conversation corpus

generation methods [Sugiyama et al., 2013], which generate sentences related to user utterances by leveraging word dependencies in a corpus; however, it is also difficult to avoid inconsistency here. The questions of reason (e) are more difficult to answer, since such questions require understanding of deep context of conversation to generate answers.

On the other hand, the questions of reasons (c) and (d) can be easily answered by substituting them with the other question categories that do not have such limitations or exclusions as *today* or *besides the clarinet*. In the questions associated with reason (f), those whose answer type is *Yes/No* and the corresponding factual questions are included in the PDB are answerable using the answers associated with such corresponding questions. For example, "*Do you have a favorite actor?*" can be answered with "*Audrey Hepburn*" that is associated with the corresponding questions like "*Who is your favorite actor?*", which can be retrieved based on the sentence similarity of the user's utterances and the question sentences in the PDB. On the contrary, when the PDB has only *Yes/No* answer type questions, since this indicates that the PDB has no information about favorite actors, *No* is a suitable answer to the questions.

By using the above solutions, we can recover 44 (15.2%) questions [(c),(d),(f)] with the PDB itself; this improvement increases the coverage rate to 50.2% (246/490), which appears to be nearly the upper bound of this approach. Even though the other

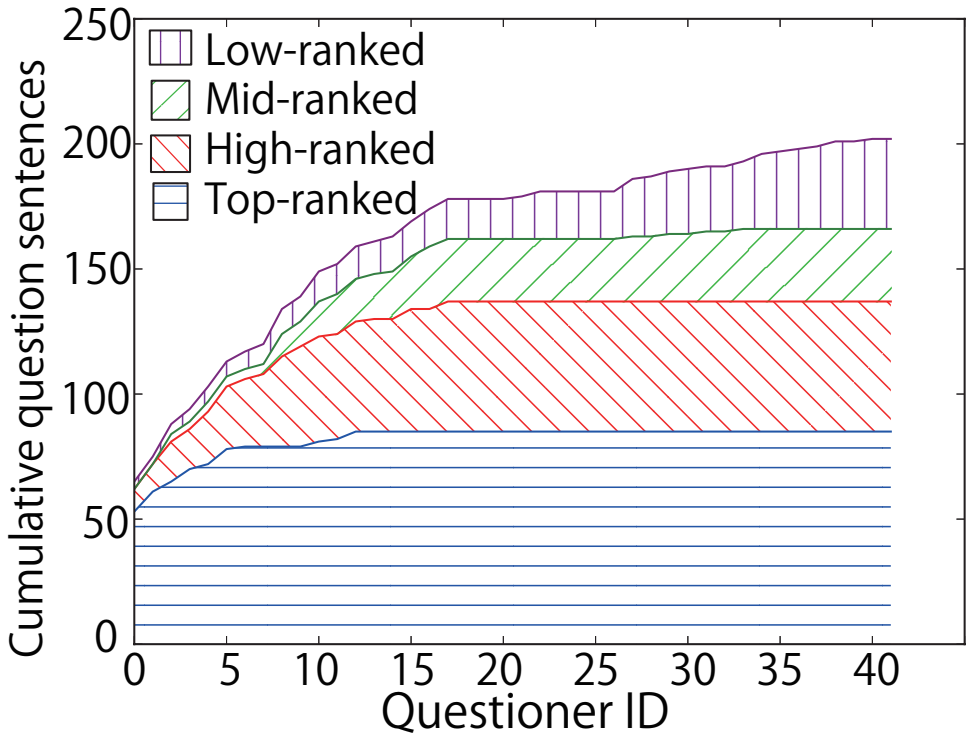


Figure 5.6. Cumulative number of question sentences included in the PDB as questioners increase

141 (48.9%) questions of (a) are answerable with a large text corpus and 64 (22.2%) questions of (b) are potentially answerable with stochastic response generation methods, they risk conflicting with other answers.

3.4 Answer types and Extended Named Entities

Classification of the answers gives us another view of the analysis that reveals the subjects of frequently asked questions. Table 5.8 shows the distribution of the answer types with associated Extended Named Entities (ENEs) [Sekine et al., 2002] that consists of about 200 Named Entity types. The answer types were originally defined by Nagata et al. [Nagata et al., 2006]. We added a small modification that integrates *Explanation: cause* and *Explanation: principle* to *Explanation: reason* and obtained 21 answer types.

The most frequent answer type was *Yes/No*, accounting for 60% of the PDB. The

Table 5.7. Reasons why questions were not included in PDB

Reason	Examples	#
(a) Limited by specific words (answer type: Yes/No)	Do you know Gero-onsen spa?	141 (48.9%)
(b) Limited by specific words (answer type: not Yes/No)	What is your recommendation for Italy?	64 (22.2%)
(c) Includes words that exclude a specific word (e.g., besides)	What other instrument have you played besides the clarinet?	7 (2.4%)
(d) Limited to a specific date or time (e.g., today's lunch)	What did you eat for lunch today?	18 (6.2%)
(e) Assumes a conversation context	What kept you up so you couldn't sleep until morning?	17 (5.9%)
(f) Different answer types (Yes/No or factoid questions)	How long have you lived at your current address?	19 (6.5%)
(g) Other	What delicious dishes have you cooked?	20(6.9%)

PDB contains many noun-driven questions such as "*Do you like sushi?*" or "*Do you know Woodstock from Snoopy?*" because they can be created to match the number of nouns. The *Yes/No* restriction also accelerates the appearance of such answer types. In the conversation corpus, the most frequent answer type is *Yes/No* (143/202, 70.7%), and the second is *Name: named entity* (41/202, 20.2%); the other answer types scarcely appeared in the corpus.

Except for *No ENE*, most ENEs are annotated to the questions whose answer types are *Name* or *Quantity*, since the ENEs are defined to categorize such entities. Table 5.9 shows the frequently appearing ENEs and example sentences for each of the answer types. Over half of the questions associated with the *Name* answer type were labeled with the most frequently 10 ENEs. In the conversation corpus, except for *No ENE*, the most frequent ones were *Dish* (10), *Movie* (9) and *Music* (5). Since the speakers in our conversational corpus used pseudonyms to speak to another speaker, the questions with ENE *Person* did not appear in the conversation corpus.

In the answer types *Explanation* and *Quantity*, the most frequent sub-type was *Other*, which indicates that our answer types were inadequate to classify them. Even though the ENEs complemented the classification of *Quantity*, they could not com-

Table 5.8. Answer types and Extended Named Entities (ENEs)

Major type	Sub-type	ENE examples
Yes/No (16,255)	Yes/No (16,255)	No ENE (16,225)
Explanation (2528)	Association (70)	Person (18), Dish (18), No ENE (16)
	Reputation (447)	No ENE (447)
	Reason (56)	No ENE (56)
	True identity (2)	No ENE (2)
	Method (243)	No ENE (242),Dish (1)
	Meaning (64)	No ENE (59),Product_Other (5)
	Other (1646)	No ENE (1632),Incident_Other (7)
Quantity (2337)	Money (235)	Money (235)
	Period (354)	Period_Time (251),Period_Year (84)
	Hour (31)	Time_Top_Other (19),Era (8)
	Time (134)	Time (134)
	Date (135)	Date (135)
	Other (1448)	Frequency (276),N_Product (239),Age (218)
Name (4339)	Organization name (320)	Company (192),Show_Organization (48)
	Location name (836)	Province (260),Country (144)
	Named entity (2571)	Dish (415),Product_Other (337)
	Person name (444)	Person (444)
	Web site (11)	Product_Other(11)
	Other (157)	No ENE (151),Name_Other (6)
Other (1136)	Selection (1012)	No ENE (275),Dish (169)
	Description (2)	No ENE (2)
	Phrase (122)	No ENE (121),Public_Institution (1)

plement the classification of the questions whose answer type was *Explanation*, since the answers associated with these questions are described with sentences, not with the words for which the ENEs are designed.

Table 5.9. Frequent ENEs

(a) Name (# is 4339)

ENE name	# of sentences
Person	444 (10%)
Dish	415 (9%)
Product_Other	348 (8%)
Province	260 (5%)
Company	192 (4%)
Position_Vocation	192 (4%)
Music	161 (3%)
Country	144 (3%)
Sport	125 (2%)
Food_Other	115 (2%)

(b) Quantity (# is 2337)

ENE name	# of sentences
Frequency	276 (11%)
Period_Time	251 (10%)
N_Product	239 (10%)
Money	235 (10%)
Age	218 (9%)
N_Person	173 (7%)
Date	135 (5%)
Time	134 (5%)
Physical_Extent	114 (4%)
Period_Year	84 (3%)

3.5 Topic labels

We annotated question sentences using the *Topic labels* to aggregate question categories. These labels are designed to integrate the question categories that have similar subjects but are divided because of the differences of limitations, answer types, and grammatical tenses. We expect these labels to enable us to investigate the subjects of the questions more clearly. The information annotator aggregated question categories into topic labels based on the following aggregation types.

(a) Specific words This aggregates question categories with specific words into one topic label with more abstract words: e.g., *Whether you tend to be angry* and *Whether you are romantic* into *Your Character*.

(b) Resemble questions This integrates similar question categories with nuanced differences: e.g., *Whether you like cooking*, *Whether you are good at cooking* and *Whether you can cook*.

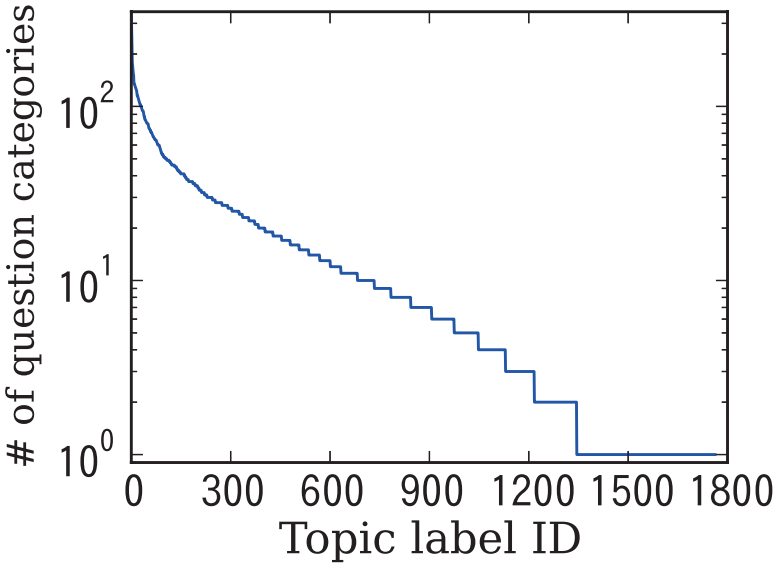


Figure 5.7. Distribution of question categories by topic labels

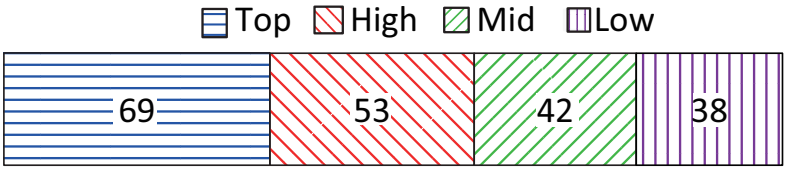


Figure 5.8. Questions in conversation corpus assigned to each topic label cluster

(c) Negative questions This integrates negative and non-negative question categories with the opposite meaning: e.g., *Whether you don't dislike some foods* with *Whether you dislike some foods*.

(b) Answer types This integrates question categories where only the answer types are different: e.g., *Whether you have a favorite sport* and *Your favorite sport*.

First, we counted the question categories included in each topic label to examine how the topic labels aggregate the question categories. Figure 5.7 illustrates the distribution of the numbers of question categories in each topic. The total was 1763, the average number of question categories by each topic was 5.71, and the average number of answer types by each topic was 1.95. While only 40% of the question categories were associated with two or more question sentences, 70% of the topics consisted of two or more question categories.

Table 5.10. Examples of topic labels

(a) Top-ranked labels (20 or more categories, 1st-77th, 2576 categories, 5779 sentences, average number in each category is 2.24)

Topic labels	# of category
Favorite foods or dishes	137
Character	121
Opinions on politics	78
Physically available movement	64
Your appearance	62
Meal style/custom	62
Policy of child care/education	55
Ideal sort of marriage partner	54
School life	54
Relationship with your family	51

(b) High-ranked labels (19 to 11 categories, 78st-171th, 2384 categories, 6682 sentences, average number in each category is 2.80)

Topic labels	# of category
Something you played with in childhood	19
Whether you play sports/Exercise habits	19
Whether you have a child/Personality of the child	19
Whether you have a favorite place	19
Using SNS	18
Person who chooses your clothes	11
Whether you like/dislike bugs	11
Whether you enjoy work	11
Weak point/Inferiority complex	11
Whether you like yourself	11

Next, we sorted and classified the topic labels into four clusters in the same manner as the question categories. Table 5.10 shows the clusters of topic labels. Figure 5.8 gives the number of questions assigned to each cluster. Table 5.10(a) shows that the top-ranked topic labels contain over 50 question categories. Since the subjects of

(c) Mid-ranked labels (10 to 6 categories, 172st-337th, 2561 categories, 6995 sentences, average number in each category is 2.73)

Topic labels	# of category
Favorite books/Whether you like reading books	10
Whether you like taking trips	10
Favorite downtown place	10
Whether you like art/art museums	10
The way you take a bath	7
Your favorite rice	7
Whether you like the sea or mountains	6
Amusement parks you have visited	6
Frequency of going to beauty shop	6
Favorite sushi	6

(d) Low-ranked labels (5 to 1 categories, 338th-1178th, 2561 categories, 7139 sentences, average number in each category is 2.78)

Topic labels	# of category
Whether you agree with holding the Tokyo Olympics	5
Whether you have a debt/Sum of the debt	5
Whether you have a credit card	5
Political affiliation	5
Whether you are a morning or nocturnal person	5
Mental age	1
Whether you want to do cosplay	1
Whether you wear sunglasses	1
Whether you have hit someone	1
Whether your first love reached	1

question categories contained with these topic labels had subject-specific words such as *Ethnic* in "*Do you like Ethnic dishes?*", these topic labels were composed of a large number of question categories. On the other hand, in the low-ranked topic labels, the categories appearing once had specific subjects such as *cosplay*; this is the reason why such categories were not integrated with the other categories.

4. Experiments

Our QA system estimates question categories to generate answers for input question sentences. In this section, first we objectively evaluate the estimation accuracy of the question categories. After that, we investigate the effectiveness of our personality QA system by examining response appropriateness and user-machine chat experiments.

4.1 Objective evaluation: Estimation accuracy of question categories

4.1.1 Experiment settings

In this section, we compare the estimation accuracy among a combination of estimation methods and features. For the estimation methods, we adopt the following three estimators: The RBF-kernel SVM, the linear SVM, and cosine-based retrieval. The former two SVM approaches directly estimate question categories, and the latter finds the most similar question sentence to an input question sentence. When we calculate estimation accuracy, we use the question category associated with the retrieved question sentence. We experimentally determined each estimator's parameters. We also compared the estimation accuracy among the following typical classification and retrieval features: Bag-of-words (unigram or uni+bigram) vectors extracted from the question sentences, and those weighted with TF-IDF, which are calculated using words contained in the PDB and the conversation corpus, which treats each sentence as a document.

To train the estimators, we limited the estimation target to frequent question categories (1,110 categories and 13,917 sentences) that appear five or more times in PDB without considering the personas, since the estimators require at least a few training examples to estimate the question categories.

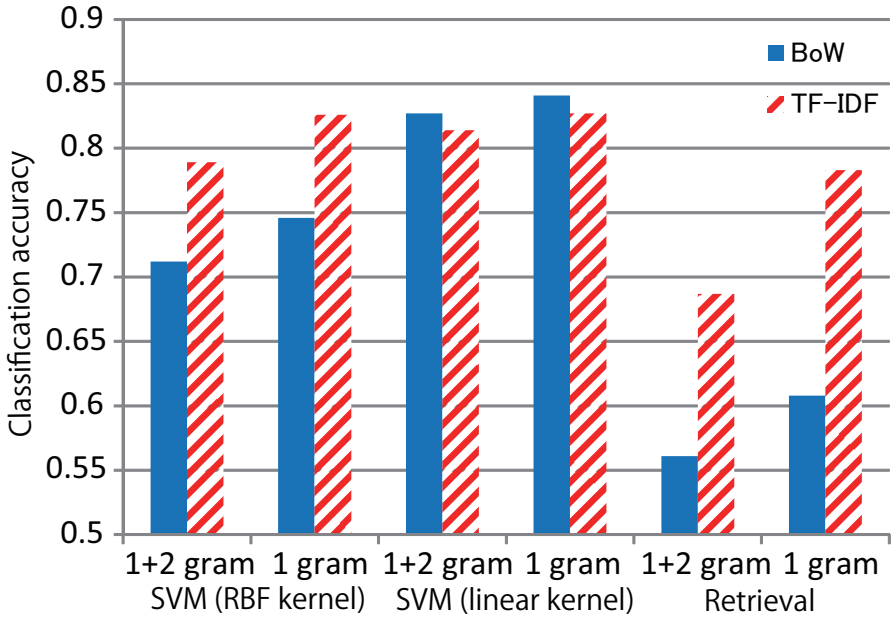
We evaluated the estimation accuracy of the question categories using two types of question sentences: Those contained in the PDB and those sampled from the conversation corpus gathered by Higashinaka et al. [Higashinaka et al., 2014], which contains 3,680 conversations based on text chats (134 K sentences). We evaluated the estimation accuracy of the PDB sentences with 5-fold cross-validation with 13,917 sentences. For the sentences in the conversation corpus, first we sampled 166 personality question sentences from the corpus and two annotators (not the authors) annotated the question categories to the sentences (Cohen's κ: 0.816). In the 166 sentences, 51 personality question sentences are associated with the frequent question categories by both the annotators. We trained the classifier with all the PDB sentences and evaluated their accuracy with the 51 personality question sentences.

4.1.2 Result and discussion

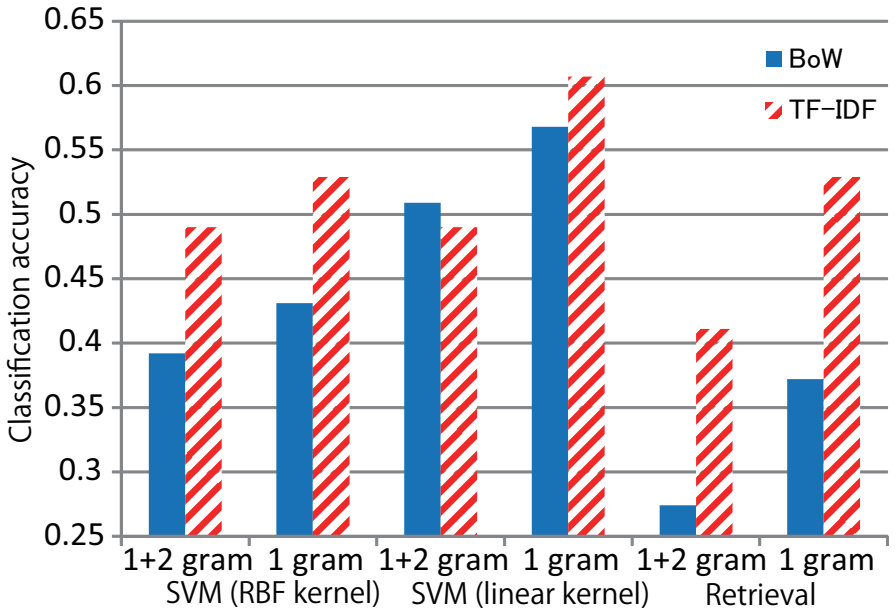
Figures 5.9 (a) and (b) show the estimation accuracy of the question categories for the PDB and the conversation corpus. The linear SVM with unigrams showed the highest accuracy with 0.841 for PDB (without TF-IDF) and 0.607 for the conversation corpus (with TF-IDF). The RBF-kernel SVM and cosine-based retrieval showed lower accuracy probably because of a lack of data for each category. In the comparison between the raw bag-of-words and the TF-IDF weighted ones, the latter outperformed the former in most of the settings, suggesting that TF-IDF compensates for the lack of data.

Figure 5.9 also shows that the estimation accuracy for the conversation corpus is lower than that for the PDB. This is because the personal question sentences sampled from the conversation corpus have many words that are not contained in the PDB. Besides, contrary to the PDB where question sentences contain only the necessary information to represent the questions, the sentences in the conversation corpus sometimes contain words unrelated to the questions, such as the introduction of question topics.

Next, we analyzed the incorrect estimations based on one trial of 5-fold cross-validation. Table 5.11 shows the error-categories, their ratios, and their examples. 29.8% of the errors can be categorized as *almost identical meaning*, where the question sentences were correctly answered with the estimated question categories, such as *whether you can ride a bicycle* as the correct category and *whether you usually ride a*



(a) PDB



(b) Conversation corpus

Figure 5.9. Accuracy of question categories: Majority baseline in (a) PDB is 0.011 (155/13917).

Table 5.11. Error-categories, their ratios, and their examples (translated by authors)

Almost identical meaning (Categories differed by tense or modalities)	29.8% 135/453	Q sent.: Have you ever ridden a bicycle? Correct: Whether you can ride a bicycle Est.: Whether you usually ride a bicycle
Different grain size of topics	9.1% 41/453	Q sent.: Are you interested in fashion? Correct: Whether you pay attention to clothes Est.: Whether you have favorite brands of clothing or bags
Different topics	35.8% 162/453	Q sent.: Do you color your hair? Correct: Whether you dye (have ever dyed) your hair Est.: Whether you have a job now
Answer-type mismatch	10.3% 47/453	Q sent.: How often do you drink a week? Correct: Frequency of drinking in a week Est.: Favorite kind of alcohol
Annotation error (Existence of more appropriate category)	7.7% 35/453	Q sent.: Do you usually drive? Correct: Whether you can drive a car/whether you have a driver's license Est.: Whether you drive a car frequently
Character mismatch (Categories varied by character)	1.5% 7/453	Q sent.: What is your favorite opposite sex type? Correct: Favorite female type Est.: Favorite male type

bicycle as the estimated category. The most frequent (35.8%) error type was *different topics*, which unfortunately cannot be answered with the estimated question categories, such as *whether you are attentive to clothes* as the correct category and *whether you have work now* as the estimated category. Such errors were caused when important words in the question sentences were not covered in the estimator's training data. We assume that thesauri like WordNet or word-clustering methods like brown-clustering [Turian et al., 2010] or word2vec [Mikolov et al., 2013] can improve the estimation accuracy of such question sentences. Answer-type mismatch explains about 10.3% of the errors. This error can be reduced by using a separate classifier for the answer-types.

4.2 Subjective evaluation 1: Response appropriateness

As a subjective evaluation, first we examined the effectiveness of a personality QA system through module-based experiments that compared the appropriateness of the sentences generated by our QA system and a conventional approach (without our per-

sonality QA system) on input personality questions.

4.2.1 Experiment settings

Based on the objective evaluation results, we adopted linear SVM with TF-IDF weighted unigrams as features of the estimators of the question category. When a personality question sentence is given, our personality QA system estimates the question category that corresponds to the question sentence and returns an answer that is randomly selected from answer sentences associated with the estimated question category.

For the conventional agent, we adopted an open-domain conversational agent proposed by Sugiyama et al. [Sugiyama et al., 2014]. Despite the wide range of topics that appear in casual dialogues, this agent generates agent utterances related to the user utterances by assembling two phrase-pairs: One extracted from user utterances and another from a Twitter corpus that has dependency relations with the former pair. Since Sugiyama's agent outperforms other rule-based and retrieval-based agents and has no question-answering function, it is reasonable to examine the effectiveness of our personality QA system.

We randomly selected 52 personality questions from the conversational corpus (Sect. 4.1), and our QA and conversational agents generated responses for each sentence. We shuffled the sentence orders for each input sentence, and two annotators (not the authors) subjectively labeled them on a 7-point Likert scale in terms of response naturalness. The Spearman's correlation coefficients of the two annotators were 0.71. We used the mean labeled scores for our evaluation.

4.2.2 Results

Figure 5.10 shows that the response sentences generated by our personality QA system acquired significantly higher scores (4.1) than the conventional agent (2.4) ($p < .05$ with a Wilcoxon signed-rank test). This indicates that the personality QA's capability adequately responds to personality questions. However, its evaluation scores are less satisfactory, since some questions are too specific and could not appropriately be classified into our question categories. For example, the question, *What are you going to do today?*, in Table 5.12 is common, so our QA system's response was appropriate.

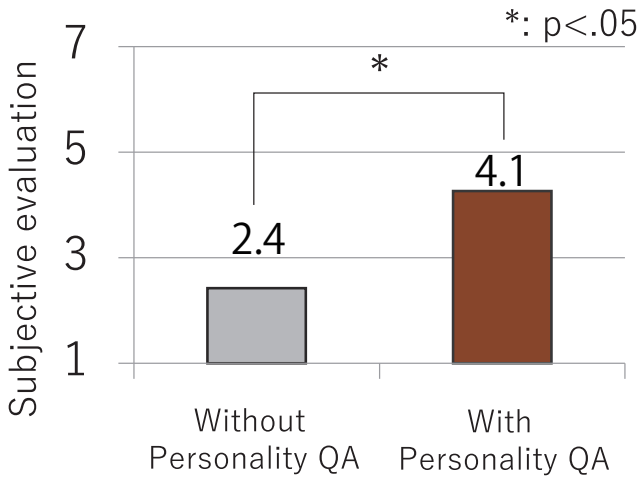


Figure 5.10. Comparison of appropriateness of one-turn responses for personality questions with 7-point Likert scale

Table 5.12. Examples of question sentences, answer sentences, and their evaluation scores (translated by authors)

Input questions	Answer sentences	Scores
What are you going to do today?	With PQA: I'm going to watch a movie.	7
	Without PQA: Yes. What are your plans?	1
Do you like wasting time?	With PQA: I like drawing postcards. How about you?	2
	Without PQA: I got it. What do you like to collect?	1

On the other hand, since the question, *Do you like wasting time?*, is not so common, neither system generated appropriate answers, which are evaluated with scores 1 or 2 that decreased the mean of the evaluation scores.

4.3 Subjective evaluation 2: Online-chat experiments

We also examined the effectiveness of our personality QA system through user-machine chat experiments.

Table 5.13. Objective evaluation scores on 7-point Likert scale ($*$: $p<.1$, $**$: $p<.05$)

	Without QA	**With QA**
(1) Naturalness of dialogue flow	2.96	3.52^{**}
(2) Grammatical correctness	3.68	4.02^{**}
(3) Appropriateness as one-turn response	3.12	3.73^{**}
(4) Semantic consistency	2.66	3.50^{**}
(5) Dialogue usefulness	2.50	3.27^{**}
(6) Ease of considering next utterance	3.04	3.75^{**}
(7) Variety of agent utterances	3.02	3.64^{**}
(8) User motivation	3.96	4.50^{**}
(9) Agent motivation felt by user	3.79	4.14^{**}
(10) Desire to chat again	2.79	3.27^{*}

4.3.1 Experiment settings

Since our personality QA system is not designed to respond to user utterances except for personality questions, we combined our personality QA system with another conversational agent that can handle such utterances to perform chat experiments. In this study, we used aforementioned open-domain conversational agent [Sugiyama et al., 2014]. When a user's utterance is estimated to be a personality question, the personality QA system generates agent utterances; if it is estimated to be a normal utterance (not a personality question), the conversational agent generates a agent utterance.

We recruited 30 native Japanese-speaking participants whose ages ranged from 20 to 50 (controlled by gender and age) from outside the authors' organization who have experience using chat agents. Each participant talked with the conversational agents with and without the personality QA system and provided subjective evaluation scores for both agents for each of the ten criteria shown in Table 5.13 using 7-point Likert scales.

Each ten-minute dialogue was conducted through a text chat interface; the presentation orders of the agents were randomized. Since it takes about 30 to 60 seconds for the agent to generate responses, about 20 utterances were performed in a dialogue.

4.3.2 Results

Table 5.13 shows that personality QA system outperformed that without the QA system in all the criteria. The dialogue examples in Table 5.14 show that the users sometimes asked the agent personality questions, which it couldn't answer without the personality QA. This failure caused the entire dialogue to collapse. On the other hand, the personality QA system answered the personality questions. In this example, although the user's factoid questions couldn't be answered, the dialogue's thread was recaptured by other personality questions. These results indicate that the capability of the personality QA is both required and useful for conversational agents to prevent dialogues from breaking down.

5. Conclusion

We developed a question-answering system for questions that ask about a conversational agent's specific personality using large-scale question-answer pairs called a Person DataBase (PDB). Our objective evaluation experiment showed that our personality QA system estimated appropriate question categories with 84.1% accuracy for question sentences in PDB and 60.7% for a conversation corpus. We also analyzed the incorrect estimations and determined that 29.8% of the errors are not critical; other errors can be improved. Furthermore, we examined our personality QA system's effectiveness through subjective evaluation experiments, which indicated that it is useful to prevent the breakdown of conversation.

Table 5.14. Example dialogues translated by authors. Underlined sentences are personality questions.

(a) Without personality QA

user	What are you doing?
agent	OK. Do you usually watch TV?
user	I don't watch TV because I don't even have one.
agent	Oh, it's a TV. Please watch TV when you get back home.
user	But I don't have a TV!
agent	Well, you're at home. It's dangerous after you get back home.
user	What's dangerous?
agent	That's right.
user	What are you talking about? What's your favorite TV program?
agent	That's right.
user	Please tell me about your favorite TV program.
agent	That's right.
user	What's your hobby?
agent	Well, what's your hobby?
user	Tell me first.
agent	Yes.
user	What do you like?
agent	Yeah. What do you like?
user	You're playing strange games with me. I like fish paste.
agent	Let's come to like fish paste parfait.
user	What's that?

(b) With personality QA

user	Hello!
agent	Yeah. What's your hobby?
user	Hobby? Well, I've been doing lots of bowling recently.
agent	OK.
user	What's your hobby?
agent	Reading.
user	What have you read recently?
agent	*Do Androids Dream of Electric Sheep?*.
user	Who is the author?
agent	Oh, who is it?
user	You seem to answer without really answering...
agent	I feel sick when it pushes and is pushed.
user	Well, what book do you recommend?
agent	*The Brothers Karamazov*.
user	That's really long, isn't it.
agent	That's right. Please recommend some more books since autumn is my favorite time for reading.
user	Well, I haven't read any books recently.
agent	Oh, reading? It's like autumn and I like it.
user	Speaking of autumn, sauries are good.
agent	It is like autumn and I like it

Chapter 6

Automatic evaluation of conversational dialogue agents using large-scale multi-references

1. Introduction

We propose a regression-based approach that automatically evaluates conversational dialogue agents using large-scale multi-references. Traditional reference-based evaluation works have two drawbacks: One is a small number of references, and another is naive handling of negative references. Unlike traditional reference-based evaluation methods that utilize a small number of references, To overcome the former drawback, we gather many references that contain both positive and negative references from human and dialogue agents. To improve the handling of negative references, we utilize discriminative approach such as regressions, instead of previously adopted generative approach that uses sentence similarities as evaluation scores like BLEU score, which is a popular metric in machine-translation area.

2. Multi-reference based evaluation

Our proposed system estimates the evaluation scores of conversational dialogue agents using many positive and negative reference sentences. This section explains how we

gather negative sentences by humans, evaluate them among the annotators consistently, and automatically estimate their evaluation scores.

2.1 Development of reference corpus

We develop a multi-reference corpus that contain both positive and negative reference sentences (responses to input sentences). To collect reasonable input-responses pairs for automatic evaluation, first we collect understandable input sentences. To collect such input sentences, we gather sentences from the Web and real dialogues between humans and rate their understandability scores. From these sentences, we randomly choose sentences that receive high understandability scores.

After collecting the input sentences, reference writers create response sentences that would satisfy users. To intentionally gather inappropriate responses, we design the following two constraints for their creation: Character-length limitation and masked input sentences. The character-length limitation, which narrows the available expressions, decreases the naturalness of the references. The "masked input sentences" means that we delete some words of the input sentences. This enables us to gather sentences that have irrelevant content. In addition, in order to add other types of inappropriate sentences to the negative references, we gather sentences that are generated by existing dialogue agents.

2.2 Evaluation of references

Human annotators evaluate reference sentences in terms of their naturalness as responses. In this work, we adopt a **pairwise winning rate** over all the other references as an evaluation score of a reference sentence. If a sentence is judged more natural than all the other references, its evaluation score is 1; a sentence that is judged the least natural obtains a 0 as its evaluation score. Our preliminary experiment shows that if the evaluation scores are rated on a 7-point Likert scale, they tend to be either the maximum or minimum; 45% were rated as 7 and 25% as 1. Hence it is difficult to determine the differences among the references by their scores. On the contrary, the winning rates can vary broadly, and we can distinguish the differences among the references.

The drawback of the winning rate is its evaluation cost; the number of pairwise

evaluations is $N(N-1)/2$. However, it has been reported that pairwise evaluations for sampled pairs are satisfactory to maintain the accuracy of the winning rates [Sculley, 2009].

2.3 Score estimation methods

To automatically evaluate the appropriateness of agent response sentences using the gathered pairs of input-references (positive and negative references) with evaluation scores, we consider the following three approaches.

Average of metrics (AM) When this method calculates an evaluation score of a agent response to a given input sentence, first it calculates sentence similarities like sentence-BLEU [Lin and Och, 2004] between the agent response and all references associated with the input sentence. Then, it outputs an average of the top-N similarities as its evaluation score. In the case of BLEU, the evaluation score E_{am}^{bleu} is defined as:

$$E_{am}^{bleu} = \frac{\Sigma_{n=1}^{N} BLEU_n}{N}, \tag{6.1}$$

where $BLEU_n$ is the nth best BLEU value. This utilizes only the similarities and resembles the approach in machine-translation. Here, we use only manually created references without the masking constraint.

Weighted scores (WS) This method outputs the weighted average of the scores r_n (winning rates) of the top-N similar references. Here, we used the similarity metric values as the weights as:

$$E_{ws}^{bleu} = \frac{\Sigma_{n=1}^{N} BLEU_n \cdot r_n}{N}. \tag{6.2}$$

Regression This estimates the evaluation scores with regression models like Support Vector Regression (SVR) [Smola and Schölkopf, 2004]. We use sentence similarities such as BLEU or Word error rates between a target sentence and the created references as features for the estimation. In this thesis, to examine the differences of the generative (WS) and discriminative (regression) models, we use the same features as the weighted scores approach; i.e., we do not use words or embedding vectors of words for the estimation. We develop a regression model for each input sentence.

3. Experiments

First we gathered the pairs of input and reference sentences with evaluation scores. Then, based on the references, we developed evaluation score estimators and examined the effectiveness of our multi-reference approach.

3.1 Settings

3.1.1 Input sentences

We sampled input sentence candidates from a chat-oriented dialogue corpus and Twitter. The chat-oriented dialogue corpus consists of 3680 text-chat dialogues between humans without specified topics [Higashinaka et al., 2014]. From the corpus, we extracted input sentence candidates whose dialogue-acts were related to self-disclosure. This eliminates such sentences like greetings and factoid questions that decrease the variety of responses. From Twitter, we sampled sentences that contain topic words, which were extracted from the top-10 ranked terms of Google trends in 2012 in Japan[1]. We sampled 10,000 tweets as the input sentence candidates.

To remove such candidates that require the contexts of the original dialogues for the writers to understand, we recruited two annotators who rated the comprehensibility of the sentences on a 5-point Likert scale and only used the sentences as candidates that both annotators rated 5 as the candidates. We used ten input sentences for the following experiments: Five candidate sentences randomly sampled from the conversational corpus and five others from Twitter. The number of input sentences may be small; however, this cannot be easily increased due to the cost of labeling as we describe in the next section.

3.1.2 Reference sentences and evaluations

We recruited ten reference sentence writers who created references. Each writer created seven reference sentences for each input sentence under the two constraints: Character-length limitation and masked input sentences. As the limitation of character length N_c, one reference writer created three sentences under $N_c < 50$ condition, three sentences under $10 \geq N_c < 50$, and one sentence under $N_c < 10$.

[1] https://www.google.co.jp/trends/topcharts#date=2012

Method	$N_c < 50$	$10 \leq N_c < 50$	$N_c < 10$	Sum
Human (no mask)	18	18	6	42
Human (30% mask)	6	6	2	14
Human (60% mask)	6	6	2	14
IR-status	10	0	0	10
IR-response	10	0	0	10
AIML	10	0	0	10
Sum	60	30	10	100

Table 6.1. Statistics of gathered references for an input sentence. Human means the number of manually created references and the others are those of automatically created references.

The followings are the details of the masked input sentences. For each input sentence, six writers create references for the input sentence without masks, two writers create references for the 30% maked input sentences, and two writers for 60% masked ones. We randomly assigned the input sentences to the writers who imagined the masked terms and created the references. They wrote 70 references for each input sentence: 42 (6 writers $\times$ 7 sentences) sentences without masks, 14 (2 writers $\times$ 7 sentences) with 30% masked, and 14 (2 writers $\times$ 7 sentences) with 60% masked (Table 6.1).

In addition to the manually created references, we gathered 30 possibly negative reference sentences that were generated by the following two retrieval-based generation methods and one rule-based generation method: IR-status, IR-response and AIML [Ritter et al., 2011, Higashinaka et al., 2014]. IR-status retrieves reply posts whose associated source posts most closely resemble input user utterances. The IR-response approach is similar to IR-status, but it retrieves reply posts that most closely resemble input user utterances. AIML represents a rule-based conversational agent described in [Higashinaka et al., 2014]. This agent uses 149,300 rules (pattern-response pairs) written in AIML [Wallace, 2004] and retrieves responses whose associated patterns have the highest word-based cosine similarity to the input sentence. Each method generated ten reference sentences for each input sentence. Table 6.1 shows the statistics of the gathered reference sentences.

After the reference collection, two human evaluators annotated the winner of each

reference pair in terms of *naturalness as a response*. Since we have 100 references for each input sentence, they annotated 4,950 pairs for each input sentence.

Input sentence	References	Agent	Rates
I don't like Disneyland when it's very crowded... そして、ディズニーランドの大混雑も苦手です…。	It's so insane when everyone starts dashing at the same time as the gates open. 開門と同時に皆が走り出す光景って、何度見てもぞっとしますよね。	Human 0%	0.96
	Oh, it was really bad.. あらら、それは大変でしたね。	Human 30%	0.43
	I'm surprised that anyone would have such a pet. 飼っている人がいると聞いてびっくりです。	Human 60%	0.01
	Yeah, I agree! あーあたしも！	IR-status	0.81
	It's so crowded! 大混雑だな！	IR-response	0.29
	Yes, I go to Disneyland over ten times a year. はい、ディズニーには年に 10 回は行きます。	AIML	0.20
I just checked my iTunes, and I know all of my songs are from animated movies, games, vocaloids, voice actors and audio dramas of comics. iTunes に入ってるの確認したらアニソンとゲーソンとボカロと声優さんとドラマ CD だらけだった	You must really like anime and games! アニメとかゲームが好きなんだね！	Human 0%	0.95
	Don't you listen to rocks or western music? ロックや洋楽は聞かないんですか？	Human 30%	0.88
	Whet is your favorite year of it? 一番あたりだった年はいつですか？	Human 60%	0.15
	That's normal for myself. いつものわたしである	IR-status	0.26
	Vocaloids and anime songs lol. ボカロとアニソンだねｗ	IR-response	0.43
	What anime songs do you like? アニソンは何が好きなんですか？	AIML	0.43

Table 6.2. Examples of input sentences, reference sentences and their winning rates.

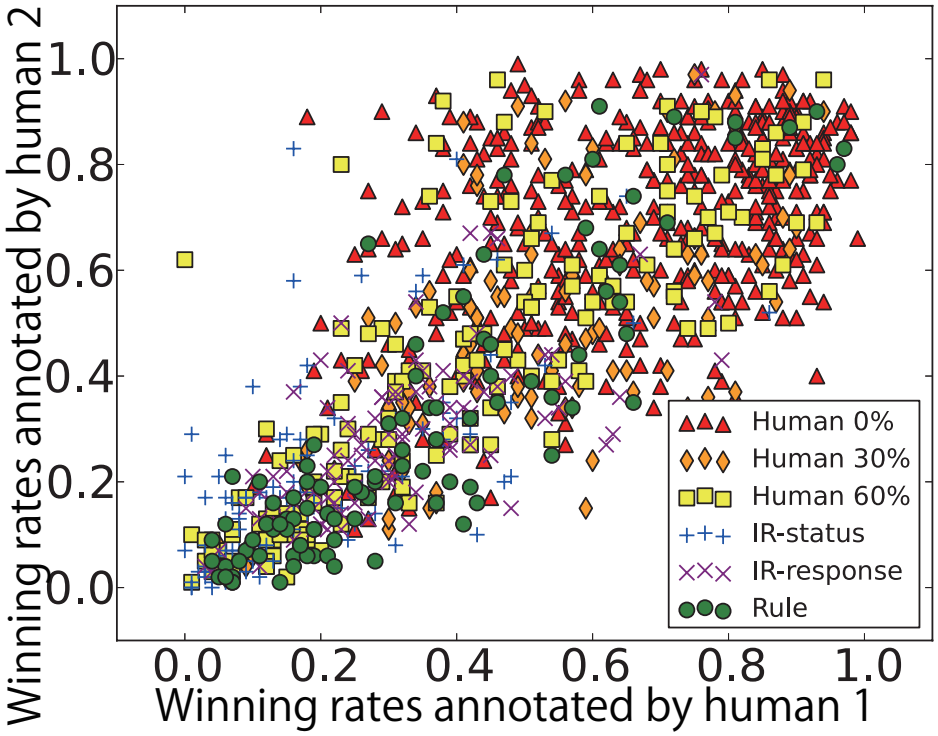


Figure 6.1. Distribution of annotated winning rates between annotators.

3.1.3 Estimation procedure

We compared the three methods described in Section 2.3 and smoothed BLEU that calculates a sentence's score over multi-references (m-BLEU) [2] and ΔBLEU [Galley et al., 2015] through the leave-one-out method; i. e., the methods estimate an evaluation score for each reference sentence using the other 99 references. Parameters of the methods are experimentally determined: We used 3 for N of AM and WS, SVR with RBF-kernel and $C = 5$. The similarity metrics used in AM, WS and Regression are either sentence-BLEU (BLEU), RIBES or Word Error Rate (WER). Here, the WER, which is calculated as the normalized Levenshtein distance NL to a reference sentence, is converted to a similarity with either $WER = 1 - NL$ (ranges from 0 to 1) or $WER = 1 - 2NL$ (ranges from -1 to 1).

[2]Here, we used NIST geometric sequence smoothing that is implemented in nltk (method 3)http://www.nltk.org/_modules/nltk/translate/bleu_score.html

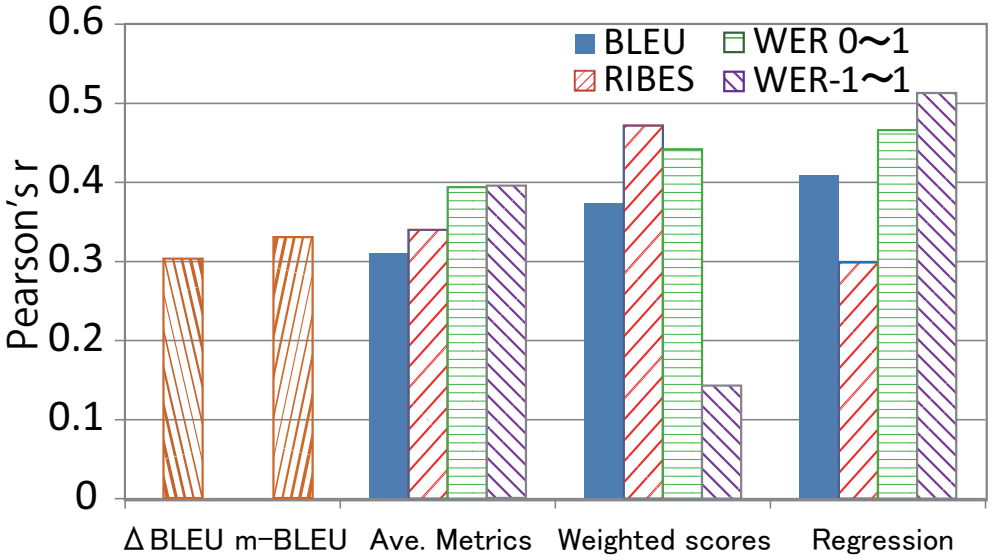


Figure 6.2. Correlation between annotated and estimated scores

3.2 Analysis of annotated evaluations

Before the experiments, we performed a brief analysis of the manually annotated scores. Figure 6.1 shows the distribution of the winning rates between the annotators. They are broadly distributed along the whole range of 0-1. Manually created references (red triangles, orange diamonds and yellow squares) were evaluated as more natural than the other agent-generated references. Comparing the agent-generated references, those generated from the retrieval-based methods (IR-status: Blue crosses, IR-responses: Purple x marks) are gathered in the low or middle winning rates, and those generated from AIML (green circles) are distributed along the whole range. This shows that AIML generates references with the same appropriateness as the manually created ones when the rules correctly match input sentences.

The Pearson's correlation coefficient between the human evaluators was 0.783. Figure 6.1 shows that the references with low winning rates show stronger correlations. This result indicates that negative input-response pairs are consistent between the evaluators, but the positive pairs are somewhat different.

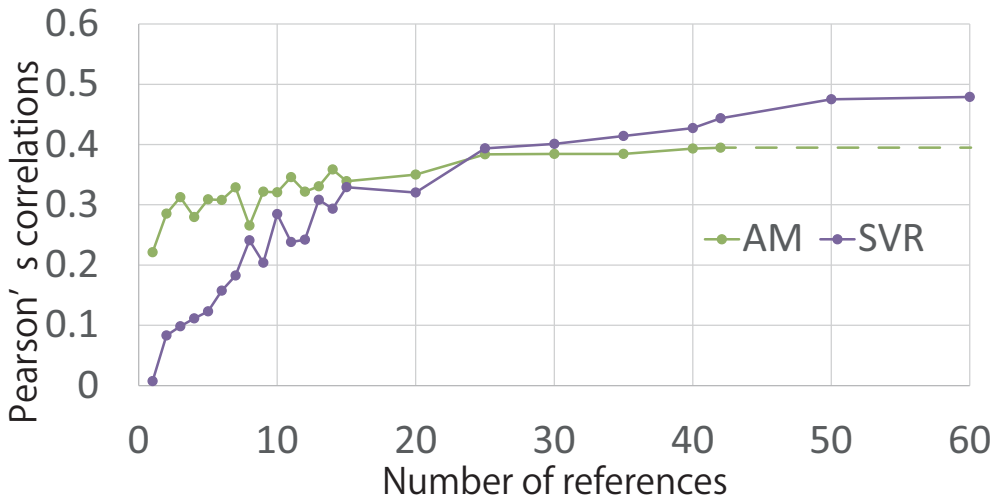


Figure 6.3. Correlations over number of references

3.3 Results

Figure 6.2 shows the correlation coefficients of the combination of the sentence similarity metrics and the proposed methods. SVR with WER (using the range from -1 to 1) shows the highest correlation ($r = 0.514$). Among the AM methods that leverage only the positive references, WER (-1 to 1) shows the highest correlation but still has lower correlations ($r = 0.399$) than those that leveraged the negative references.

Figure 6.3 shows the relations between the number of references and the correlations of SVR with WER and AM with WER. With fewer references, AM shows higher correlations than SVR, because it requires some training samples for accurate estimations, while AM outputs reasonable estimations even with just one reference. The performance of SVR becomes higher than AM with over 25 references and continues to improve. This indicates that we should use the average of metrics if we have only a few references and use a regression-based method if we have many references.

Figure 6.4 shows the agent-wise evaluation scores between manual annotation and SVR estimation. Each point is calculated as the means of ten scores; each score is sampled from estimated scores of an input-references pair whose references are associated with certain generation types (e.g., human 30% mask or AIML). The scores are highly correlated with Pearson's $r = 0.772$. Figure 6.4 illustrates that the references generated from human 60% mask, AIML, and IR-status are estimated with higher scores

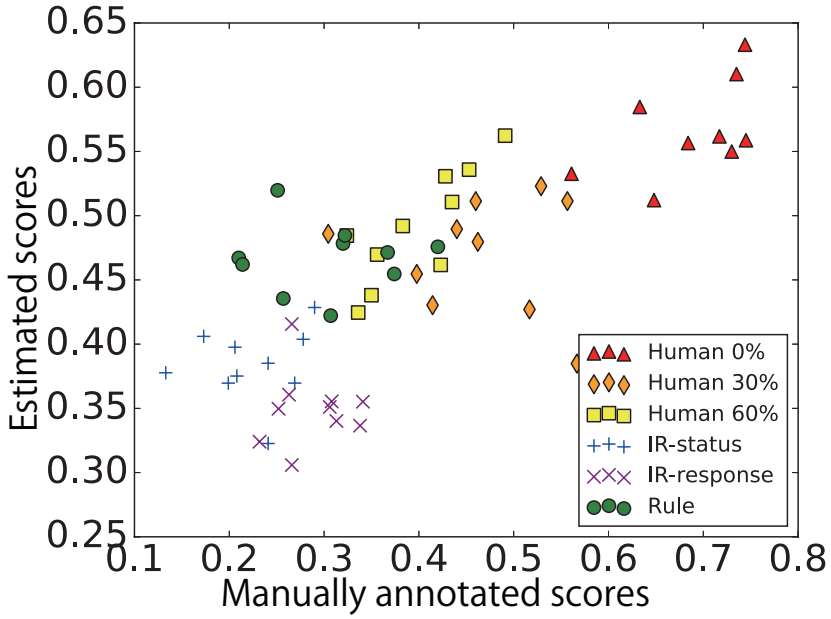


Figure 6.4. Annotated scores vs. estimated scores. The Pearson's $r = 0.772$.

than the manual scores. This is because that most low-evaluated references of human 60% mask and AIML have wrong contents but correct grammar, which are hardly distinguished with WER that only considers edit counts. IR-status has many expressions that did not appear in other references such as *lol* (*www* in Japanese) and emoticons like *:-)*. The differences between these expressions and the references are difficult to evaluate with WER and BLEU because they depend on word matching. This problem can be solved using character N-gram and the proportions of the character types as regression features.

4. Conclusion

We proposed a regression-based evaluation method for conversational dialogue agents. The sentence-wise correlation coefficient between our proposed and human annotated scores reached 0.514 and the agent-wise correlations were 0.772. Future works will compare our method through a dialogue-wise evaluation, increase our input-references pairs, and introduce other features for regression like character n-grams. Besides, we try to utilize all the references for all input sentences to increase negative references,

which possibly improve the estimation performance.

Chapter 7

Conclusion

1. Summary of this study

This study addresses the development of conversational dialogue agents that aim to make rapport with users. The problems arise from *considerable variations* of available dialogue-acts, topics, personality questions. With these variations, it is difficult to generate agent utterances even if we focus on the generation of one-turn responses. These variations also make it difficult to evaluate developed agents. In this thesis, we focus on the one-turn response generation to suppress the complexity of information that agents are required to deal with. Under this limitation, we categorized user utterances in open-domain conversation in terms of roughly grained dialogue-acts, and made directions to generate responses for each type of utterances. To realize the directions, we focus on the four key problems and proposed approaches to solve them, which leverage large-scale corpora that are respectively designed according to the characteristics of the problems.

First of all, in conversation that does not have obvious goals, it is difficult to define the appropriateness of conversational agents' actions. Besides, since dialogues between humans are not equally adequate, mimicking human dialogues is not enough to estimate appropriate agent actions; however, people can evaluate appropriateness of the dialogues. To estimate the appropriateness of the agent actions automatically from dialogue sequences and their manually annotated evaluations, we propose preference-learning based inverse reinforcement learning (PIRL) that estimates a reward function of reinforcement learning. This is important expansion for previous IRL to leverage

unequally adequate sequences for the estimation of a reward function. We examine the advantages of PIRL through comparisons between competitive RL and IRL based algorithms and our experiments show that our PIRL outperforms the other algorithms.

Second, we proposed an open-domain utterance generation method for conversational agents, whose requirements are derived from Grice's Maxims. The primary problem is the wide variety of topics of user utterances that cannot be covered by hand. Our proposed method automatically retrieves new topics (phrase-pairs that have dependency relations) relevant to the topics in user utterances. The relations between topics are defined as the number of dependency relations between the topics in a large-scale corpus. Our method creates agent utterances with combining both of the topics extracted from the user utterance and newly retrieved using dependency relations. Our experiment shows that the proposed method generates more appropriate utterances, with which the users can easily continue to talking, than the other conventional retrieved- and rule-based methods.

Third, we proposed a question-answering system for questions that ask conversational agent's specific personalities. We adopt QA methods based on manually created question-answer pairs; however, this approach generally lacks the coverage of user utterances. To improve the coverage, we developed Person DataBase (PDB), which consists of 26595 personality questions gathered from many questioners that are categorized by hand into 10082 question categories. First we investigated the distributions of frequently asked questions. The distribution of the questions are so long-tailed, so our PDB covers 42% of questions that appear in conversation between humans. Our analysis also shows that 20 questioners are cost-effective numbers of our approach. Our objective evaluation experiment showed that our personality QA system estimated appropriate question categories with 84.1% accuracy for question sentences in PDB and 60.7% for a conversation corpus. Furthermore, we examined the effectiveness of our personality QA system through subjective evaluation experiments. Our result indicated that it is useful to prevent the breakdown of conversation.

Finally, we proposed an automatic evaluation method for conversational agents based on large-scale reference sentences. Previous reference-based automatic evaluation methods, which adopt generative approach that returns sentence similarities themselves as evaluation scores, have two difficulties to estimate the ratings of agent utterances: One is insufficient references to cover the wide range of agent utterances,

and another is the naive handling of negative references, with which inappropriate utterances that are far from any references are evaluated as medium scores. To intensively gather references around positive utterances, we gathered manually created 70 reference sentence under the character-length limitations and the masking of input sentences constraints in order to collect inappropriate references. To enlarge negative references, we added automatically retrieved 30 reference sentences to the references. Two human evaluators annotated evaluations for each pair of the references. Our regression-based method estimates scores (pairwise winning rates) of response sentences generated by agents using Support Vector Regression, which utilizes similarities between target sentences and reference sentences as estimation features. The sentence-wise correlation coefficient between the scores, which proposed method and human annotator rated, reached 0.514, and the agent-wise were 0.772. These scores shows that our method has a potential to be substitute for manual evaluations.

With these improvements, our approaches generate appropriate responses for each type of user utterances in conversation when we focus on the generation of one-turn responses. We believe that this thesis gives a clue for the further development of conversational agents, which model dialogue contexts and generate consistent responses for open-domain user utterances.

2. Remaining problems and future directions

In this thesis, we discus about the improvements of the literal qualities of one-turn responses. Although it is necessary to consider multimodal factors such as natural turn-taking and entailment between talkers to develop conversational agents that make rapport with users, much work still remains to improve the literal qualities. For example, since our study focuses only on the generation of one-turn responses, much work still remains to treat multiple turns of utterances. Besides, we think that the appropriateness of one-turn responses can be improved. Here we discuss the following three problems to realize conversational agents.

First of all, even with the improvement of utterance generation methods, there remains a risk of generating syntactically and semantically inappropriate utterances. Users will be disappointed with such inappropriate utterances, which denote that the agents do not understand the user utterances. To suppress such utterances,

filtering approach of inappropriate utterances has recently gathered attentions [Higashinaka et al., 2016]. Now we have so many utterances generated by several utterance generation methods, some utterances will be retained after the filtering of potentially inappropriate utterances. As the need and performance of conversational agents increase, this filtering will become more important topic.

Second, the user state that our agent can model is limited to dialogue-acts and topic phrase-pairs. Although these are almost enough to generate one-turn responses, of course it is not enough to capture the sequence of user utterances. Memorizing the other talker's personalities is a fundamental requirement to build relationship with each other; thus, this function is necessary to develop conversational agents that make relationship with users. We assume a key for the user state modeling is personality information. We already have a huge database about personalities, which enables us to store the user's personality information like slots in task-oriented dialogues. If we model a user state with such slots, it is possible to use the same approach as dialogue control to control the dialogue topics and generate utterances. Besides, we also assume that the personality information can be separated into two types of information: Contents and modalities. If we define contents using predicate-argument structures that represents events and modalities using dialogue-acts of self-disclosure, we can treat the whole space of user's personality information. For example, Hirano et al. proposed a user modeling method that stores the user's preferences or experiences with such approach on the basis of our analysis of the Person DataBase [Hirano et al., 2015]. When we model user states with this approach, we can generate questions about the details of the user's personalities, with which users feel that the agent attentively listen user's talk.

Third, our conversational agent lacks the direction of dialogues, and sometimes generate utterances such that users cannot understand the reason why the agent said. Current utterance generations are based on question-answer style, so it cannot capture and deal with the topic flows of several dialogue turns. This drawback makes it difficult to build relationship between agents and users. One idea to make dialogue directions is to consider larger size of dialogue units like scenarios with a few utterances, instead of *center* words or phrases. We assume that each topic of conversation has a small story that users would like to share with agents, and conversation consist of a flow of topics. Storytelling or narrative generation methods that retrieve many sto-

ries from weblogs seem promising way to develop scenarios with wide range of topics [Swanson and Gordon, 2012]. The transition of the topic flows will be realized with RL-based approaches in the similar way as dialogue control. Another direction is to build agents on the basis of more precise user states [Zhao et al., 2014]. Belief, Desire and Intention (BDI) model [Cohen and Levesque, 1990] is also an interesting model, with which agents can be designed to have meta-level of objectives such as *knowing each other* or *building confidence*.

Acknowledgements

First of all, I would like to thank to Professor Satoshi Nakamura and faculty members of AHC laboratory for their careful reviews and fruitful discussions on my research. Thanks to Professor Nakamura, who gives me an admission to join his laboratory. He also gave me many comments in daily discussions and for my thesis. I appreciate Assistant Professor Koichiro Yoshino, who gave me a fundamental directions to improve the structures of my thesis. He also supported much of paperwork related to my Ph.D. course. From Assistant Professor Sakriani Sakti, I received the detailed comments and directions to make my thesis much clear. I thank her of careful and kind comments.

I would also thank to Professor Yuji Matsumoto and Associate Professor Masahiro Araki for agreeing to join the member of my doctoral committee and giving insightful comments on my thesis. I thank to Professor Yuji Matsumoto, who pointed out ambiguous writings of my thesis, especially from technical viewpoints. I gratitude Associate Professor Msahiro Araki for his fruitful discussions about problems that will arize in practical use.

All the research in my thesis are performed at NTT Communication Laboratories. Thanks to Professor Yasuhiro Minami, who helped me to start the research career of first a few years. I wish to thank Dr. Ryuichiro Higashinaka for his kind supervision. Most of my work are proceeded with him. I also thank to Ms. Toyomi Meguro for daily discussion about research, and childcare recently. I appreciate Professor Junji Yamato, who was the director of my research group, discussed our research directions and helped my office work.

Dr. Toru Hirano, Ms. Chiaki Miyazaki and Mr. Atsushi Otsuka, the members of dialogue sub-group in the NTT Media Intelligence Laboratory, gave me much discussion on my research. I thank them for the discussion and their suggestions.

References

[Abbeel and Ng, 2004] Pieter Abbeel and Andrew Y. Ng (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of 21st international conference on Machine learning*, pages 1–8.

[Batacharia et al., 1999] Bobby Batacharia, David Levy, Roberta Catizone, Alex Krotov, and Yorick Wilks (1999). CONVERSE: a conversational companion. *Machine conversations*, pages 205–215.

[Bickmore and Cassell, 2000] Timothy Bickmore and Justine Cassell (2000). "How about this weather?": Social Dialogue with Embodied Conversational Agents. In *Proceedings of the 2000 Fall Symposium, Socially Intelligent Agents: The Human in the Loop*, pages 4–8.

[Bickmore and Cassell, 2001] Timothy Bickmore and Justine Cassell (2001). Relational Agents: A Model and Implementation of Building User Trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 396–403.

[Bickmore et al., 2005] Timothy W. Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren (2005). 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers*, 17(6):711–735.

[Boularias et al., 2010] Abdeslam Boularias, Hamid R. Chinaei, and Brahim Chaibdraa (2010). Learning the Reward Model of Dialogue POMDPs from Data. In *Proceedings of NIPS 2010 Workshop on Machine Learning for Assistive Technologies*, pages 1–9.

[Carberry et al., 1999] Sandra Carberry, Jennifer Chu-Carroll, and Stephanie Elzer (1999). Constructing and Utilizing a Model of User Preferences in Collaborative Consultation Dialogues. *Computational Intelligence*, 15(3):185–217.

[Caspi et al., 2005] Avshalom Caspi, Brent W. Roberts, and Rebecca L. Shiner (2005). Personality Development: Stability and Change. *Annual Review of Psychology*, 56(1):453–484.

[Cassell, 2000] Justine Cassell (2000). *Embodied conversational agents*. MIT Press.

[Chandramohan et al., 2011] Senthilkumar Chandramohan, Matthieu Geist, Fabrice Lefevre, and Olivier Pietquin (2011). User Simulation in Dialogue Systems using Inverse Reinforcement Learning. In *Proceedings of International Conference of the International Speech Communication Association*.

[Cheepen, 1988] Christine Cheepen (1988). *The predictability of informal conversation*. London: Pinter.

[Cheng et al., 2011] Weiwei Cheng, Johannes Fürnkranz, Eyke Hüllermeier, S.H. Sang Hyeun Park, and F Johannes (2011). Preference-based policy iteration: leveraging preference learning for reinforcement learning. In *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 312–327.

[Cohen and Levesque, 1990] Philip R. Cohen and Hector J. Levesque (1990). Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261.

[DeVault et al., 2014] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-philippe Morency (2014). SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support. In *Proceedings of International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1061–1068.

[Eggins and Slade, 1997] Suzanne Eggins and Diana Slade (1997). *Analysing casual conversation*. London: Cassell publishing.

[Engelbrech and Hartard, 2009] Kp Engelbrech and Felix Hartard (2009). Modeling user satisfaction with hidden markov model. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pages 170–177.

[Freire da Silva et al., 2006] Valdinei Freire da Silva, Anna Helena Reali Costa, and Pedro Lima (2006). Inverse Reinforcement Learning with Evaluation. In *Proceedings of International Conference on Robotics and Automation*, pages 4246–4251.

[Fuchi and Takagi, 1998] Takeshi Fuchi and Shinichiro Takagi (1998). Japanese morphological analyzer using word co-occurrence: JTAG. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 409–413.

[Galley et al., 2015] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan (2015). Delta-BLEU : A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. In *Proceedings of the Association for Computational Linguistics*, pages 445–450.

[Grefenstette, 1988] John J. Grefenstette (1988). Credit assignment in rule discovery systems based on genetic algorithms. *Machine Learning*, 3(2):225–245.

[Grice, 1975] Herbert Paul Grice (1975). Logic and Conversation. In *Syntax and semantics. 3: Speech acts*, pages 41–58.

[Hasegawa et al., 2013] Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda (2013). Predicting and Eliciting Addressee's Emotion in Online Dialogue. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 964–972.

[Herbrich et al., 2000] Ralf Herbrich, Thore Graepel, and Klaus Obermayer (2000). Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, volume 88, pages 115–132. MIT Press.

[Higashinaka et al., 2016] Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi., and Michimasa Inaba (2016). The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference*, pages 3146–3150.

[Higashinaka et al., 2014] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo (2014). Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 928–939.

[Higashinaka et al., 2010] Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro (2010). Issues in predicting user satisfaction transitions in dialogues: Individual differences, evaluation criteria, and prediction models. In *Proceedings of International Workshop on Spoken Dialogue Systems*, pages 48–60.

[Hirano et al., 2015] Toru Hirano, Nozomi Kobayashi, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo (2015). User Information Extraction for Personalized Dialogue Systems. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 67–76.

[Huang et al., 2011] Lixing Huang, Louis Philippe Morency, and Jonathan Gratch (2011). Virtual rapport 2.0. In *Proceedings of International Workshop on Intelligent Virtual Agents*, pages 68–79.

[Imamura et al., 2014] Kenji Imamura, Ryuichiro Higashinaka, and Tomoko Izumi (2014). Predicate-Argument Structure Analysis with Zero-Anaphora Resolution for Dialogue Systems. In *Proceedings of the International Conference on Computational Linguistics*, pages 806–815.

[Imamura et al., 2007] Kenji Imamura, Genichiro Kikui, and Norihito Yasuda (2007). Japanese Dependency Parsing Using Sequential Labeling for Semi-Spoken Language. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 225–228.

[John and Srivastava, 1999] Oliver P. John and Sanjay Srivastava (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research*, volume 2, pages 102–138.

[Kobayashi et al., 2005] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima (2005). Collecting evaluative expressions for

opinion extraction. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 596–605.

[Kolomiyets and Moens, 2011] Oleksandr Kolomiyets and Marie-Francine Moens (2011). A Survey on Question Answering Technology from an Information Retrieval Perspective. *Information Sciences*, 181(24):5412–5434.

[Laver, 1975] John Laver (1975). Communicative functions of phatic communion. In *Organization of behavior in face-to-face interaction*, pages 215–238.

[Lee et al., 2006] Cheongjae Lee, Sangkeun Jung, Jihyun Eun, Minwoo Jeong, and Gary Geunbae Lee (2006). A Situation-Based Dialogue Management using Dialogue Examples. In *Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing*, pages 69–72.

[Leuski and Traum, 2011] Anton Leuski and David Traum (2011). NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.

[Lin and Och, 2004] Chin-Yew Lin and Franz Josef Och (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 605–612.

[Liu and Nocedal, 1989] Dong C. Liu and Jorge Nocedal (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45:503–528.

[Mairesse and Walker, 2007] Francois Mairesse and Marilyn Walker (2007). PERSONAGE: Personality generation for dialogue. In *Proceedings of the Annual Meeting of the Association For Computational Linguistics*, pages 496–503.

[Meguro et al., 2010] Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka (2010). Controlling Listening-oriented Dialogue using Partially Observable Markov Decision Processes. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 761–769.

[Meguro et al., 2011] Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka (2011). Evaluation of listening-oriented dialogue control rules based on the analysis of HMMs. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 809–812.

[Mikolov et al., 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 1–9.

[Misu et al., 2012] Teruhisa Misu, Kallirroi Georgila, Anton Leuski, and David Traum (2012). Reinforcement Learning of Question-Answering Dialogue Policies for Virtual Museum Guides. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 84–93.

[Misu et al., 2011] Teruhisa Misu, Komei Sugiura, Tatsuya Kawahara, Kiyonori Ohtake, Chiori Hori, Hideki Kashioka, Hisashi Kawai, and Satoshi Nakamura (2011). Modeling spoken decision support dialogue and optimization of its dialogue strategy. *ACM Transactions on Speech and Language Processing*, 7(3):10:1–10:18.

[Nagata et al., 2006] Masaaki Nagata, Kuniko Saito, and Yoshihiro Matsuo (2006). Japanese natural language retrieval system: Web Answers. In *Proceedings of the Annual Meeting of the Association of Natural Language Processing*, pages B2–2.

[Nakano et al., 2000] Mikio Nakano, Noboru Miyazaki, Norihito Yasuda, Akira Sugiyama, Jun-ichi Hirasawa, Kohji Dohsaka, and Kiyoaki Aikawa (2000). WIT: A toolkit for building robust and real-time spoken dialogue systems. In *Proceedings of the 1st SIGdial workshop on Discourse and dialogue*, pages 150–159.

[Ng and Russell, 2000] Andrew Y. Ng and Stuart Russell (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 663–670.

[Nio et al., 2014] Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura (2014). Improving the robustness of example-based dialog retrieval using recursive neural network paraphrase identification. In *Proceedings of Spoken Language Technology Workshop*, pages 306–311.

[Nisimura et al., 2003] Ryuichi Nisimura, Yohei Nishihara, Ryosuke Tsurumi, Akinobu Lee, Hiroshi Saruwatari, and Kiyohiro Shikano (2003). Takemaru-kun : Speech-oriented Information System for Real World Research Platform. In *Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction*, pages 70–78.

[Oh and Rudnicky, 2000] Alice H. Oh and Alexander I. Rudnicky (2000). Stochastic language generation for spoken dialogue systems. In *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems*, volume 3, pages 27–31.

[Papineni et al., 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics*, pages 311–318.

[Pérez-Agüera et al., 2010] José R. Pérez-Agüera, Javier Arroyo, Jane Greenberg, Joaquin Perez Iglesias, and Victor Fresno (2010). Using BM25F for semantic search. In *Proceedings of the 3rd International Semantic Search Workshop*, pages 1–8.

[Ritter et al., 2010] Alan Ritter, Colin Cherry, and Bill Dolan (2010). Unsupervised modeling of Twitter conversations. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

[Ritter et al., 2011] Alan Ritter, Colin Cherry, and William B. Dolan (2011). Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593.

[Robinson et al., 2008] Susan Robinson, David Traum, Mdhun Ittycheriah, and Joe Henderer (2008). What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In *Proceedings of the anguage Resources and Evaluation Conference*, pages 1125–1131.

[Schmitt et al., 2011] Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker (2011). Modeling and predicting quality in spoken human-computer interaction.

Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue, pages 173–184.

[Scott Thornbury, 2006] Diana Slade Scott Thornbury (2006). *Conversation: From Description to Pedagogy*. Cambridge University Press.

[Sculley, 2009] D Sculley (2009). Large Scale Learning to Rank. In *Proceedigs of the NIPS 2009 Workshop on Advances in Ranking*, pages 1–6.

[Sekine et al., 2002] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata (2002). Extended Named Entity Hierarchy. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1818–1824.

[Shibata et al., 2009] Masahiro Shibata, Tomomi Nishiguchi, and Yoichi Tomiura (2009). Dialog System for Open-Ended Conversation Using Web Documents. *Informatica*, 33:277–284.

[Singh et al., 1999] Satinder Singh, Michael Kearns, Diane Litman, and Marilyn Walker (1999). Reinforcement learning for spoken dialogue systems. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

[Smola and Schölkopf, 2004] Alex J. Smola and Bernhard Schölkopf (2004). A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222.

[Stolcke et al., 2000] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

[Sugiyama et al., 2013] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami (2013). Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures. In *Proceedings of the 14th annual SIGdial Meeting on Discourse and Dialogue*, pages 334–338.

[Sugiyama et al., 2014] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami (2014). Open-domain Utterance Generation using Phrase

Pairs based on Dependency Relations. *Proceedings of the 2014 IEEE Workshop on Spoken Language Technology*, pages 60–65.

[Sutskever et al., 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of the Advances in neural information processing systems*, pages 3104–3112.

[Swanson and Gordon, 2012] Reid Swanson and Andrew S. Gordon (2012). Say Anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling. *ACM Transactions on Interactive Intelligent Systems*, 2(3):1–35.

[Takeuchi et al., 2007] Shota Takeuchi, Tobias Cincarek, Hiromichi Kawanami, Hiroshi Saruwatari, and Kiyohiro Shikano (2007). Construction and optimization of a question and answer database for a real-environment speech-oriented guidance system. In *Proceedings of the International Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques*, pages 149–154.

[Tidwell and Walther, 2002] Lisa C. Tidwell and Joseph B. Walther (2002). Computer-Mediated Communication Effects on Disclosure, Impressions, and Interpersonal Evaluations. *Human Communication Research*, 28(3):317–348.

[Traum et al., 2015] David Traum, Kallirroi Georgila, Ron Artstein, and Anton Leuski (2015). Evaluating Spoken Dialogue Processing for Time-Offset Interaction. In *Proceedings of the 16th Annual SIGdial Meeting on Discourse and Dialogue*, pages 199–208.

[Turian et al., 2010] Joseph Turian, Lev Ratinov, and Yoshua Bengio (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linquistics*, number July, pages 384–394.

[Vardoulakis et al., 2012] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L. Sidner, and Timothy Bickmore (2012). Designing relational agents as long term social companions for older adults. In *Proceedings of the International Conference on Intelligent Virtual Agents*, pages 289–302.

[Walker et al., 1998] Marilyn A Walker, Aravind Aravind Krishna Joshi, and Ellen Ellen Friedman Prince (1998). *Centering theory in discourse*. Oxford University Press on Demand.

[Walker et al., 1997] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280.

[Wallace, 2004] Richard S. Wallace (2004). The Anatomy of A.L.I.C.E. *ALICE Artificial Intelligence Foundation, Inc.*

[Wang, 2006] Mengqiu Wang (2006). A survey of answer extraction techniques in factoid question answering. *Computational Linguistics*.

[Wärnestål, 2007] Pontus Wärnestål (2007). *Dialogue Behavior Management in Conversational Recommender Systems*. PhD thesis, Linköping University.

[Weizenbaum, 1966] Joseph Weizenbaum (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

[Williams, 2007] Jason D. Williams (2007). Applying POMDPs to dialog systems in the troubleshooting domain. In *Proceedings of the Workshop on Bridging the Gap*, pages 1–8.

[Williams and Young, 2005] Jason D. Williams and Steve Young (2005). The SACTI-1 Corpus: Guide for Research Users. Technical report.

[Williams and Young, 2007] Jason D. Williams and Steve Young (2007). Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

[Zhao et al., 2014] Ran Zhao, Alexandros Papangelis, and Justine Cassell (2014). Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*, pages 514–527.

[Ziebart et al., 2008] Brian D. Ziebart, Andrew Maas, Andrew Bagnell, and Anind K. Dey (2008). Maximum entropy inverse reinforcement learning. *Proceedings of the Twenty-Second Conference on Artificial Intelligence*, pages 1433–1438.

Publication list

Refereed Journals

[1] 杉山弘晃, 目黒豊美, 東中竜一郎. 対話システムのパーソナリティを問う質問の大規模な収集と分析. 人工知能学会論文誌　論文特集「知的対話システム」, Vol. 31, No. 1, pp. 1-9, 2015.

[2] 松崎拓也, 横野光, 宮尾祐介, 川添愛, 狩野芳伸, 加納隼人, 佐藤理史, 東中竜一郎, 杉山弘晃, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩. 「ロボットは東大に入れるか」プロジェクト：代ゼミセンター模試タスクにおけるエラーの分析. 自然言語処理　論文特集「エラー分析」, Vol. 23, No. 1, pp. 119-159, 2016.

[3] 藤田早苗, 小林哲生, 南泰浩, 杉山弘晃. 幼児を対象としたテキストの対象年齢推定方法, 認知科学, Vol. 22, No. 4, pp. 604-620, 2015.

[4] 杉山弘晃, 目黒豊美, 東中竜一郎, 南泰浩. 任意の話題を持つユーザ発話に対する係り受けと用例を利用した応答文の生成. 人工知能学会論文誌　論文特集「Web インテリジェンスとインタラクションの新展開」, Vol. 30, No. 1, pp. 183-194, 2015.

[5] 小林哲生, 南泰浩, 杉山弘晃. 「語彙爆発の新しい視点」のさらなる検証. ベビーサイエンス, Vol. 12, pp. 55-58, 2013

[6] 小林哲生, 南泰浩, 杉山弘晃. 語彙爆発の新しい視点：日本語学習児の初期語彙発達に関する縦断データ解析 ベビーサイエンス, Vol. 12, pp. 34-49, 2013.

[7] 杉山弘晃, 南泰浩. 情報提示対話を主導するシステムのためのユーザの潜在

的情報要求意図の推定. 電子情報通信学会論文誌 A（人とエージェントのインタラクション特集号）, J95-A, No.1, pp. 74-84, 2012.

Refereed Conferences

[8] Hiroaki Sugiyama. Utterance Selection based on Sentence Similarities and Dialogue Breakdown Detection on NTCIR-12 STC Task. NII Testbeds and Community for Information access Research Conference (NTCIR), 2016.

[9] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka. Evaluation of Responses for Questions about Conversational Agent's Personality. International Workshop on Spoken Dialogue Systems (IWSDS), 2015.

[10] Ryuichiro Higashinaka, Toyomi Meguro, Hiroaki Sugiyama, Toshiro Makino, Yoshihiro Matsuo. On the difficulty of improving hand-crafted rules in chat-oriented dialogue systems. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1014-1018, 2015.

[11] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami. Open-domain Utterance Generation using Phrase Pairs based on Dependency Relations. Spoken Language Technologies, 2014.

[12] Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, Yoshihiro Matsuo. Evaluating Coherence in Open Domain Conversational Systems. Interspeech, 2014.

[13] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, Yoshihiro Matsuo. Towards a fully-NLP based open domain conversational system. Coling, 2014.

[14] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami. Large-scale Collection and Analysis of Personal Question-answer Pairs for Conversational Agents. Intelligent Virtual Agent, 2014.

[15] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami. Open-domain Utterance Generation for Conversational Dialogue Systems using Web-scale Dependency Structures. The 14th annual SIGdial Meeting on Discourse and Dialogue, pp. 334-338, 2013.

[16] Hiroaki Sugiyama, Tessei Kobayashi, Yasuhiro Minami. Individual variation of word acquisition age: a comparison of Japanese- and English-speaking infants. Workshop on Infant Language Development, 2013.

[17] Yasuhiro Minami, Tessei Kobayashi, Hiroaki Sugiyama. Cross-linguistic universality of word acquisition ages in comprehension and production. Workshop on Infant Language Development, 2013.

[18] Tessei Kobayashi, Yasuhiro Minami, Hiroaki Sugiyama. Vocabulary spurt and noun acquisition: Empirical evidence from longitudinal data in Japanese-speaking children. Child Language Seminar, 2013.

[19] Tessei Kobayashi, Yasuhiro Minami and Hiroaki Sugiyama. Word-class composition in first 20 words predicts later word acquisition rate. SRCD, 2013.

[20] Hiroaki Sugiyama, Toyomi Meguro and Yasuhiro Minami. Preference-learning based Inverse Reinforcement Learning for Dialog Control. Interspeech, 2012

[21] Tessei Kobayashi, Yasuhiro Minami and Hiroaki Sugiyama. Vocabulary spurt and word-class composition: Further evidence for a model of plateaus and linearity in early vocabulary growth. AMLaP, 2012.

[22] Hiroaki Sugiyama, Tessei Kobayashi and Yasuhiro Minami. Prediction of Vocabulary Growth Using Locally Linearity. ISSBD, 2012

[23] Yasuhiro Minami, Tessei Kobayashi and Hiroaki Sugiyama. Plateaus and linearity of early vocabulary growth. ISSBD, 2012

[24] Yasuhiro Minami, Hiroaki Sugiyama and Tessei Kobayashi. Multiple vocabulary spurts in Japanese children. IASCL, 2011

[25] Hiroaki Sugiyama, Tessei Kobayashi and Yasuhiro Minami. Analysis of Vocabulary Spurt From Prediction Performance Evaluation SRCD, 1-088-190, 2011

[26] Hiroaki Sugiyama and Yasuhiro Minami. Information Provision-timing Control for Informational Assistance Robot, HRI, pp259-260, 2011

Unrefereed Conferences

[27] 堯天貴之, 植田佳文, 東中竜一郎, 杉山弘晃, 平博順. 述語項構造解析を用いた英語長文読解問題の自動解法. 言語処理学会年次大会，2016.

[28] 北条伸克，井島勇祐，杉山弘晃. 対話行為情報を表現可能な音声合成の検討 人工知能学会全国大会，2016.

[29] 杉山弘晃. 異なる特性を持つデータの組み合わせによる雑談対話の破綻検出. 第 75 回 人工知能学会 言語・音声理解と対話 処理研究会（SIG-SLUD）, pp. 51-56, 2015.

[30] 太田昌克, 松田昌史, 小林哲生, 奥村優子, 杉山弘晃. Bluetooth ビーコンを用いたポスター会場における来場者の移動軌跡取得 〜HCS2015 年 1 月研究会における実証実験〜. HCS 研究会, 2015.

[31] 東中竜一郎, 杉山弘晃, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩. センター試験における英語問題の回答手法. 言語処理学会第 21 回年次大会, 2015.

[32] 松崎拓也, 横野光, 宮尾祐介, 川添愛, 狩野芳伸, 加納隼人, 佐藤理史, 東中竜一郎, 杉山弘晃, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩. 『ロボットは東大に入れるか』プロジェクト 代ゼミセンター模試タスクにおけるエラーの分析. 言語処理学会第 21 回年次大会 エラー分析ワークショップ, 2015.

[33] 杉山弘晃, 目黒豊美, 東中竜一郎. 雑談対話中の発話文に対する多面的評価の分析. 人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD 72), pp. 31-36, 2014.

[34] 杉山弘晃, 目黒豊美, 東中竜一郎. 大規模マルチリファレンスに基づく雑談対話システムの自動評価に向けた実験的検討. 人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD-B401), pp. 1-6, 2014. （2014 年度人工知能学会研究会優秀賞）

[35] 目黒豊美, 杉山弘晃, 東中竜一郎, 南泰浩. ルールベース発話生成と統計的発話生成の融合に基づく対話システムの構築. 人工知能学会全国大会, 2M5-OS-20b-2, 2014.

[36] 杉山弘晃, 目黒豊美, 東中竜一郎, 南泰浩. 対話システムのパーソナリティを問う質問に対する応答生成. 人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD)，2014.

[37] 杉山弘晃, 目黒豊美, 東中竜一郎, 南泰浩. 任意の話題を持つユーザ発話に対する係り受けを利用した応答文の生成. 人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD)，2013.

[38] 目黒豊美, 東中竜一郎, 杉山弘晃, 南泰浩. 意味属性パターンを用いたマイクロブログ中の発言に対する自動対話行為付与. 研究報告音声言語情報処理 (SLP), Vol. 1, pp. 1-6, 2013.

[39] 杉山弘晃, 小林哲生, 南泰浩. 語彙の身体性が獲得時期の個人差に与える影響. 日本赤ちゃん学会第 13 回学術集会，2013

[40] 南泰浩, 小林哲生, 杉山弘晃語の学習では本当に幼児は名詞を早く獲得する？—語の理解・発話日齢の推定による名詞優位性の言語間比較—. 日本赤ちゃん学会第 13 回学術集会，2013

[41] 小林哲生, 南泰浩, 杉山弘晃. 初期語彙発達の急増期における語彙カテゴリー構成の特徴. 思考と言語研究会，2012

[42] 南泰浩, 小林哲生, 杉山弘晃. 幼児早期出現語の理解-発話指標による幼児語彙学習特徴の検証. 思考と言語研究会, 2012

[43] 南泰浩, 小林哲生, 杉山弘晃. カルマンフィルタを用いた語彙発達におけるプラトー時期の推定. 音響学会秋季大会, 2012

[44] 杉山弘晃, 小林哲生, 南泰浩. 語彙学習速度の線形性を利用した語彙学習日齢の予測. 日本赤ちゃん学会　第 12 回学術講演会, 2012

[45] 南泰浩, 小林哲生, 杉山弘晃. 線形関数とプラトー割り込みによる語彙発達モデルの検証—幼児の語彙発達におけるポアソン過程性の検証—. 日本赤ちゃん学会　第 12 回学術講演会, 2012

[46] 小林哲生, 南泰浩, 杉山弘晃. 幼児の語彙学習速度と語彙カテゴリー構成 (優秀発表賞). 日本赤ちゃん学会　第 12 回学術講演会, 2012

[47] 杉山弘晃, 目黒豊美, 南泰浩. 順序学習に基づく逆強化学習による対話制御. 第 26 回人工知能学会全国大会, 2012.

[48] 南泰浩, 小林哲生, 杉山弘晃. 線形関数とプラトー割込による幼児語彙発達のモデル化（優秀賞）. 言語処理学会第 18 回年次大会, 2012

[49] 南泰浩, 小林哲生, 杉山弘晃. 折れ線近似による語彙爆発開始時期の推定. 電子情報通信学会音声研究会 (SP), 2012

[50] 木村昭悟, 南泰浩, 坂野鋭, 前田英作, 杉山弘晃. 対話型映像認識理解における動的学習戦略に関する試み. 電子情報通信学会技術報告, PRMU2010-135, 2010

[51] 杉山弘晃, 南泰浩. ユーザ支援システムのための人の行動タイミング決定方策の分析. 日本ロボット学会学術講演会, 2010