NAIST-IS-DD1461010

Doctoral Dissertation

Adaptive conversational agent considering user preferences

Masahiro Mizukami

March 16, 2017

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

Masahiro Mizukami

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Assistant Professor Koichiro Yoshino	(Co-supervisor)
Assistant Professor Graham Neubig	(Carnegie Mellon University)
Assistant Professor Sakriani Sakti	(Co-supervisor)
Dr. Ryuichiro Higashinaka	(NTT Media Intelligence Laboratories)

Adaptive conversational agent considering user preferences*

Masahiro Mizukami

Abstract

Conversation is an important communication channel to build a social relationship between participants. Building a social relationship helps to make conversation smoother, more connected and comfortable. This effect is called rapport in social psychology. This thesis presents an adaptive conversational agent system considering user preference. Studies of establishing rapport showed an importance of the adaptivity to user preferences to achieve a satisfactory conversation. If user preferences are promptly extracted and adopted to the conversational system's responses, these system responses are able to evoke high engagement to continue the conversation in the long term, and high satisfaction to make the conversation more comfortable. User preference has a variety of speaking style, dialogue strategies, and communication distance. In order to utilize these user preferences, it is necessary to know how to extract and handle those preferences. In this thesis, we studied four cooperative approaches in example-based dialogue modeling to build a conversational agent system considering user preference. First, we proposed a linguistic individuality transformation method to transform the speaking style of conversational agent's responses. This method makes it possible for the conversational agent to talk with the preferred individuality. Second, we proposed a satisfaction prediction method for the example database that the conversational agent holds inside to achieve a conversation with higher satisfaction. This method enables selecting a response that increases the user satisfaction. Third, we proposed an adaptive response selection method considering user preferences

*Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1461010, March 16, 2017.

to select the best response to the specific user preference. This method enables for the agent to respond with the most satisfactory response to the user. Fourth, we proposed a response selection method based on an entrainment analysis. This method selects words given a dialogue act depending on the user's dialogue act. Entrainment is a conversational phenomenon in which dialogue participants synchronize to each other, which induces an increase of naturalness and engagement on a conversation. This response selection method based on an entrainment analysis realizes the conversational agent to synchronize appropriately with the user through a dialogue. We examine these proposed methods and confirmed effectiveness in evaluation experiments.

Keywords:

conversational agent, example based dialogue modeling, response selection, individuality modeling, entrainment

ユーザの選好を考慮した適応的対話エージェント*

水上 雅博

内容梗概

雑談は話者同士の社会的関係を構築し、その後の対話をスムーズに進めるために重要な行為である。雑談を通して構築される社会的関係は、社会心理学においてラポールとも呼ばれており、ラポールによって対話相手との間に信頼感や一体感、快適さなどが生じる。本研究では、この雑談の重要性に着目し、より快適な雑談が可能な対話エージェントの構築を目指す。より快適な雑談を実現するために、ラポールの形成に必要な要素について着想を得て、ユーザの選好を考慮して快適な会話を行う適応的対話エージェントを提案する。対話エージェントの応答においてユーザの選好を考慮することは、対話エージェントに対するユーザのエンゲージメントを高め、長期的に会話を継続し、対話を好意的に進行するためにも重要な要素である。ユーザの選好はユーザによって異なり、選好の対象は話し方、対話の進め方、距離感の取り方など多岐にわたる。これらの異なる複数の選好に対してそれぞれ考慮した上で処理を行う必要がある。この問題に対して、本研究では用例ベース対話システムを対象とし、以下の 4 つの協調要素について述べる。一つ目は、ユーザの望む話し方を持った対話システムを構築するために、システムの発話候補に対して変換処理を行う、言語的個人性変換である。これによって、対話エージェントがユーザの望む話し方で対話を行うことができる。二つ目は、用例データベースに対する快適度推定である。ここではユーザにとって快適に対話を進めることができるように、対話エージェントが用いる用例データベースに対して事前に快適度を推定する。三つ目は、ユーザの反応を考慮して応答を行うことで、対話中のユーザに合わせて適応的に最適な応答を選択する、適応的応答選択である。これによって、対話エージェントが対話中のユーザに合わせて快適な応答を行うことができる。四つ目は、対話行為レベルのエントレインメントを考慮した応答選択である。エントレインメントは対話を通して話者同

*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-DD1461010, 2017 年 3 月 16 日.

士が同調する現象であり、対話の自然性やエンゲージメントの増長と関係している。対話行為レベルでのエントレインメントを考慮することで、同調すべきところとそうでないところを考慮した応答選択を行うことができる。我々は、これらの協調要素についてそれぞれ実験を行い、その評価結果から提案手法の有効性を示した。

キーワード

対話エージェント, 用例ベース対話モデリング, 応答選択, 個人性モデリング, エントレインメント

Contents

List of Figures

List of Tables

Acknowledgements

本論文は筆者が奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 博士後期課程に在籍中の研究結果をまとめたものです．本論文の執筆にあたり，多くの方のご支援，ご協力を賜りました．謹んで御礼申し上げます．

同専攻教授 中村 哲先生には指導教官として本研究に取り組む機会を与えていただき，その遂行にあたって多数のご指導，ご助言をいただきました．博士前期課程入学から博士後期課程修了までの 5 年間，貴重な時間を割いて本研究をご指導をいただき，また，時には叱咤激励をいただきました．先生の熱意とお心遣いによって，本研究の成果が得られたものと思っております．心から感謝の意を表します．

同専攻教授 松本 裕治先生には副査としてご助言をいただくとともに，本論文の細部にわたり多数の指導をいただきました．先生の多数のご助言によって，本論文がより良いものとなったと思っております．心から感謝の意を表します．

同専攻助教 吉野 幸一郎先生には副査としてご助言をいただくとともに，日々の研究においても多数のご指導をいただきました．また，本論文を含め，これまでの論文執筆において，非常に多くの時間をご指導にあててくださいました．先生のご指導のおかげで，本論文を執筆することができました．心から感謝の意を表します．

Carnegie Mellon University Linguistic Technologies Institute Assistant Professor Graham Neubig 先生には副査としてご助言をいただくとともに，本研究の遂行および論文の執筆に関して多くのご指導をいただきました．論文執筆におけるご助言と，プログラミングに関する技術のご指導のおかげで，本研究の成果が達成できました．心から感謝の意を表します．

同専攻助教 Sakriani Sakti 先生には副査としてご助言をいただくとともに，発表では多数の議論やご意見をいただきました．また，より分かりやすい発表を行えるように，多数のご指導をいただきました．心から感謝の意を表します．

NTT Media Intelligence Laboratories 東中 竜一郎先生には副査としてご助言いただくとともに，本論文の執筆に関わる多数の指導をいただきました．先生には，本研究以外にも様々な研究テーマおよび研究プロジェクトにお誘いいただき，また多くのご助言と知見をいただきました．心から感謝の意を表します．

I would like to thank Professor David Traum and the members of USC Institute for Creative Technologies for the careful and lively discussion on my internship research project. Professor David Traum has provided me with the

opportunity to study at USC, giving me a chance to challenge the new research topic. The experience of studying abroad in USC has strengthened my confidence in the research.

知能コミュニケーション研究室 秘書 松田 真奈美様，知能コミュニケーション研究室の先生，学生の皆様には大変お世話になりました．本論文の執筆まで研究を続けることができたのは，先生方のご指導と研究室の環境，そして皆様のおかげであると思っています．特に，同研究室の対話研究グループの学生一同には，多数の議論と，研究のお手伝いをしていただきました．心から感謝の意を表します．

これまで温かく見守り，支援をしてくれた家族には本当に感謝しています．この博士課程を全うできたのも，家族の理解あってこそであると思っています．

最後に，本研究に関わった皆様への深い感謝の意を表して謝辞といたします．

Chapter 1

Introduction

1.1 Background

Why do we have a conversation? In the Oxford dictionary, a "*conversation*" is defined as "*A talk, especially an informal one, between two or more people, in which news and ideas are exchanged.*" This "*informal conversation*" plays an important role in building a social relationship with a conversational partner. The preliminary conversation helps to build a social relationship with the partner to enable the main subject of the meeting to be covered smoothly. The informal conversation is one of the most important factors enabling the conversation, by building and maintaining social relationships.

The social relationship through a conversation is called "*rapport*" in social psychology. [Spencer-Oatey, 2005] defined rapport as "*Rapport refers to the relative harmony and smoothness of relations between people, and rapport management refers to the management (or mismanagement) of relations between people.*" [1] Rapport is closely related to the trust, sense of unity, sense of connection, and comfortableness of the attendees in the conversation. [Tickle-Degnen and Rosenthal, 1990, Cassell et al., 1999, Huang et al., 2011] analyzed rapport in conversations, and they clarified the necessity of several factors to build it: the success of chatting, eye contact of the attendees, and back channels at appropriate points. In the research fields of interface and communication, [Tickle-Degnen and Rosenthal, 1990] tried to reveal the relationship between rapport and conversation. These studies investigated that positivity, mutual attentiveness, and coordination are important factors in building rapport in conversation.

By advances in computer science, the definition of the word "*conversation*" is changing and has been extended from human-human to human-machine. Such a machine is called a "*conversational agent.*" The conversational agent can build a social relationship with users to achieve task success and make the user more comfortable through the process of engagement and naturalness of the system

[1] Excerpts from "(Im)Politeness, Face and Perceptions of Rapport: Unpackaging their Bases and Interrelationships", pp. 96

Table 1. Grice's Maxims

Maxim of		Supermaxim
Quantity	Information	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.
Quality	Truth	Do not say what you believe to be false. Do not say that for which you lack adequate evidence.
Relation	Relevance	Be relevant.
Manner	Clarity ("be perspicuous")	Avoid obscurity of expression. Avoid ambiguity. Be brief (avoid unnecessary prolixity). Be orderly.

[Bickmore and Cassell, 2000, Takeuchi et al., 2007, Meguro et al., 2010, Vardoulakis et al., 2012].

In sociolinguistics, it is said that everyone must follow rules called "cooperative principle" to achieve smooth communication in common social situations. Paul Grice [Grice, 1975] introduced Grice's maxims "*Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.*" [2] Table 1 shows the details of Grice's Maxims. These are minimum rules to achieve a basic conversation in common social situations. We should add some other rules for a conversational social agent to deploy positivity, mutual attentiveness, and coordination to build a social relationship. Keeping the positivity is an important factor in building a social relationship, and it is directly related to user feeling on conversation content. Following the user's request inspires positivity in the user, for example, by changeing the speaking styles according to the user's request. We have to listen to the partner attentively and to react to the partner by considering user preferences. Good reactions considering mutual attentiveness let conversational partner feel connectedness.

Synchronizing with the partner through a conversation is an important social phenomenon. It is well known as entrainment, synchrony, and coordination in linguistics and social psychology. A study of analyzing entrainment shows this

[2]Excerpts from "Logic and Conversation", pp. 45

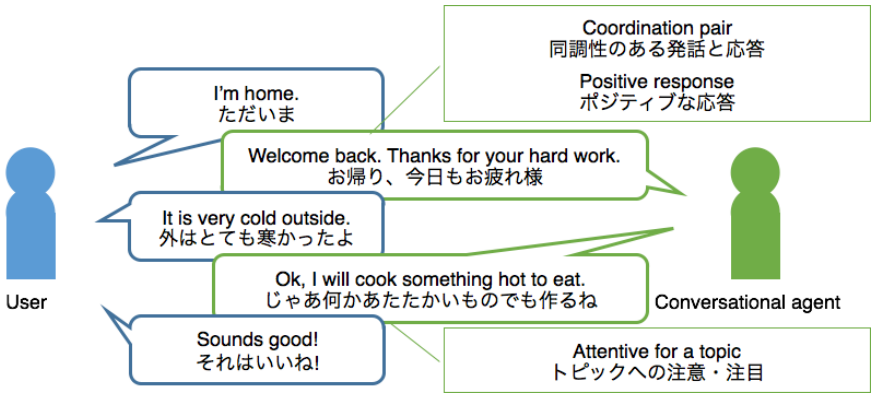

Figure 1. Example of good conversation between humans

phenomenon induces an increase in naturalness and engagement during a conversation. Entrainment causes adaptation in the speaking style of the system, in the quality of the response, in the attentiveness of the partner, and in the synchrony of the response. These rules are focused on adapting to a user such as the speaking style of the system, quality of the response, attentiveness of the partner, and synchrony of the response. We assume that adaptiveness to the user preferences is the noticeable factor for satisfactory conversations.

Some studies showed the importance of adaptability in conversational agents and used dialogue strategies to adapt the agent to user preferences. For example, some conversational agents have dialogue strategies to remember user information like names, hobbies, birthdays and more and to generate system responses based on user information [Elzer et al., 1994, Wärnestål et al., 2007]. These studies mainly deployed adaptability to user information and did not focus on user preferences. Adaptability to not only user information but also user preferences helps conversational agents build a social relationship with a user. In this thesis, we focus on conversational agent's adaptability to user preferences because such adaptability is an important factor in establishing a social relationship with a user in studies of social psychology. We show examples of the conversation between a user and a conversational agent with/without adaptability to clarify the importance of adaptability for the conversational agent.

Figure 1 is an example; the conversational agent has adaptability to the user. First, the user says "I'm home.", and the conversational agent responds "Welcome

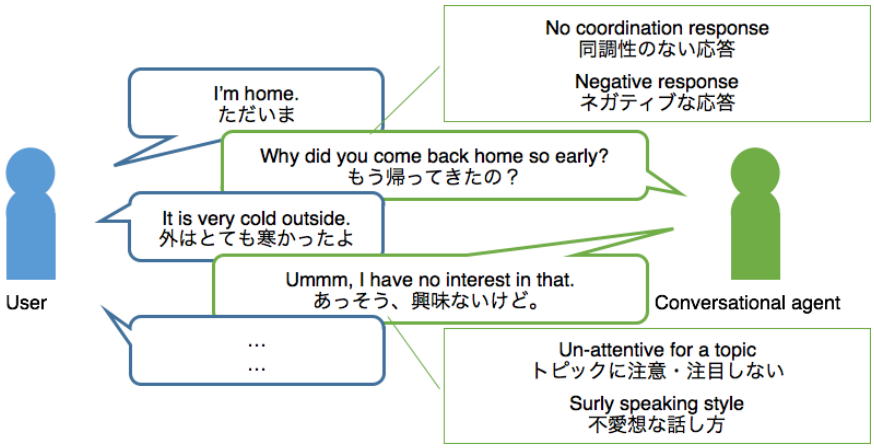

Figure 2. Example of bad conversation between humans

back. Thanks for your hard work." This pair of "I'm home" and "Welcome back" is not only an appropriate pair but also a coordination pair which makes the user be in sync. The response of "Thanks for your hard work." expresses considerateness that relates to the positivity of important factors in building a social relationship.

Next, the conversational agent responds "Ok, I will make something hot to eat." to "It is very cold outside." This response focuses on "cold," which is provided from the user utterance. A topic transition that focuses on the conversational partner is mutual attentive action, and it makes the user feel connectedness.

Finally, the user responds "Sounds good!" which expresses positivity and involves sharing the positivity between the user and the conversational agent. The user and the conversational agent build a social relationship through a conversation that includes adaptability to the user.

Figure 2 is a bad example because the conversational agent has no adaptability to the user. The first pair of "I'm home" and "Why did you come back home so early?" is an appropriate pair; however, it includes an incoordination and negative response. The incoordination and negative response make the user feel dissatisfaction, which inhibits building a social relationship.

The next pair of "It is very cold outside. " and "Ummm, I have no interest in that." is a surly and inattentive response to the user; such a surly speaking style

makes the user feel dissatisfaction, and an inattentive response does not help the user be in sync during a conversation.

Finally, the user stops speaking and gives up on the conversation. The social relationship is not built through this conversation, and the user feels annoyed with the conversational agent.

Modeling a relationship between conversation and building a social relationship has been studied. [Matsuyama et al., 2014] proposed a computational model of rapport enchantment, maintenance, and destruction with dialogue strategies.

Studies of conversational agents have tried to use a social relationship to proceed with the main subject of the meeting smoothly and effectively. [Bickmore et al., 2011] show the effects of relationship-building behaviors in a museum guide agent, and these effects enable not only engagement but also learning gains. In these studies, the conversational agent tries small talk with a user by utilizing pre-defined rules or Wizard of Oz methods. These methods are effective in only limited domains and tasks and require well-prepared situations. These studies show us a large and difficult problem in developing a conversational agent that builds a social relationship with a user through an open-domain conversation without tasks.

We show that a conversation with a user and a conversational agent is not only for entertainment but also for building a social relationship with a user, and it helps to increase engagement, satisfaction, naturalness and smoothness. The social relationship that is built through small talk also helps a task-oriented conversational agent to proceed with the main subject of a meeting smoothly. To build a social relationship requires certain factors through a conversation, and we believe that a factor in building a social relationship that current conversational agents lack is the adaptability to user preferences. From related studies and examples, we show the adaptivity to user preferences is an important factor in a satisfactory conversation and in building a social relationship. These benefits and required factors in building social relationships are supported by the theory of rapport in social psychology.

In this thesis, we propose an adaptive conversational agent considering user preferences to increase the user satisfaction with social relationships. Previous conversational agents provide a specific response to a specific user utterance with-

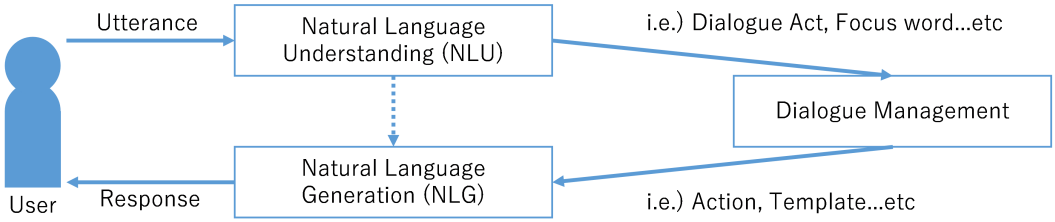

Figure 3. General architecture of the task-oriented conversational agent.

out considering user preferences and progress in the conversation. An adaptive conversational agent that considers user preferences can generate responses that are appropriate for user preferences and for enabling progress in the conversation.

1.2 Fundamental approaches

Studies of conversational agents are roughly classified into two types: task-oriented and non-task-oriented.

The task-oriented conversational agent tries to solve a task through a conversation, such as airplane ticket booking and acting as a travel guide. They are based on the architecture that has the following three modules: Natural Language Understanding (NLU), Dialogue Management (DM), and Natural Language Generation (NLG). NLU analyzes a user utterance and extracts features that are machine understandable for DM and NLG. DM decides the agent action by considering NLU features and a dialogue state of the current conversation. NLG generates an agent response based on NLU features and the agent action provided from the DM. In the task oriented conversational agents, DM mainly works to solve tasks and basically bring a dialogue to a goal state. The DM is the core module of this architecture, and some studies developed DM based on machine learning such as a partially observable Markov decision process (POMDP) to increase the performance especially for task-oriented conversational agents [Young et al., 2013, Yoshino and Kawahara, 2015]. These studies provided training in DM to increase the number of task successes and used only limited dialogue states to avoid the problem of data sparseness. We show this architecture of the task-oriented conversational agent in Figure 3.

The non-task oriented conversational agent tries to interact with users. Thus,

they are also called chat-oriented conversational agents, and their role is to make users enjoy a conversation. As mentioned in Section 1.1, the non-task oriented conversational agent refers to small talk used to build a social relationship before solving tasks. Therefore, an abstract task assigned to the non-task oriented conversational agents is to satisfy the user with a conversation.

The abstract task includes difficulties that are not included in the task oriented conversational agent's task. One difficulty is how to define the task success. The progress of tasks is observable for task-oriented conversational agents from dialogue states or the number of the solved small tasks. Defining the progress of the abstract task is difficult for non-task oriented conversational agents because the satisfaction of the talk depends on each user's subjectivity. Some studies try to define the objective evaluation function by using other kinds of measures such as turn-length.

Another difficulty is data sparseness. The non-task oriented conversational agent has to track topics that are provided by a user. These topics might be unlimited in scope because it is an open domain. We need an architecture that handles an open domain for the non-task oriented conversational agent. We should note that expanding topics requires a lot of training data because dialogue states will be expanded in proportion to the topic size.

Studies on the non-task oriented conversational agent have proposed some kinds of architecture. ELIZA is one of the most famous rule-based conversational agents, which were originally used for counseling. This conversational agent tends to respond to user utterances with repetition or general questions [Weizenbaum, 1966], and it does not require training data. The rule-based architecture was also regarded as one of the first programs capable of passing the Turing Test; however, most users noticed its simple behavior and were disappointed. Subsequently, studies of this type of conversational agent proposed an architecture that enables processing on a minimum unit of conversation. The minimum unit of conversation means a turn consisting of an utterance and a response. Specifically, this architecture tackles a minimum task like a question and answering system that chooses an appropriate response to an input utterance from an example database including a collection of pairs of an utterance and a response. This minimum task is defined more than the abstract task, is not related to a di-

alogue state, and is easy to evaluate clearly. The non-task oriented conversational agent based on this architecture selects a response by considering only an input utterance, and it is not related to a dialogue state. Using no dialogue state means avoiding the problem of data sparseness. This architecture which tackles the minimum task, is called "Example Based Dialogue Modeling (EBDM)" and is widely used to develop the non-task oriented conversational agent [Lee et al., 2009, Kim et al., 2010]. The current study of non-task oriented conversational agents such as "Rinna (りんな)" [Wu et al., 2016] is based on the EBDM architecture and machine learning to choose an appropriate response.

In almost all cases, the EBDM architecture aggregates modules of NLU, DM, and NLG into one core module as the response selection module, and it changes the aim from handling appropriate agent actions and responses to finding appropriate responses. In the EBDM architecture, a response is chosen from pairs of query utterance q, and response utterance r if the query utterance q is the most similar to the current user query utterance q' according to a pre-defined similarity score. The corresponding response utterance r is selected as the conversational agent response. The selection is defined as:

$$\langle \hat{q}, \hat{r} \rangle = \underset{\langle q,r \rangle \in \mathbf{e}}{\mathbf{argmax}} \, \mathrm{sim}(q', q). \tag{1}$$

For the pre-defined similarity score, previous studies proposed useing TF-IDF based cosine similarity [Banchs and Li, 2012], syntactic semantic similarity [Nio et al., 2012], or recursive neural network-based paraphrase detection [Nio et al., 2014b]. In this thesis, we used cosine similarity as the similarity measure $\mathrm{sim}(q', q)$ because it is one of the simplest and most effective algorithms to measure similarity [Navarro, 2001]. The performance of the selection directly affects the performance of the conversational agent.

This response selection module is often compared with a response generation module in studies of conversational agents. These two modules have their own advantages and disadvantages, and we use either module depending on the purpose and task. The response generation module, which generates a new response to the user utterance, has the advantages of high adaptability and diversity of responses. However, the response generation module requires a lot of training

data and annotations to train models, and we cannot completely handle the module. Furthermore, this module has a fatal risk that often generates some ungrammatical sentences. The response selection module has lower adaptability and diversity of responses than the response generation. However, this module can respond with satisfactory and natural responses by using a lot of examples, and we can handle the module easily. Therefore, an EBDM architecture with a response selection module is used in studies of non-task oriented conversational agents.

We note that almost all studies using EBDM are different from the architectures of the task oriented conversational agent because the latter architecture has no DM. In this case, the conversation agent does not have any states of users or systems. Therefore, EBDM can work on a simple architecture. Few studies of the non-task oriented conversational agent use DM to handle the user's or conversational agent's errors; instead they use some strategies that consider user engagement or response appropriateness [Yu et al., 2016b,a].

In EBDM architecture, the size and quality of the example database are important factors for the quality of the non-task oriented conversational agent. The example database is often constructed using an existing data source such as human-human conversation logs [Murao et al., 2003], movie or television scripts [Banchs, 2012, Nio et al., 2012], or Twitter logs [Bessho et al., 2012]. They extract tri-turns, which are three turns consisting of dialogue from two people (i.e., A:"I'm hungry", B:"Me too. Let's go to restaurant", and A:"Sounds good!" is collected as a tri-turn; however A:"I'm hungry", B:"Me too. Let's go to restaurant" and C:"Can I join you?" is not collected as a tri-turn). These construction methods do not consider the example quality explicitly. While using dialogue corpora that are well disciplined, these heuristic rules work well in constructing an example database.

The standard EBDM architecture basically has one modules and one database. In Figure 4, we show the standard EBDM architecture.

EBDM for the non-task oriented conversational agent presents a light-weight and highly portable yet feasible alternative to more conventional methods that require NLG. A lot of studies of EBDM have tried to increase the accuracy of selecting an appropriate response and to construct a high-quality example database.

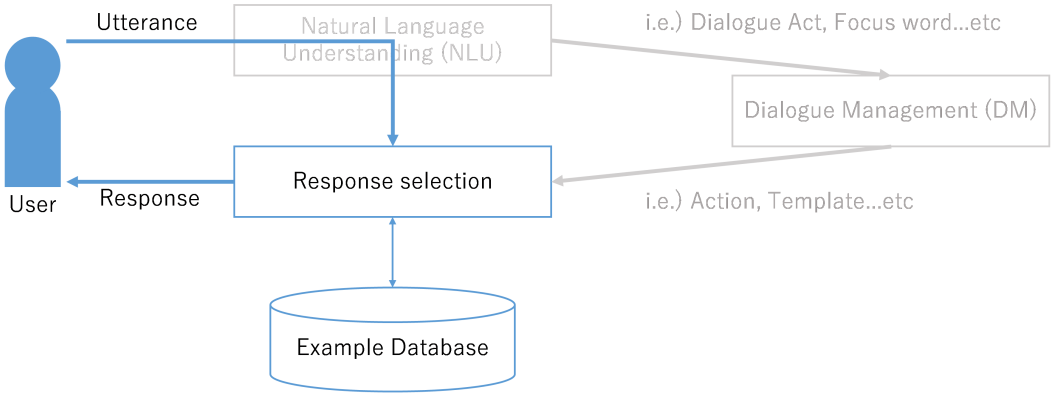

Figure 4. Overview of EBDM architecture.

The many attempts of these studies show the importance and practicability of EBDM.

In this thesis, we used EBDM to develop an adaptive conversational agent considering user preferences; however, various problems lie ahead. The first problem is the linguistic individuality of agent responses. The conversational agent changes its speaking style to adapt to the request of users. However, an example database is constructed from corpora without considering the speakers, and it does not allow us to control the speaking style.

The second problem is the quality of agent responses. The conversational agent response has to be not only appropriate but also include a coordination, positivity, and considerateness.

The third problem is adaptability to user preferences using agent responses. The conversational agent selects the best response by considering user preferences to increase user satisfaction.

The final problem is synchronized responses between a user and a conversational agent. EBDM always outputs the same response to the specific utterance because it does not consider the dialogue state. We cannot consider all of the dialogue states because of a problem of data sparseness. However, some states that are related to building a social relationship like entrainment can be used to increase user satisfaction.

1.3 Problems and related studies

Positivity, mutual attentiveness, and coordination are important factors to enable developing an adaptive conversational agent that considers user preferences. As mentioned in Sections 1.1 and 1.2, we focus on four factors: the speaking style, quality of the examples, the adaptiveness of the response selection, and the adaptiveness of the response coordination. Related studies and problems are described in the following sections.

1.3.1 Linguistic individuality of agent responses

Some studies report that the speaking style of the conversational agent affects the impression that a user feels toward the conversational agent; this effect of the speaking style can be used to make conversational agents more entertaining, attractive, friendly, and humanlike through conversation [Miyazaki et al., 2016]. The effect includes not only producing entertaining and/or friendly expression but also building a social relationship [Ogan et al., 2012].

To express the appropriate speaking style, these related studies use attributes such as sex, age, or traits (i.e., Big Five Traits [Gosling et al., 2003]). These studies assume that the speaking style is decided by the defined attributes. However, the specific speaking style is affected not only by attributes of age, sex, and trait but also individuality. To handle a speaking style for conversational agents, some studies generated sentences according to a certain speaking style based on the rule-based sentence generation [Mairesse and Walker, 2011], rule-based sentence transforming [Miyazaki et al., 2016], and personality infused language models selection [Isard et al., 2006]. The rule-based sentence generation method and the language models selection provide various speaking styles based on the Big Five Traits and the Costa and McCrae's five-factor model; however, they require a large number of rules or language models, which are trained from corpora or made by humans. The rule-based sentence transforming method provides speaking styles based on the properties of sex and age, making it possible to transform an utterance that expresses a certain speaking style from an original sentence.

These definitions and limitations of attributes inhibit increasing the variation in speaking style. Specifically, we have costs to analyze and annotate for each target person and to define attributes that are necessary to express a target

speaking style. These costs need to be decreased to give the variation of speaking style in conversational agents.

In this thesis, we handled the specific speaking style that is closely related to the "individuality." Specifically, we define the "individuality" as the speaking style that is recognized from a particular single speaker's corpus. The user's favorite speaking style of individuality may helps them feel more entertained and engaged with the conversational agents. An architecture needs to be proposed to transform a speaking style using a speaker's small corpus and statistical machine translation for reducing the costs of data preparation and for handling various speaking styles of individuality.

1.3.2 Quality of agent responses since construction

Previous studies of EBDM defined the research task as selecting the most appropriate response to the input utterance and avoided problems of evaluation and data sparseness. Some studies improved response selection functions with various metrics: TF-IDF weighted vector space similarity [Banchs and Li, 2012], WordNet-based syntactic-semantic similarity [Nio et al., 2012], or recursive neural network-based paraphrase detection [Nio et al., 2014b].

The other important factor to improve EBDM is example database construction. The example database has been constructed using corpora that have pairs of a query and a response. In most EBDM studies, it is based on the intuition of the engineer who built the system, and a post-hoc subjective evaluation is used to validate its correctness. Some studies tried to clarify this intuition by using heuristic rules by considering the turn changes of speakers, question-answer pairs, or tweet ids [Murao et al., 2003, Banchs, 2012, Nio et al., 2012, Bessho et al., 2012]. These construction rules consider the appropriateness of an example pair alone and not their quality. To make matters worse, these methods cannot evaluate the appropriateness of responses without running the system.

Considering not only the appropriateness but also the positiveness, attentiveness, and coordination of the response to the utterance helps to increase the quality of example pairs. We herein propose a method for predicting user satisfaction, one that can evaluate the quality of examples immediately based on the evaluation and error analysis of an already-finished dialogue [Ultes and Minker,

2014, Schmitt et al., 2011, Higashinaka et al., 2010, Engelbrech et al., 2009] to predict user satisfaction. This prediction method enables securing and increasing the performance of a conversational agent.

1.3.3 Adaptation to User Preferences in Agent Responses

Some studies of conversational agents have reported that adaptability to users contributed to increasing task success, engagement, and user satisfaction. These studies propose that a conversational agent adapts to personal knowledge and information mainly by using some personal questions such as a user's hobby, birthday, and name [Elzer et al., 1994, Wärnestål et al., 2007]. These studies focus on specific entities of user information and adapt to the user by utilizing handcrafted rules. These studies consider user preferences in indirect ways; however, through some properties, this study directly focuses on such preferences to the user's response. The EBDM architecture always gives a specific response to an utterance, and this property makes adapting to user preferences difficult. Through directly considering user preferences enables avoiding the problem, we just introduce a new state of use preference to cope with this problem of sparsity.

As mentioned in Section 1.2, the non-task oriented conversational agent cannot track all of the user states to consider user preferences because it assumes open-domain and dialogue states are sparse. Only user preferences need to be modeled without sparse states to adapt to user preferences on the non-task oriented conversational agent.

1.3.4 Synchroneity responses to user

Entrainment (synchrony) is a conversational phenomenon in which dialogue participants synchronize with each other. Previous studies reported that entrainment modeling helps to improve the performance of speech recognition and turn taking [Campbell and Scherer, 2010, Fandrianto and Eskenazi, 2012, Levitan, 2013]. Other studies analyzed lexical entrainment and found that entrainment is correlated with dialogue success, naturalness, and engagement [Nenkova et al., 2008].

Previous studies showed the importance of entrainment; however, a conversational agent that synchronizes with users to increase naturalness, engagement, and satisfaction has not yet been proposed. Therefore, we propose a function to

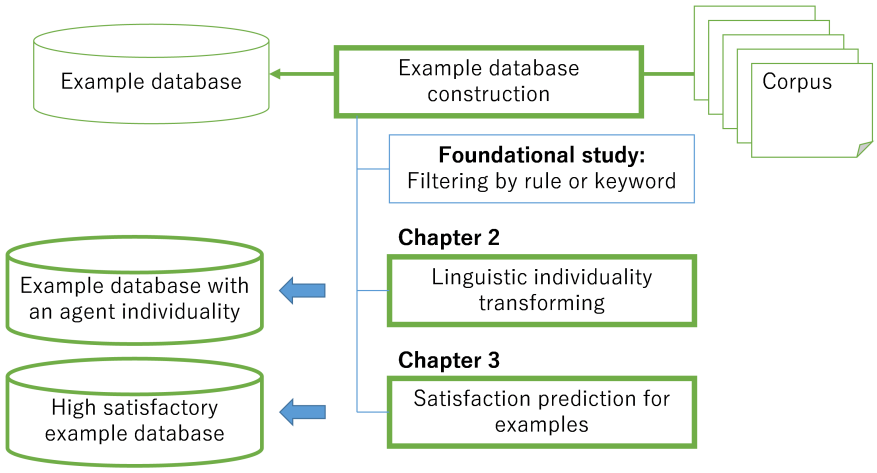

Figure 5. Construction of example databases

consider the appropriateness of synchronization to the user regarding the response selection of EBDM.

1.4 Approaches in this thesis

While EBDM works to enable conversations with a human and a conversational agent, a lot of problems remain, as mentioned in Section 1.3. To cope with these problems, we separated improvements into two parts. The first part is an improvement in the example database construction by considering agent individuality and satisfaction with the example database. The second part is an improvement in response selection. We propose adaptive response selection methods to consider user satisfaction and dialogue entrainment. Each consideration corresponding to the problem defined in section 1.3 is described as follows.

1.4.1 Linguistic individuality transformation based on statistical machine translation

Previous studies adopted a personality to conversational agent responses by using some attributes such as age, sex, or traits, based on transforming, generation, or selection. These methods enable expressing speaking style based on attributes in conversational agents; however, it cannot generate responses according to the specified individuality with these attributes.

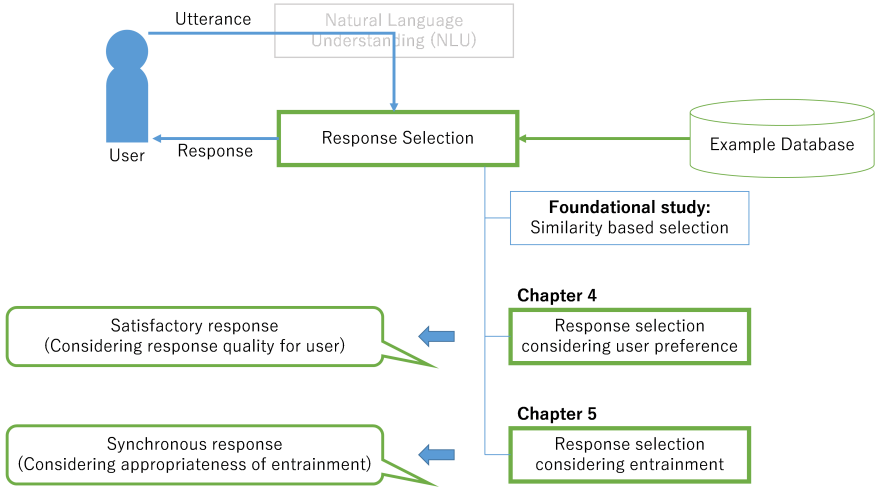

Figure 6. Response selection

In this study, we define individuality as a specific person's speaking style expressed without attributes. Expressing individuality helps conversational agents to make themselves more attractive, friendly, humanlike, and entertaining because imagining a certain character is easy if the conversational agent has that individuality. The proposed framework enables transforming a response based on translation rules and language models, and this framework works as a data-driven approach. Therefore, the proposed method handles individuality directly from corpora without considering actual attributes that existing approaches consider, and this property enables the system to learn individuality totally.

1.4.2 Satisfaction prediction for examples

Previous studies of EBDM collected better examples based on some heuristic rules, making it difficult to consider the quality of responses. In contrast, we consider the quality of responses on the example database to increase the quality of an example database. The proposed method predicts user satisfaction expected for a response and selects examples based on the predicted user satisfaction. In the prediction, we annotated the general score of a query-response pair using the averaged value of annotated satisfaction for the example.

Our proposed method estimates the general user preference scores to predict user satisfaction for a response based on support vector regression, which

estimates the satisfaction score annotated for the query-response pair.

1.4.3 Adaptive response selection based on collaborative filtering using user feedback

We utilized general user preferences as scores for query-response pairs; however, once the conversational agent starts to talk with the user, the conversational agent can get information to estimate the personal satisfaction scores linked with the personal preference using the responses and reactions of the user. In the response selection, our method predicts user's personal satisfaction and adapts to the user based on collaborative filtering. For our method to work as desired, we divided the process into two parts. The first part is a multi-response example database, which has multiple responses for one utterance. The second part is a new response selection that considers not only similarity but also a tendency of the preference of the user – what kind of response the user prefers. We developed our proposed adaptive response selection based on the technique of collaborative filtering.

1.4.4 Response selection based on entrainment analysis

Previous studies showed us that entrainment is strongly related to rapport; however, we did not know how to use entrainment in a conversational agent. The phenomenon of entrainment is related to user preferences, and using entrainment in dialogue enables improving conversational agents and making them friendlier. Our response selection method is based on entrainment analysis to increase the performance of conversational agents.

The method considers words and dialogue acts in user utterances and selects a response according to the type of dialogue act and the appropriateness of entrainment (synchronization) that is expected for the previous dialogue act. We investigated the relationship between dialogue acts and lexicons on the basis of the appropriateness of synchronization clarified using the analysis of lexical entrainment. We developed a method for selecting responses by considering language models to synchronize with a user on a lexical level appropriately.

The method adapts to the user through synchronizations of conversations.

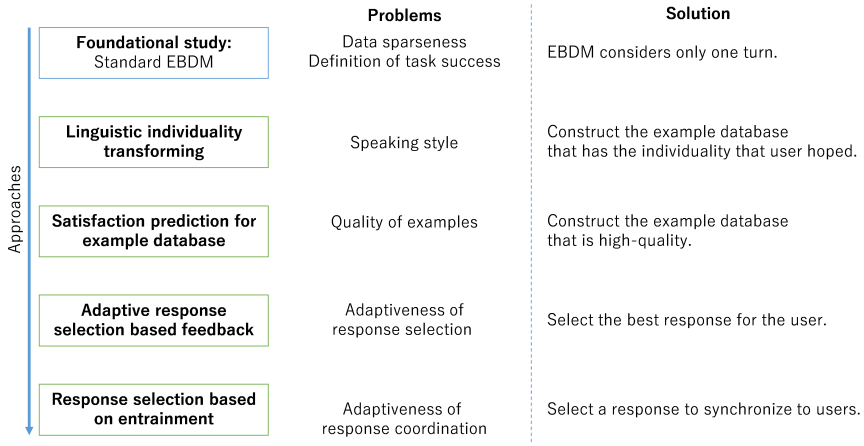

Figure 7. Focuses and approaches of our research

This adaptivity enables synchronizing with users and selecting appropriate responses.

1.5 Contributions of this thesis

We summarized problems and solutions in Figure 7, and described the details. Previous studies constructed an example database based on rule-based filtering, and they selected responses considering the similarity between user utterances and example queries. These studies have various problems that are related to the speaking style, quality of examples, adaptiveness of the response selection, and the adaptiveness of response coordination. These problems inhibit developing the adaptive conversational agent.

We focused on the individuality of agent responses and constructed an example database that has individual responses, that the user hoped. Differences from previous rule-based transforming or language model based filtering are cost and portability, and our machine translation based individuality transforming is a way to improve both. The proposed method requires only corpora, which are without annotations of attributes, and enables giving various kinds of individuality for conversational agents at low-cost. The variety of individuality helps the conversational agent be more entertaining, attractive, friendly, and human-like. In Chapter 3, we describe our linguistic individuality transformation method

and the construction of translation model. Finally, we examine this individuality transformation method using two groups of target speakers and show that translated utterances are highly subjective scores of individuality.

We proposed satisfaction prediction for constructing an example database based on filtering low-quality examples. Previously proposed methods evaluated an example, and the example is used in any dialogue; however, our method often predicts the expected satisfaction. Securing the quality of the example database and predicting the expected satisfaction of the user in advance is an important sub-task of EBDM. The quality of the example database is one of the most important factors in improving the performance of the EBDM, and enabling the agent to have a satisfactory conversation with users is a necessity. In Chapter 4, we describe our satisfaction prediction method for an example database and evaluate the filtering based on our proposed prediction. An experimental evaluation showed that user satisfaction improved significantly using an example database constructed by filtering based on our proposed prediction.

In chapter 5, we also propose a method for adaptive response selection based on user feedback and collaborative filtering to consider user preferences of responses. Foundational studies of EBDM have no adaptability on response selection. We developed a multi-response example database and response selection modules to select a response from the example database in accordance with user preferences. When the conversational agent obtains user feedback, the conversational agent adapts to user preferences and selects the best response for a user who is just talking. The adaptiveness of the response selection is the core of the adaptive conversational agent, and it is the basis of the new modeling of EBDM. Finally, we examined the effectiveness of an adaptive response selection and showed that the user satisfaction was improved by the response selected by the proposed method in comparison with the response selected by conventional methods in a subjective evaluation.

We proposed a response selection method based on the entrainment analysis of combinations of lexical and dialogue acts. Previous studies used the effect of entrainment to improve acoustic speech recognition or turn-taking prediction. We focused on the effect of entrainment on response selection and developed a module to select more appropriate responses by considering the similarity to the

language model of the user. Entrainment is an important factor to increase dialogue success, naturalness, and engagement, and it makes the conversational agent more adaptable. In Chapter 6, we describe our response selection method based on entrainment analysis. An experimental evaluation showed that our response selection method selects an appropriate response considering entrainment in the objective evaluation.

In this thesis, we proposed four improvements enabling the development of a conversational agent that is adaptive to user preferences, thereby increasing user satisfaction. We assume these proposals solve problems that have not been achieved in previous studies. We examined the proposed methods, and each result showed its effectiveness: the linguistic individuality transforming substantially changes and increases the subjective score of individuality, the satisfaction prediction for examples increases the user satisfaction with a filtered example database, the adaptive response selection increases the overall user satisfaction, and the response selection based on entrainment analysis increases the performance of the response selection. These results show that the proposed methods solve problems that have not been achieved in previous studies and enable developing an adaptive conversational agent considering user preferences. While these improvements in modules work a conversational agent can benefit by increasing user satisfaction based on adaptability to user preferences.

Chapter 2
Linguistic Individuality Transformation

2.1 Introduction

In this chapter, we describe a probabilistic framework for transforming linguistic individuality that creates individual responses for an example database. Linguistic individuality transformation uses a technique inspired by statistical machine translation and paraphrasing. We try to transform a response into many kinds of individuality by using a general framework and plural different way of paraphrasing correspond.

The proposed method enables increasing the number of responses, which increase the potential to interact with the specific individuality that the user hopes. The proposed method has innovations that are the data-driven approach, the general statistical framework, and paraphrasing techniques of unique expressions, to make it possible to train an individuality transforming from specific speaker ' s corpus with a low-cost.

2.2 Linguistic Individuality

In language, the words chosen by the speaker or writer transmit not only semantic content but also other information such as aspects of their individuality, personality, or characteristics. While not directly related to the message, these aspects of language are extremely important to build a social relationship between the person creating the message and its intended target. These speaking styles affect not only building a social relationship but also making a user to attractive, friendly, humanlike, and entertaining. We can assume that this observation will also carry over to human-computer interaction [Metze et al., 2009].

Previous studies tried to convert to an utterance that expresses specific speaking style, and define the speaking style as personality, which is generalized by using some attributes such as sexes, ages, or traits (i.e., Big Five Traits [Gosling

et al., 2003]). These studies show the importance of controlling a speaking style on the conversational agent, and the necessaries of the ability to express a more rich variety of individuality and atmosphere depending on the type of user or scene [Isard et al., 2006, Mairesse and Walker, 2011]. For example, in a situation where a conversational agent is used to represent famous characters in movies or comics to give more great impressions for a user, we would like to reproduce the character well knows and unique expressions.

These previous studies tried to paraphrase an utterance to convert a speaking style based on words or phrases level. Words or phrases level conversion is not enough to convert the speaking style at all, it is necessary to convert based on the grammar or more abstract level. However, grammar level conversion without changing the semantics of an utterance is difficult. We propose a linguistic individuality transforming method based on paraphrasing with the word and phrase level as same as previous studies.

In order to tackle this challenge of unique expressions, we must propose a method to reproduce the speaking style of a specific speaker. In this thesis, we define the individuality that is extracted speaking style from a specific speaker's corpus, and we don't generalize the individuality by using some attributes like previous studies. This definition makes it possible to transform and distinguish two persons who are resembled on attributes as separate human beings. We clarify the effectiveness of the proposed data-driven definition and approach to transform linguistic individuality.

2.3 Proposed method

2.3.1 A probabilistic framework for transforming linguistic individuality

We describe the proposed method for transforming of speaker individuality. To create a method capable of this conversion, we base the previous studies that have tried conversion of writing or speaking style [Xu et al., 2012, Brill and Moore, 2000, Neubig et al., 2012]. These studies enable converting written texts from spoken texts by using the framework of statistical machine translation (SMT). The SMT generates translations based on statistical models that derived from

bilingual parallel corpora [Brown et al., 1990, 1993, Och and Ney, 2004, Koehn, 2009]. The SMT is the data-driven method that has no constraint of a specific pair of languages, and is widely used to try transforming in pairs of various languages.

We build upon the study of [Neubig et al., 2012], which was originally conceived for translation from spoken to written text, or for translation of text from one style to another. Given a string of input words V (representing a sentence) and a string of words W (representing a sentence in target speaking style), we transform V to W using the noisy channel model. In consideration of available corpora, the posterior probability $P(W|V)$ is decomposed into the translation model probability $P(V|W)$, which must be estimated from a corpus of parallel sentences, which is more difficult to find, and language model probability $P(W)$, which can be estimated from a corpus of only output side text that we can secure in large quantities:

$$P(W|V) = \frac{P(V|W)P(W)}{P(V)}. \tag{2}$$

Given this probabilistic model, the output is found by searching for the output sentence $\hat{W}$ that maximizes $P(W|V)$. $P(V)$ is not affected by choice of W, so this maximization is expressed as follows:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(V|W)P(W). \tag{3}$$

We note that the language model probability $P(W)$ tends to prefer shorter sentences, we also follow standard practice in machine translation [Och and Ney, 2002] in introducing a word penalty proportional to sentence length $|W|$. We combine these three elements in a log-linear model, with parameters λ_{tm}, λ_{lm}, and λ_{wp} as follows:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, \lambda_{tm} \log P(V|W) + \lambda_{lm} \log P(W) + \lambda_{wp}|W| \tag{4}$$

Following this framework, we consider a setting in which we translate from utterance V that expresses the individuality of the source speaker to utterance W that expresses the individuality of target speaker. However, compared to the previously mentioned style transformation or standard SMT, we are faced with a drastic lack of data. The amount of target side data W is limited, and we will

often have no parallel data with identical semantic content expressed with the individuality of the target and source speakers. In fact, when we had one author of the section attempt to make this data in preliminary experiments, we found that even when an annotator is available, a creation of the data is quite difficult and time-consuming. If the annotator attempted to follow the semantic content of the input faithfully, it was difficult to express a rich variety of individuality, and when the annotator attempted to edit more freely, the individuality was expressed abundantly, but in many cases the semantic content changed too much to be used reliably training or testing data for the system.

2.3.2 Language model

For transforming individuality, it is necessary to build a language model that expresses the individuality of the target speaker. To build a language model, we need to collect data that expresses the target speaker's speaking style. It is better if the data used to train the language model matches the content of the data to be converted. Therefore, an initial attempt to create a language model that expresses the speaking style of the target will start with gathering data from the speaker, and training an n-gram language model on this data.

When we collect the utterance of only one target speaker and build a language model, it is difficult to collect a large number of utterances from any one speaker. Thus the contents covered by the language model are restricted. Therefore, a language model made with only data from the target speaker cannot estimate the language model probability $P(W)$ accurately. To remedy this problem, in this section, we build a target language model that interpolates a small language model $P_t(W)$ that is trained as explained in the previous section and a language model $P_g(W)$ that is trained from a large-scale corpus. Using an interpolation coefficient λ, we combine these two models using linear interpolation

$$P(W) = \lambda P_t(W) + (1 - \lambda) P_g(W). \tag{5}$$

We calculate λ to generate language model $P(W)$, such that we achieve the maximum language model probability on a held out development set also created using data from the target speaker. Note that this framework is flexible, so we

could also add an additional language model considering the personality of the speaker [Isard et al., 2006], but in this section for simplicity we only use two models: the general domain, and with the target speaker's individuality.

2.3.3 Translation model

Now that we have modeled individuality in the language model, next, we must create a translation model $P(V|W)$ that expresses the possible transformations changing the speaking style, but not the semantic content, of the utterance. However, as mentioned in Section 2.3.1, it is non-trivial to collect a corpus of sentences spoken by the source and target speaker while having the same meaning, thus, we will have to create translation models without relying on a parallel corpus. In this section, we solve this problem by building the translation model using techniques of paraphrasing. We define 4 translation models with different roles, and describe details as follows:

1. **Content** is the thesaurus based translation model. Thesauri are language resources, specifying groups of synonyms, and are a good resource for reliably finding semantically plausible transformations. The most widely used thesaurus in the NLP community is Wordnet [Miller, 1995], and its counterpart in Japanese, our target language, is Japanese Wordnet [Bond et al., 2009]. The **Content** translation model has the advantage of providing broad coverage, but it also consists mainly of synonyms for content words, and does not have data regarding synonymy of fillers, exclamations, particles and other function words.

The **Content** translation model, which is built by using the thesaurus, is used to find replacement candidates based on synonyms for content words, similarly to previous studies on paraphrasing using thesauri [Inui and Fujita, 2004]. We build **Content** translation model according to the following procedure.

1. For each word in the input, search the WordNet with the word as the query.
2. When the word is found, acquire all synonyms from WordNet using the synset.
3. Calculate the translation model probability for all words, and store them in the translation model.

We note that this **Content** method can find can find potential candidates for translation, it gives us no mechanism to determine how reliable these candidates are. We have to calculate translation model probabilities according to any basis. In this thesis, we base our method on techniques to acquire synonyms from non-parallel corpora [Dagan et al., 1999, Barzilay and Lee, 2003]. In the previous studies, similarity of the word itself is calculated from a non-parallel corpus according to the contextual similarity of words.

In order to calculate this contextual similarity, we prepare a bigram language model with vocabulary L, and decide the similarity $\mathrm{Sim}(w, v)$ for two words w and v as follows:

$$\mathrm{Sim}(w, v) = 1 - \frac{1}{2|L|}\Big(\sum_{l \in L}\Big|P(w|l) - P(v|l)\Big| + \sum_{l \in L}\Big|P(l|w) - P(l|v)\Big|\Big). \quad (6)$$

Similarity $\mathrm{Sim}(w, v)$ is decided by the similarity of n-gram distributions, based on the distributional hypothesis that words that appear in similar contexts have a similar role. For the calculated similarity $\mathrm{Sim}(w, v)$, we normalize over values of $\mathrm{Sim}(w, v)$ for all words, so that the probabilities sum to one

$$P(w|v) = \frac{\mathrm{Sim}(w, v)}{\sum_{L \in l} \mathrm{Sim}(l, v)}. \quad (7)$$

From these calculation, we can approximate translation model probability of words w and v without using a parallel corpus. In this calculation, we can extract similar words more strictly according to the language model by using the larger n of the language model length, however, we use n=2 to find various paraphrases because we cannot use the combination of target speaker's language model and source speaker's language model.

We show an example of the translation model acquired by **Content** method in Table 2.

2. **Particle** is the translation model that is collected according to POS tags and n-gram clustering. The **Particle** translation model mainly targets particles of the end of the sentence, auxiliary verb particles, and fillers. These particles have been noted as playing an important role in expressing individuality especially in Japanese [Teshigawara and Kinsui, 2012], and these elements are very important

Table 2. A sample of the Content translation model.

Source	Target	TM prob.
カメラ (camera)	カメラ (camera)	0.95
	キャメラ (kamera)	0.01
	ビデオカメラ (video camera)	0.01
	写真機 (photo machine)	0.01
	and other 2 words	
良い (good)	良い (good)	0.4
	いい (nice)	0.4
	よろしい (fine)	0.01
	見事 (excellent)	0.01
	and other 42 words	

in expressing a number of aspects of language [Chung and Pennebaker, 2007]. The **Particle** translation model covers high frequency function words except **Content**.

We build **Particle** translation model according to the following procedure.

1. Prepare a list of function words by performing POS tagging on the training corpus and extracting all non-content words.

2. Count all 3-grams in the target speaker's utterances.

3. Find groups of 3-grams that have a function word in the second position and the same first and third words, and add them to the set of potential synonyms.
   e.g.) は とても よい (that's so great), は かなり よい (that's really great)

4. Calculate the translation model probability for all words, and store them in the translation model.

We note that the **Particle** cannot determine how reliable these candidates are, and have to calculate translation model probabilities according to the n-gram based similarity translation model probability as same as the **Content** method.

We show an example of a translation model acquired by this method in Table 3, and show the extracted paraphrases of function words. In this method, we

Table 3. A sample of the Particle translation model.

Source	Target	TM prob.
です (is)	です (is)	0.7
	だ (is: informal)	0.3
けど (but)	けど (but)	0.8
	よ (yes)	0.2
も (also)	も (also)	0.6
	で (at)	0.4
が (SUBJ)	が (SUBJ)	0.6
	は (SUBJ)	0.4

don't consider meaning of words, and we sometime get wrong paraphrases of the meaning, for example, "それ は あなた へ 。 (it for you.)" and "それ は あなた から 。 (it is from you.)". We check this problem by evaluating transforming word error rate.

3. **PPDB** is the translation model based on [Bannard and Callison-Burch, 2005]'s method for using a bilingual corpus to train. Paraphrases acquired by this method have the advantage of providing broad coverage (theoretically it is possible to cover both content and function words) and allowing for the acquiring of multi-word transformations.

Assume we have two phrases v and w in the language under consideration (in our case, Japanese), and also have a phrase-based translation model indicating the translation probabilities to and from a phrase e in a different language (in our case, English). We decide the paraphrase probability $P(w|v)$ using translation probabilities $P(w|e)$ and $P(e|v)$ by using the English phrase e as a pivot as follows:

$$P(w|v) = \sum_{e} P(w|e)P(e|v). \tag{8}$$

The translation model probability can be computed using standard methods from SMT alignment [Koehn et al., 2003]. The details of the phrase table that we used in the construction of paraphrases for this study is shown in Table 4, and the details of creation and definition is shown in Appendix A. We show an example of a paraphrasing model acquired by this method in Table 5.

4. **Characteristic word** is the translation model based on target speaker's

Table 4. The details of the phrase table.

Corpus	BILINGUAL corpus including Wikipedia, lecture, newspaper, magazine and dialogue
Words	24.2M (en) 29.6M (ja)
Phrases	67.1M
Max length	7 words
Alignment	Nile [Riesa et al., 2011]
Parsing	Kytea [Neubig et al., 2011]

Table 5. A sample of PPDB, for "翻訳 さ れ た (translated)".

Translation	TM prob
翻訳 さ れ た (translated)	0.083
に 翻訳 さ れ た (translated to)	0.034
翻訳 (translate)	0.012
共訳 (joint translation)	0.011
訳 さ れ る (was translated)	0.011
と 訳 さ れ た (was translated to)	0.002
and 20 other phrases	

unique expressions. We call these unique expressions as characteristic words, and it means a stereotype speaking style like as a character in comic, movie, or novel talks. [Teshigawara and Kinsui, 2012] raised a stereotype speaking style of a samurai as an example of these characteristic words that are reflected speaker's individuality. "Sessya (I; primarily used by samurai)" and "Gozaru (to be, to go, to come; honorific for)" are used in characteristic words to appeal own stereotype individuality.

These characteristic words are not included in existing language resources. We have to extract these characteristic words directly from target speaker's corpus, and build the paraphrase model between normal Japanese words and characteristic words. In extracting of characteristic words, we assume that these characteristic words are substituted for some frequently normal Japanese words. In the previous example, the samurai individuality converts "I", which is the frequently normal word, into "Sessya", which is the characteristic word, to express own stereotype individuality. Therefore, we have to build a specific translation model to convert these characteristic words from frequently normal Japanese words by finding characteristic words and calculating translation probability.

In previous studies of the text analysis, the method to extract characteristic words is proposed. We use this method to extract characteristic words. This study calculates the χ^2 score between target language model and source language model to extract characteristic words that are included in only target language model. This χ^2 score is the statistic of Pearson's chi-square test, and it means a ratio to contribute to a difference between these two language models. χ^2 score is calculated as follows:

$$\chi^2 = \sum_{w \in L} \frac{\Big(P_t(w) - P_g(w)\Big)^2}{P_g(w)} \tag{9}$$

We extract top 100 words in decreasing order in χ^2 score as characteristic words.

Next, we calculate paraphrase probability for these paraphrase candidates. We note that we focus on high frequent words to collect paraphrases, which are useable frequently. In collection of candidates of paraphrases, the max phrase length is 7, and the number of candidates is 30. Specifically, we calculate para-

phrase probability based on techniques to acquire synonyms from non-parallel corpora [Dagan et al., 1999, Barzilay and Lee, 2003] according to the following procedure.

1. Extract $\mathbf{v_{pp}}$ that is top 30 phrases in decreasing order in frequency from a large-scale corpus.
2. Extract $\mathbf{w_{pp}}$ that is top 30 phrases that include a characteristic word in decreasing order in frequency from a large-scale corpus
3. Calculate a Jensen-Shannon divergence with $\mathbf{w}$ and $\mathbf{v}$.
4. Marginalize Jensen-Shannon divergences with $\mathbf{w}$ as paraphrase probability.
5. Collect paraphrases that have paraphrase probability more than the threshold.

In paraphrases of characteristic words, there are no restrictions on the part of speech of function words, and these paraphrase candidates are not necessarily paraphrased. Therefore, it is necessary to calculate under the constraint stronger than the paraphrase probability defined in other translation models. In this case, we use Jensen-Shannon divergences to evaluate the mismatch of the conditional language model probabilities of phrases.

In order to calculate Jensen-Shannon divergence for phrase $\mathbf{w}$ and $\mathbf{v}$, we prepare two language models, and decide the conditional language model probability for two words $\mathbf{w}$ and $\mathbf{v}$ as follows:

$$\begin{aligned} D_{JS}(\mathbf{w}||\mathbf{v}) &= \frac{1}{2}D_{KL}(\mathbf{w}||\mathbf{v}) + \frac{1}{2}D_{KL}(\mathbf{v}||\mathbf{w}) \\ &= \sum_{x,y\in\mathcal{X},\mathcal{Y}} \Big(P_t(x,y|\mathbf{w}) - P_g(x,y|\mathbf{v})\Big) \log \frac{P_t(x,y|\mathbf{w})}{P_g(x,y|\mathbf{v})} \end{aligned} \tag{10}$$

$$P_g(x,y|\mathbf{v}) = \frac{C_g(x,\mathbf{v},y)+1}{C_g(\mathbf{v})+N_g} \tag{11}$$

$$P_t(x,y|\mathbf{w}) = \frac{C_t(x,\mathbf{w},y)+1}{C_t(\mathbf{w})+N_t} \tag{12}$$

In this formula, $C_g(\mathbf{v})$ is number of occurrences of a phrase $\mathbf{v}$ in general speaker corpus, and $C_g(x,\mathbf{v},y)$ is the number of occurrences of words x, y before and

Table 6. Sample of paraphrasing with Characteristic words

v	w	$P(\mathbf{w}\|\mathbf{v})$
。(.)	モン 。(MON .; to be)	0.108
。(.)	モン ！(MON !; to be)	0.106
。(.)	モン ☆ (MON ☆; to be)	0.076
。(.)	だ モン ！ (MON !; to be)	0.029
。(.)	た モン ！ (MON !; "was")	0.027
	計 30 語	

after a phrase $\mathbf{v}$. In a similar manner, $C_t(\mathbf{v})$ is number of occurrences of a phrase $\mathbf{v}$ in target speaker corpus, and $C_t(x, \mathbf{v}, y)$ is the number of occurrences of words x, y before and after a phrase $\mathbf{v}$. We calculate conditional language model probabilities $P_g(x, y|\mathbf{v})$, $P_t(x, y|\mathbf{w})$ where words x, y occur before and after the phrases $\mathbf{v}$ and $\mathbf{w}$. Note that N_g and N_t are the number of different words in n -gram in each corpus. $D_{JS}(\mathbf{w}||\mathbf{v})$ that calculated by substituting for Eqs. (12) and (13), is affected by conditional language model probabilities $P_g(x, y|\mathbf{v})$, and $P_t(x, y|\mathbf{w})$ as contextual words x and y. Therefore, if $D_{JS}(\mathbf{w}||\mathbf{v})$ is small, phrases $\mathbf{w}$ and $\mathbf{v}$ have similar contexts before and after, and it means phrases $\mathbf{w}and\mathbf{v}$ have high paraphrase probability.

Next, we calculate a paraphrase probability with using $D_{JS}(\mathbf{w}||\mathbf{v})$. A paraphrasing probability $P(\mathbf{w}|\mathbf{v})$ is calculated that marginalize $\mathbf{w_{pp}}$ with fixed $\mathbf{w_{pp}}$ as follows:

$$P(\mathbf{w}|\mathbf{v}) = \frac{\exp(-D_{JS}(\mathbf{w}||\mathbf{v}))}{\sum_{\mathbf{w}' \in \mathbf{w_{pp}}} \exp(-D_{JS}(\mathbf{w}'||\mathbf{v}))} \tag{13}$$

When calculation of paraphrase probabilities for all candidates is finished, we collect paraphrase pairs of $\mathbf{w}, \mathbf{v}$ if paraphrase probability $P(\mathbf{w}|\mathbf{v})$ is larger than 0.01. We show an example of a translation model acquired by this method in Table 6.

2.4 Experimental result

In order to evaluate the proposed method, we performed an evaluation focused on how well the proposed model can reproduce the individuality of a particular

speaker. In the evaluation, we target 2 kinds of speakers, which are speakers of camera sales clerks and Twitter characters.

2.4.1 Evaluation Measures

In studies of statistical machine translation, they often use automatic evaluation measures, for example BLEU. These automatic evaluation measures require a parallel corpus to evaluate. However, we mentioned in Section 2.3, collecting a parallel corpus on individuality is difficult. Therefore, we also perform a manual evaluation to evaluate correctness and individuality of the output. Specifically, we evaluate two following factors.

Individuality In order to evaluate the individuality, subjects read the training data (It is the same as the corpus, which is used to train the language model of linguistic individuality transforming) to learn the individuality of target speaker. Subjects are shown the system output and try to answer the question: "does this sentence reflect the individuality of person who wrote the training data?" subjects give a score of 1 (do not agree) to 5 (do agree).

Word Error Rate; WER This is the ratio of words in a converted sentence that are syntactically or semantically incorrect from the post-conversion sentence. This is calculated by having the subject look at the sentence before and after conversion and point out conversion mistakes.

We find the confidence interval of each evaluation measure using bootstrap resampling [Koehn, 2004] with significance level $p < 0.05$. We note that we evaluate these factors on only utterance to avoid by errors in response selection and the effect from contents of the response.

Subjects don't evaluate "entertaining or not" because subjects have each preference to speaking style, and the difference of target speaking individuality makes unfair conditions. In this evaluation, we assume that subjects get entertaining conversation if responses have the target individuality without regard to subject's preference.

Table 7. Number of utterances and words in camera sales dialogue corpus.

	Clerk	Utt.	Word
Train	A	238	11,758
	B	240	12,495
	C	228	9,039
Dev.	A	65	3,016
	B	43	2,271
	C	37	1,462
Test	A	9	173
	B	9	134
	C	9	148

Table 8. Number of sentences and words in BTEC, and REIJIRO.

Corpus	Sent.	Word
BTEC	465k	4.11M
REIJIRO	424k	8.90M
SUM	889k	13.01M

2.4.2 Targeting for speakers of camera sales clerks

As data for our research, we use a camera sales dialogue corpus [Hiraoka et al., 2014] that consists of one-on-one sales dialogues between three salesclerks and 19 customers. We split the corpus of three salesclerks into one corpus for every speaker each and further divide each of these corpora into training, development, and evaluation data. The details of the data for each of the salesclerks is shown in Table 7. All conversations were performed in Japanese by native or highly fluent Japanese speakers. As mentioned in Section 2.3.2, in order to create an language model that is both sufficiently accurate and expresses the personality of the speaker, we use multiple language models created using data from the target speaker and a larger background corpus. As our target speaker data, we use the training data from the previously described camera sales corpus. As our large background corpus, we use data from the BTEC [Takezawa et al., 2002], and the REIJIRO[3] dictionary example sentence corpus. The size of these background corpora are also shown in Table 8. We calculate the linear interpolation parameter to maximize likelihood on the development data.

We perform an evaluation over 4 combinations of translation models for conversion of individuality as shown in Table 9. We compare the four methods for

[3] `http://eijiro.jp`

Table 9. Translation models and paraphrasing targets

Methods	Target
Source	–
Content	Content words
Particle	Function words
Content + Particle	Combination of Content and Function words
PPDB	Paraphrasing database

Table 10. An example of transforming for speakers of camera sales clerks

Methods	Transformed result (Underlines are transformed words)
Source	ま 値段 的 に ね た ぶん 希望 として は たぶん B あたり や と 思う ん です Ah, I think that perhaps B is your hoped one perhaps in this price range.
Particle	そうですね 値段 的 に ね た ぶん 希望 として は たぶん B あたり や と 思う ん です Yes, I think that perhaps B is your hoped one perhaps in this price range.
Content	ま 値段 的 に ね た ぶん 人間 として は たぶん B パリ や と 思う ん です Ah, I think that perhaps B is your hoped one paris in human.
PPDB	ま 値段 的 に ね , これ を 希望 として は たぶん B あたり や と 思う ん です Ah, I think in this case, B is your hoped one perhaps, in this price range, isn't it?
Source	ちょっと 今 ね A と B と 比較 見 てる ん です けども そうですね Just now, I am comparing A and B, so,
Particle	ちょっと 今 ね A と B と 比較 見 てる ん です けど あの Just now, I am comparing A and B, but ah-
Content	ちょっと 今 ね A と B と 間 見 てる ん です けども そうですね Just now, I am comparing between A and B, so,
PPDB	ちょっと 今 ね A と B と 比較 を 見 てる ん です けども そうですね Just now, I am comparing on A and B, so,

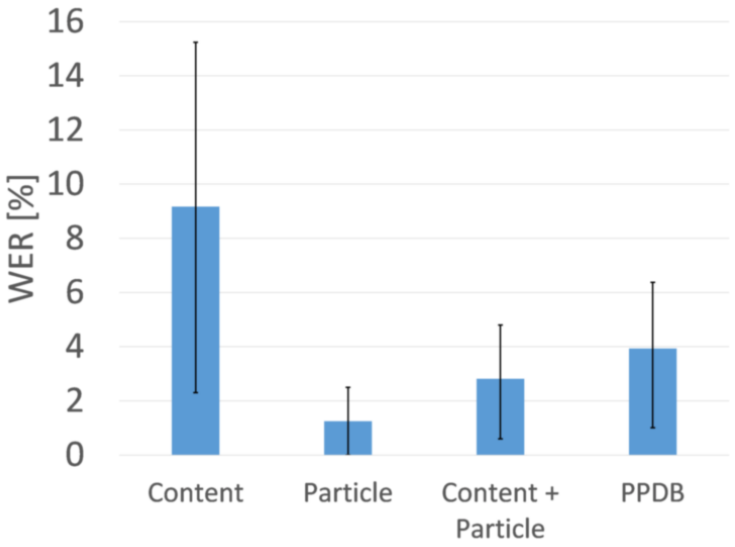

Figure 8. WER of transforming for speakers of camera sales clerks

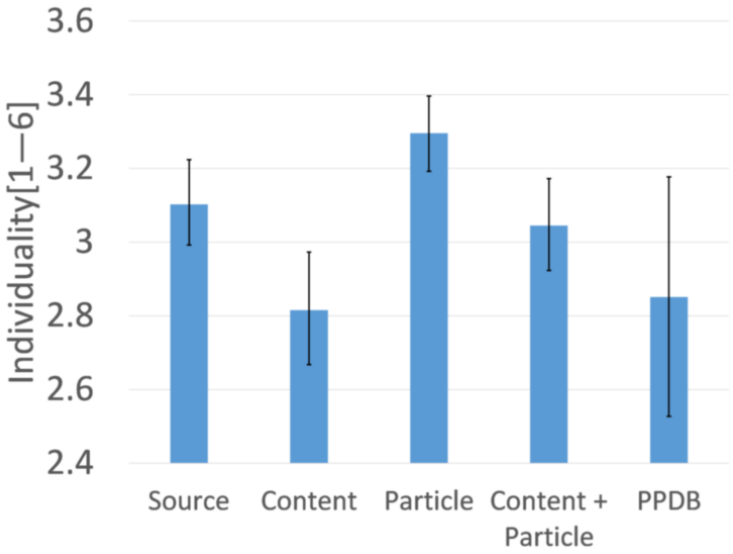

Figure 9. Individuality score of transforming for speakers of camera sales clerks

constructing the translation model using the CONTENT, PARTICLE, combination of CONTENT and PARTICLE, and PPDB. We also compare with a baseline method that does not perform any conversion at all (SOURCE).

In the experimental evaluation, we first have subjects read the training data of the target speaker. Next, we prepare an input sentence that is selected randomly from other salesclerks. Based on this input sentence, we use the three methods described in the previous paragraph to convert it into the target speaker's individuality. The subject reads these three results. The subject estimates WER and individuality for each of these four conversion results according to the measures described in Section 2.4.1. In this evaluation, three subjects evaluate result for 3 speakers, each with 9 utterances, 27 conversion results in total.

We show the results of manual evaluation of WER in Figure 8, and individuality in Figure 9. The first result to be noted is that SOURCE is the middle individuality score of 3.1. As a cause of this, these staffs who are target speakers consistently use honorific expression, and they speak with similar speaking style. However, transformation using PARTICLE is able to raise the individuality to 3.3 from the SOURCE of 3.1, a significant difference. From Table 10, we can obtain importance of non-content words to express target speaker's individuality. In

Table 11. Number of utterances and words in the character corpus.

	Character	Utt.	Word
Train	A	880	17.4k
	B	276	5.5k
	C	288	3.2k
Dev.	A	220	3. 2k
	B	69	1.1k
	C	72	0.8k

addition, the order of the WER and the order of the individuality score are in agreement, and the individuality score of the model with the high WER is low.

In this case of using PARTICLE, the individuality is significantly improved to 3.3, however it is not enough to high. To analyze this cause, we calculate the ratio of words that can transform in each method. Using Particle, which transforms filler, particles, and exclamations, has 19% of transformable words. In other hands, using CONTENT, which transforms content words, has 34% of transformable words, and using PPDB has 80% of transformable words. From these, using particles is able to convert function words that influence linguistic individuality with fewer errors, and improves individuality. However transformable words are few, and the individuality score due to the final conversion did not exceed 4.

2.4.3 Targeting for speakers of Twitter characters

In this evaluation, we target for 3 characters who are active in Twitter to public relations. It is because these characters have the stereotype speaking style that depends on the character's motive, and we can collect these speaker's utterances easily. These target speaker's corpora are collected by their monologue tweets without URLs, Retweets, Mentions, and Hashtags. We show the details of Twitter character's corpora in Table 11. We built language models as mentioned in Section 2.3.2.

The subject estimates WER and individuality for each of these four conversion results according to the measures described in Section 2.4.1. In this evaluation, three subjects evaluate result for 3 speakers, each with 10 utterances, 30 conver-

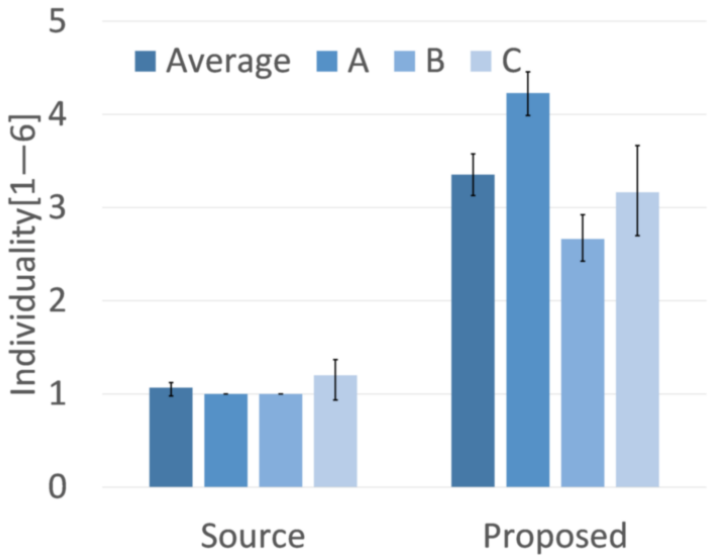

Figure 10. Individuality score of transforming for speakers of Twitter characters

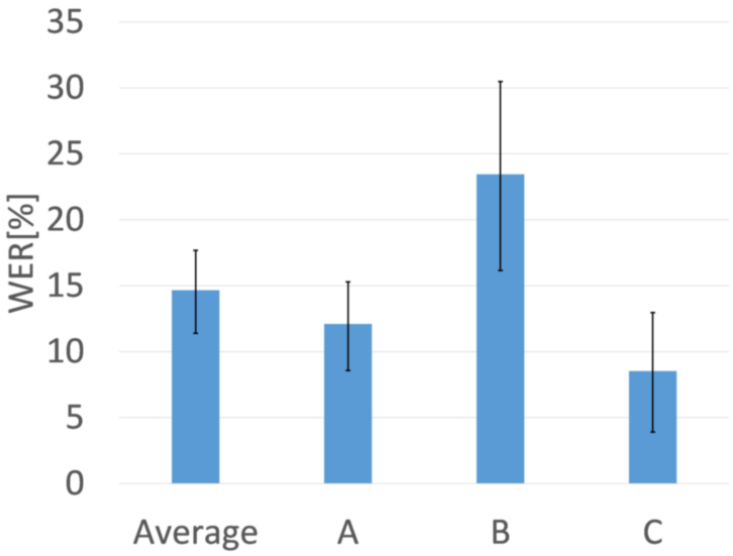

Figure 11. WER of transforming for speakers of Twitter characters

Table 12. An example of transforming for speakers of Twitter characters

Speaker	Transformed result (Underlines are transformed words)
Source	無駄 な 時間 を 費やし たく も あり ませ ん 。
	I do not want to spend wasted time.
A	無駄 な 時間 を 費やし たく も あり ませ ん だ モン ！
	I do MON not want to spend wasted time !
B	無駄 な 時間 を 費やし たく も あり ませ ん ゜ ▽ ゜) ノ
	I do not want to spend wasted time ゜ ▽ ゜) ノ.
C	無駄 な 時間 を 費やし たく も あり ませ ん ッ ！
	I do not want to do spend wasted time.
Source	ウィンドウ の 中 に ある の が 欲しい の です が 。
	I would like to get one that is in this show case.
A	ウィンドウ の 中 に ある の が 欲しい の くま ー が だ モン ！
	I would kuma- like to get MON one that is in this show case !
B	ウィンドウ の 中 に ある の が 欲しい の ゜ ▽ ゜) ノ が ゜ ▽ ゜) ノ
	I want to get one that is in this show case ゜ ▽ ゜) ノ ゜ ▽ ゜) ノ.
C	ウィンドウ の 中 に ある の が 欲しい の だ が 。
	I'd like to get one that is in this show case.

sion results in total. We show the results of manual evaluation of WER in Figure 11, and individuality in Figure 10.

From Figure 10, transformation is able to raise the individuality to 3.4 from the SOURCE of 1.0, a significant difference. We obtain a difference between target speakers for improvement of the individuality, and it means the effect of the transformation is different depending on the individuality of target speaker.

From Figure 11, an average WER of proposed method is 15%. The reason why the WER is higher than the result of targeting for speakers of camera sales clerks, is the weakness of the constraint in extracting paraphrases of characteristic words. In paraphrasing of characteristic words, extraction method considers only words before and after phrases, therefore it cannot consider part of speech or agreement with a wide context. This weakness of the constraint causes high WER in transformation. We may solve this weakness by calculating similarity considering longer n-gram and considering POS in characteristic words. Decreasing WER is one of the most important future work in the linguistic individuality transforming.

From Table 12, proposed method transforms end particles mainly. It is well

known that function words that include end particles, affects individuality in linguistics [Teshigawara and Kinsui, 2012, Chung and Pennebaker, 2007], and we well known that function words affect individuality from the result of the targeting for speakers of camera sales clerks.

2.5 Summary

In this chapter, we proposed a method for transforming individuality, and proposed techniques to train language models and paraphrase models using few target speaker's corpus and large general speaker's corpora to use for the transforming method.

In experimental evaluation, we assumed subjective evaluation in 2 speaker groups. In speakers of camera sales clerk, the proposed method by paraphrasing of particles improved the individuality significantly. In speakers of Twitter character, the proposed method by paraphrasing of characteristic words improved the individuality significantly. These results are consistent with linguistic findings. Therefore, we analyzed the relationship between the WER, the ratio of transformable words, and the individuality score, and showed future tasks.

Chapter 3
Satisfaction prediction for example database

3.1 Introduction

In this chapter, we describe a satisfaction prediction method for an example database, it achieves to construct the high satisfactory example database for the EBDM framework. The proposed method makes it possible to use a large amount of example that have no annotations or evaluations by using a small amount of example that are annotated. Specifically, this satisfaction prediction model takes as input user utterance q, agent response r and some external linguistic resources such as lexicons, and learns a function $s(q, r)$ to predict user satisfaction. Because this function can be calculated using only the dialogue example $\langle q, r \rangle$, and doesn't rely on any information about the surrounding dialogue context, it is appropriate for applications such as an example database construction, where the dialogue context is not available at the time the database is constructed. It can also be easily incorporated into response selection by predicting the goodness of a response before presenting it to the user. In this thesis, we assume the average of users as the general user; however, we can assume the specific user and train this satisfaction prediction model for the specific user if we have enough training data.

In an experimental evaluation using two diverse corpora in two languages, we show that the proposed prediction model is able to reduce the error between predicted satisfaction and annotated satisfaction. We also apply the proposed model to EBDM example selection and find that it is effective, improving a satisfaction score evaluation from 4.04 to 4.26 on a scale of 1–6.

3.2 Construction of example database

We need example databases to train and to test out proposed satisfaction prediction method. In this section, we construct two kinds of example databases to

use. We show that the proposed method is effective with both of Japanese and English, and is not dependent on specific data.

The first corpus is a manually constructed example database covering everyday conversations. Given 14 events that occur in daily life, we had 7 human creators create user utterances related to each event. To create system responses for these utterances, we asked 15 human response creators to fill in blanks following every user utterance, finally obtaining an average of 12 unique responses for each user utterance. It should be noted that each query q in this database has multiple responses $\mathbf{r}$. The construction of examples with multi writers and events is a contrivance to inhibit the over fitting in a small amount of example and to give diversity or variety. This contrivance makes it possible to prevent that the proposed method estimates a quality of examples from specific few features, which are not related to quality ordinarily.

The aim of the proposed method is to predict user satisfaction for system responses, and thus the next step in our data collection is to collect a corpus that includes annotated satisfactions for each response. To annotate satisfaction, we must first have a definition of satisfaction. In the well-known PARADISE framework [Walker et al., 1997, Hajdinjak and Mihelič, 2006] for task-based dialogue, satisfaction is calculated by asking the user several subjective questions after the dialogue completes, and averaging the scores for each question into a total satisfaction score. These questions are related to task success, response delay, response quality, and other topics, with a heavy weight on task success. However, in the case of non-task-oriented dialogue, as handled in this section, these questions cannot be applied directly. Therefore, following Yang et al. [Yang et al., 2010], we judge overall satisfaction with responses with a single question "Do you think that this is a satisfactory response?", and have the user reply to this single question on a 1–6 scale. We then had 5 users annotate the collected corpus with satisfaction scores for each response, resulting in 2,555 user utterance/response pairs. We show an example of this annotated data in Table 16.

To analyze the relationship between examples and annotators, we show the pairwise correlations between annotators in Figure 12. The inter-annotator correlation generally is in the range of 0.3–0.5 (except for annotator 3, who is somewhat of an outlier), demonstrating that while the trends are the same, there is still a

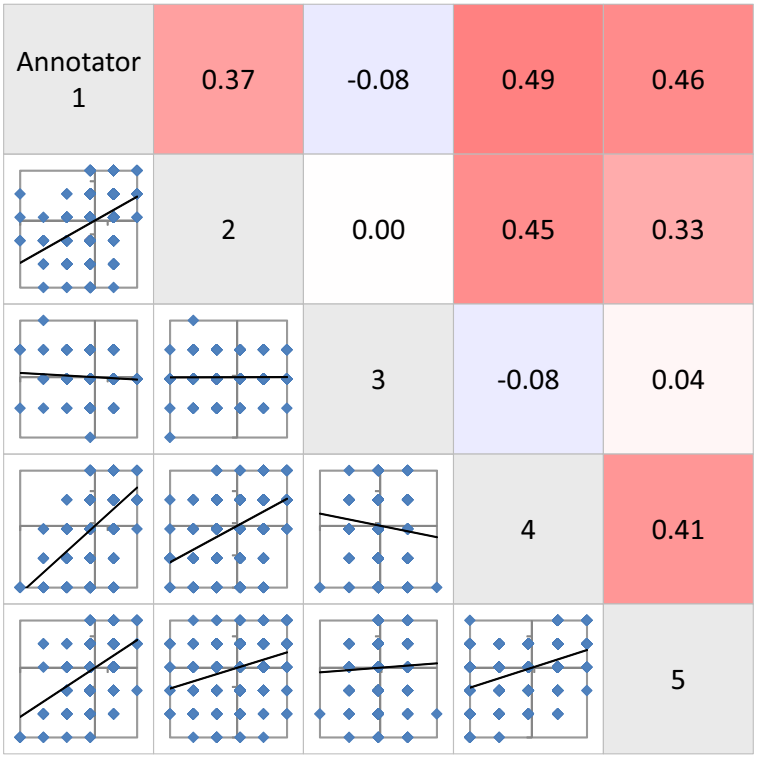

Figure 12. Correlation between annotators

significant amount of subjectivity in rating satisfaction with dialogue responses.

The second corpus is an English utterance-response pair corpus constructed by collecting utterance-response pairs from a chat-oriented EBDM system trained using examples from movie or drama scripts [Nio et al., 2014a]. This corpus has 791 examples annotated with subjective scores for each example. Subjective scores were annotated on a scale of 1–5 regarding the naturalness of each example, but there is no information of which annotators annotated which utterance. In addition, in some cases, multiple annotators annotated single examples, and in this case, we take the average score as the true score. We show an example of this annotated data in Table 15.

3.3 Proposed method

In this section, we describe about the proposed satisfaction prediction method. This satisfaction prediction method is same as learning a function $s(q, r)$ that

Table 13. Examples of events and pairs of utterance and responses (translated from Japanese)

Events	Utterances	Responses
Eating dinner	What shall we eat today? 今日は何食べようか?	Today is cold, let's make stew. 今日は寒いので、シチューはどうですか.
		Anything to eat is good for me. なんでもいいです.
		I know you're going to make me cook dinner. どうせ私に作らせるんでしょ
	It looks delicious. おいしそうだね	How about the taste? 味はどうですか？
		Thank you! ありがとうございます！
	Let's eat. いただきます	Please eat a lot. たくさん召し上がれ
		Sure, let's eat. いただきます
Drinking	Let's have some liquor. お酒飲もうかな	What will you have to drink? 何を飲みますか?
		Me too. 私も.
		Try to drink in moderation. 飲み過ぎには気をつけて
	Beer! まずはビールでしょ！	Right! いーねー

Table 14. Examples of utterance-response pairs and annotations (translated from Japanese)

Utterance	Response	Annotations
What shall we eat today?	Today is cold, let's make stew.	5 6 4 5 6
What shall we eat today?	Be careful not to eat too much!	4 4 4 2 4
How about you?	I'm fine.	4 3 4 2 4
What time is it now?	I don't know.	1 3 5 2 1

Table 15. Examples of pairs of utterance and response

User query	System Response	Score
Ah, I got it.	Got what?	5.0
Where are you going?	I don't know. But I'm going on foot.	5.0
You're a monster.	No – I 'm a survivor.	3.5
So what do you make of it?	I already told you what I think.	1.0

predicts user satisfaction from only information of examples. This can be defined as a regression problem from input features derived from $\langle q, r \rangle$ to a real-valued satisfaction score assigned to the example through manual evaluation.

To achieve this regression, we train regression function $s(q, r)$ with Support Vector Regression (SVR) [Basak et al., 2007], which has previously seen success in dialogue quality estimation [Schmitt et al., 2011]. In order to train an SVR model, it is necessary to define features over the dialogue example $\langle q, r \rangle$, and these features are detailed in this section.

Note that in this prediction model, satisfaction is predicted only from query q and response r, independent from other information such as dialogue context and previous user states. In itself, dialogue is a consecutive string of utterances, but by creating a context-independent estimator of satisfaction, it allows for simple integration into other applications such as database construction and response selection, as described in Section 3.2. Thus, the proposed prediction model tries to predict potential satisfaction directly from information about the example itself, and thus is essentially different from models that predict the trajectory of user satisfaction in an already-completed dialogue [Ultes and Minker, 2014, Schmitt et al., 2011, Higashinaka et al., 2010, Engelbrech et al., 2009].

In the setting of predicting regression function $s(q, r)$, we must use features that can be derived solely from q and r. In this section, we use occurrences of words, word classes defined by WordNet [Bond et al., 2009, Bird et al., 2008], and sentiment orientation scores from a sentiment lexicon [Takamura et al., 2005] as features for this prediction model. Specifically, we define these features as follows:

- Counts of n-grams in example query q and system response r.
- Counts of word classes in example query q and system response r.
- Counts of word pairs co-occurring in example query q and system response r.
- Flags of whether a word in the sentiment lexicon exists in example query q or not, and in system response r or not.
- Maximum, minimum and average of sentiment scores in example query q, and in system response r.

- Flags of who annotated user satisfaction.

Here, the word n-gram features allow the classifier to flexibly learn expressions that affect user satisfaction, and word classes allow these features to generalize. The co-occurrence word pair features express relationships between words in the utterance and response. The sentiment lexicon features intuitively capture information such as "utterances including negative words cause the user to feel negative." The annotator features help to capture the tendency of likes and dislikes for each annotator.

When using the prediction model for response selection in a dialogue system, we can also obtain the user query q' and define features over it. These features include the above features with q' replacing q, as well as the following feature:

- Similarity scores between example query q and user query q'.

The similarity feature expresses the reliability of the match between the two queries. In this section, we use similarity scores of TF-IDF weighted vector space similarity [Banchs and Li, 2012], WordNet-based syntactic-semantic similarity [Nio et al., 2012], or recursive neural network-based paraphrase detection [Nio et al., 2014b].

In construction of example databases, the proposed satisfaction prediction model can be used to filter examples that may result in low user satisfaction. Previous research about example database construction in EBDM is based on harvesting examples from a corpus, and using rules or heuristics to filter examples that are obviously bad. Specifically, it is possible to gather only utterances in which a user is explicitly responding to another user, making it possible to easily gather a relatively clean example database [Bessho et al., 2012]. Furthermore, Nio et al. proposed heuristic rules that use only utterance/response pairs that are performed by 2 speakers in 3 consecutive turns, which helps avoid noisy examples due to switches of topic or scene in a movie/drama corpus [Nio et al., 2012]. While these rules and heuristics guarantee some level of naturalness in the dialogue examples, they do not consider user satisfaction directly.

In contrast, the proposed model can be used to directly filter examples with low predicted user satisfaction. The simplest method is to gather examples that

predicted satisfaction score is better than threshold t for the new database $\mathbf{e}'$:

$$\mathbf{e}' = \{\langle q, r\rangle \in \mathbf{e} | s(q, r) > t\}. \tag{14}$$

3.4 Experimental result

We evaluated the proposed model from two viewpoints: accuracy of satisfaction prediction, and effectiveness of response selection.

For evaluation, we normalized the satisfaction score to have a mean of 0 and variance of 1. Normalization was done for every annotator for the multi-response corpus, and for the whole corpus for the movie/drama corpus, which don't have extensive annotator data. In each evaluation measure, error bars represent 95% confidence intervals according to bootstrap resampling [Koehn, 2004].

3.4.1 Accuracy of Satisfaction Prediction

For the accuracy of satisfaction prediction, we measured the Mean Squared Error (MSE) of predicted satisfaction for each example according to 50-fold cross validation. We compared with a baseline that always chooses the average satisfaction.

The results for the multi-response corpus in Figure 13 show that the proposed prediction model decreased prediction error significantly ($p < 0.05$) to 0.90 from 1.00 of the baseline. Looking at the individual annotators, we can see that all but annotator 3 saw an increase in prediction accuracy. The lack of a gain for annotator 3 can be explained by the lack of correlation with other annotators shown in Figure 12.

The results for the chat-oriented dialogue corpus in Figure 14 are similar, a significant ($p < 0.05$) decrease of MSE to 0.96 from 1.00 of the baseline. In this case, the proposed prediction model was able to achieve better predictions than the baseline in the majority of cases (55.7%), and the rate of large prediction errors over 1.0 also decreased to 39.1% from 42.1% of the baseline. This shows that the proposed model was consistently more accurate, and helped to reduce the number of examples, which have large errors, resulting in a decrease in overall prediction error.

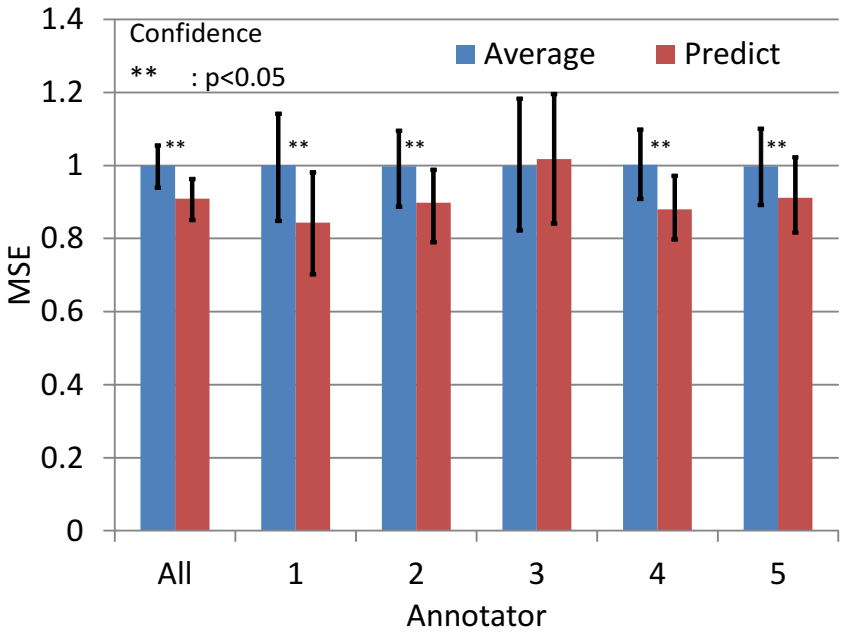

Figure 13. Evaluation for satisfaction prediction on the multi-response corpus

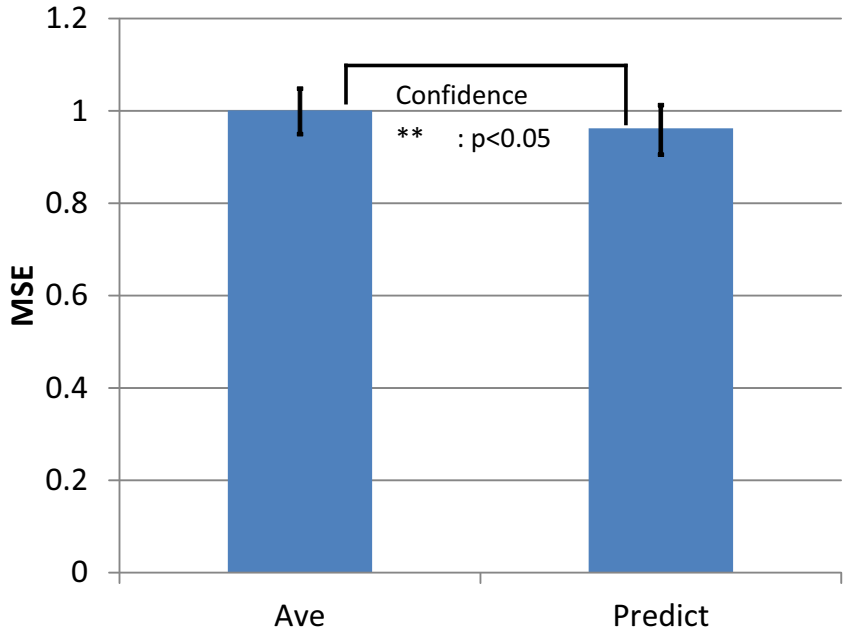

Figure 14. Evaluation for score prediction on the chat-oriented dialogue corpus

In this section, we evaluate whether the proposed model can be used as a criterion for response selection. Specifically, using the multi-response corpus, we used the proposed model to estimate satisfaction for each response in response set r for query q and return the response $\hat{r}$ that has the highest estimated satisfaction[4]. We evaluated the satisfaction of response selection by referring to the annotated satisfaction score. Evaluation was performed with 50-fold cross validation, comparing to the same random response selection baseline from the previous section.

Figure 15 shows the improvement of user satisfaction of selected responses using the proposed model. User satisfaction improved significantly ($p < 0.05$) to 4.26 from 4.04 of the baseline. We can also note that the results in this figure and Figure 13 are quite similar, including the difficulties with annotator 3, indicating that the success of the response selection is closely related to success of the satisfaction prediction. Overall 31.9% of selected responses of the proposed model had a better score than the baseline, 49.7% were the same, and 18.4% were worse, indicating that the proposed model chooses better or similar responses in majority of cases. There was also a small decrease in the number of unpleasant responses with satisfaction scores of 3.5 or lower, decreasing to 15.7% from 21.6% of the baseline.

3.5 Summary

In this chapter, we proposed a model to predict user satisfaction with dialogue examples for example-based dialogue systems. An evaluation showed that the proposed model was effective for both satisfaction prediction and response selection.

Also, while the features in the model used here were only based on the user query and dialogue example, it is notable that our framework is easily extensible to use other features. For example, we could use features of the user such as sex or age, or other salient features about the environment such as the time of day or location.

While the experimental results showed that the prediction model is able to suc-

[4]We only evaluate the response selection with the multi-response example database, because the chat-oriented dialogue corpus does not have multiple responses $\mathbf{r}$ for one utterance q.

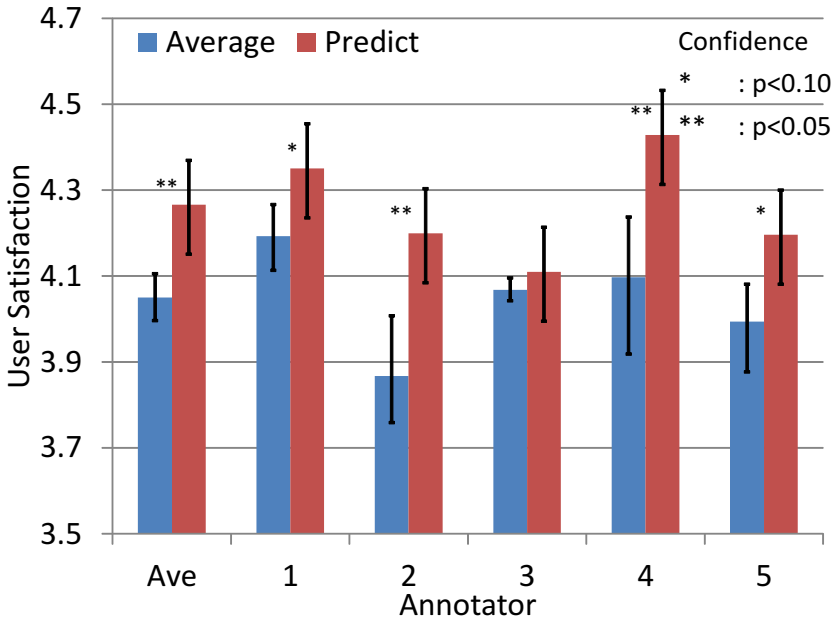

Figure 15. User satisfaction when the proposed model is used to select responses

cessfully predict expected user satisfaction, there are still some future challenges related to refining the prediction model. The main potential for improvement lies in the creation of better features (such as the original source of the dialogue example, etc.), and the creation of training labels without requiring the explicit annotation of satisfaction scores. For example, if we could predict user satisfaction by the user's reaction to system responses, it may be possible to annotate system responses with satisfactions automatically and implicitly. Furthermore, the proposed model has an advantage that also works with small data, and the advantage makes it possible to develop a high-quality conversational agent by a few resources. We also plan to utilize the proposed model for example selection in real-time, during dialogue with an actual example-based dialogue system.

Chapter 4

Adaptive response selection

4.1 Introduction

In this chapter, we describe the adaptive response selection module that considers user preference and user feedback. The satisfaction prediction for examples selects an appropriate response from plural appropriate response candidates that may satisfy users. This method tries to increase user satisfaction by using a large amount of annotated examples. Both methods could select satisfactory responses if we used enough training data that annotated by the target user in question. The difference of these methods is adaptivity. The satisfaction prediction for examples assumes the specific user (or the typical/averaged user) in all of the dialogue; however, this method can adapt to the user who is not in training data. We can select a method depends on a condition of data to improve user satisfaction.

This proposed module has a multi-response example database, and selects an appropriate response based on collaborative filtering. We evaluate two points of this proposed method, one is the accuracy of satisfaction prediction for user feedback, and one is the effectiveness of response selection.

4.2 Construction of feedback corpus

As mentioned in chapter of conversational agent, databases are generally constructed from available data sources such as human-to-human conversation log databases [Murao et al., 2003], movie or television scripts [Banchs, 2012, Nio et al., 2012], or Twitter logs [Bessho et al., 2012]. There are many methods to construct large and high-quality example databases for EBDM. However, it does not try to construct the multiple response example database.

In this chapter, we use multi-response example database, that constructed in Section 4.2. This improvement of EBDM affect not only research but also engineering. It is because proposed multi-response example database makes it possible to solve the problem of monotonicity on EBDM. The problem of mono-

tonicity on EBDM is caused by poor pairing with a response to an utterance on the example database. It does not allow a conversational agent to respond various responses to a specific user utterance. However, the proposed architecture has a multi-response example database and selects a variety of response to a specific user utterance.

The aim of the proposed method is to find a response to a user utterance from the multi-response example database that maximize user satisfaction, and thus the next step in our data collection is to collect responses annotated with satisfactions, as well as annotator feedback utterances. Following Yang et al. [Yang et al., 2010], we judge overall satisfaction with responses with a single question "Do you think that this is a satisfactory response?", and have the user reply to this single question on a 1–6 Likert scale, where 1 is "I don't think so" and 6 is "I think so."

To collect this data, the annotator first views the reply of the dialogue system for each input and decides a *satisfaction score* on a scale of 1–6. The user then can make a *feedback utterance*, which is a verbal expression of their satisfaction with the system's response. We show an example of inputs, system responses, feedback utterances, and annotated satisfactions in Table 16, and we explain these in detail in the following paragraphs.

In most cases it is an unreasonable burden on the user to annotate explicit satisfaction scores while the dialogue progresses. On the other hand, in many cases the user may provide implicit feedback regarding the goodness of the response. For example, when the dialogue system makes a funny joke, the user may laugh or praise the system. These user feedback utterances express the user's opinion or feeling about the response, and it is useful to estimate satisfaction at test time using these utterances, removing the need for explicit annotation. Thus, we propose to predict user satisfaction score directly by analyzing this feedback. The satisfaction prediction method is used with the proposed adaptive method (Section 4.3.2), which requires knowledge of satisfaction of the actual user. To create the data for the satisfaction prediction method, annotators are told to perform a feedback utterance as an option (non-essential). When there is no annotator feedback, it is treated as an instance of "null" feedback.

For convenience, we define a triplet of user utterance, response utterance,

Table 16. A sample of tri-turns and annotation results (translated from Japanese)

User utterance	System response	User feedback	score
Do I have any plans today?	Please check your calendar.	No, you tell me!	1
今日は何かあったっけ?	カレンダーを確認してください	教えてよ!	
Please be quiet.	Umm... I'm sorry...	"null"	4
静かにして.	あ... ごめんなさい...	"null"	
What shall we eat today?	Today is cold, let's make stew.	Nice idea!	6
今日は何食べようか?	今日は寒いので、シチューはどうですか.	いいね!	

and annotator feedback utterance as a "tri-turn" [Nio et al., 2012]. In the end, we collected satisfaction annotated tri-turns from 5 annotators for 15 example databases corresponding to each response writer. The corpus totals 2,555 tri-turns including 2,056 non-null annotator feedbacks. We normalize satisfaction by Z-score[5] for each annotator for the purpose of reducing differences between annotators. In the annotation, annotators viewed response and gave feedback by using a text-based chat interface, using each database separately.

4.3 Adaptive satisfaction prediction based on using collaborative filtering and user feedbacks

4.3.1 Satisfaction prediction for user feedbacks

In this section, we describe a proposed method for predicting a user satisfaction score using a user feedback utterance. Some previous studies tried to predict user satisfaction using n-gram-based dialog history [Hara et al., 2010], collaborative filtering [Yang et al., 2010], or analyzing "competence" and "certainty" [Engelbrecht and Möller, 2010]. However, these studies predicted the satisfaction in batch processing after each dialogue. In contrast, our method predicts turn-by-turn while the dialogue is progressing for use in response selection.

We predict user satisfaction in each tri-turn using Support Vector Regression [Basak et al., 2007], which has proven effective in previous study on dialogue quality estimation [Schmitt et al., 2011]. For the t-th tri-turn in the training data, we have a labeled satisfaction score s_t which is to be estimated by a regression model $\mathrm{R}(m_t)$ given the user feedback utterance m_t as input. As input variables of the regression, we use occurrences of words, word classes defined by Japanese Word Net [Bond et al., 2009], and sentiment orientation scores calculated by

[5]Z-score is a method that normalizes score so $\mu = 0$, $\sigma^2 = 1$.

a sentiment lexicon [Takamura et al., 2005]. Specifically, we use the following features:

- Flag about whether user feedback m_t exists or not.
  $f_{m_t} \in \{0, 1\}$
- Counts of n-grams in user feedback m_t.
  $\mathbf{w_t} = \{w_{t,1}, w_{t,2}, \ldots, w_{t,N}\}$
- Counts of word classes in user feedback m_t.
  $\mathbf{c_t} = \{c_{t,1}, c_{t,2}, \ldots, c_{t,M}\}$
- Flag about whether a word in the sentiment lexicon s_t exists in user feedback m_t or not.
  $f_{s_t} \in \{0, 1\}$
- Vector containing maximum, smallest and average of sentiment scores for user feedback m_t.
  $\mathbf{s_t} = \{s_{t,max}, s_{t,min}, s_{t,ave}\}$

Here, the word n-gram features allow the classifier to flexibly learn expressions that represent user satisfaction, and word classes allow these features to generalize. The sentiment lexicon features intuitively capture information such as "utterances including sentimentally charged words express positive or negative opinions about the previous utterance."

Based on these features, we construct the user satisfaction prediction model with Support Vector Regression (SVR) [Basak et al., 2007], which has previously seen success in dialogue quality estimation [Schmitt et al., 2011].

4.3.2 Satisfaction prediction by using collaborative filtering

First, we describe the baseline on response selection method. In our actual data, we have multiple responses **r** for each query, so we create two baselines to simulate how standard EBDM systems would act in this situation. The first, RANDOM, randomly chooses from the potential responses **r**, simulating a situation where we don't consider quality of responses at all. The second baseline, MAXDB notes that in Section 3.2, we have 15 different writers who create example bases, and

selects the single *writer* that achieves the highest satisfaction. This simulates a situation where we can collect a high quality single example base from a skilled writer.

Next, we describe two proposed methods for selection from multiples responses in EBDM. Both methods select the query q that has the highest similarity to user utterance q', and obtain its corresponding response set $\mathbf{r}$ from multi-response example database $\mathbf{e_{multi}}$:

$$\langle \hat{q}, \hat{\mathbf{r}} \rangle = \underset{\langle q, \mathbf{r} \rangle \in \mathbf{e_{multi}}}{\operatorname{argmax}} \ \operatorname{sim}(q', q). \tag{15}$$

Next, we select a response r that has the highest expected satisfaction $\mathrm{C}(q, r)$ in response utterance candidates $\mathbf{r}$:

$$\langle \hat{q}, \hat{r} \rangle = \underset{\langle q, r \rangle \in \langle \hat{q}, \hat{\mathbf{r}} \rangle}{\operatorname{argmax}} \mathrm{C}(q, r). \tag{16}$$

We detail methods to calculate expected satisfaction $\mathrm{C}(q, r)$ as follows:

Our first scoring method is entitled MAXR, for "maximum response," MAXR chooses the response that has the highest average evaluation score by human annotators. This method is similar to MAXDB, but instead of having a single skilled writer create an example base, we have multiple writers create examples, and select the best example for each particular query.

Every pair of query q and response r has several scores annotated by different annotator, thus, we calculate the average satisfaction $\overline{s}_{\langle q,r \rangle}$ from the annotated satisfaction score $s_{u,\langle q,r \rangle}$ of each annotator $u \in U$:

$$\overline{s}_{\langle q,r \rangle} = \frac{1}{|U|} \sum_{u \in U} s_{u,\langle q,r \rangle}. \tag{17}$$

We then define $\mathrm{C}_{\mathrm{maxr}}(q, r) = \overline{s}_{\langle q,r \rangle}$ for the estimated satisfaction in Equation (16). While it considers multiple response candidate, the selected response is static. It is invariant throughout the dialogue, and not tailored to a specific user.

The other method, named ADAPTIVE, is an adaptive method that uses the satisfaction prediction explained in Section 4.3.1, and collaborative filtering to adapt the response utterance to the user based on annotators who has similar

preference.

Collaborative filtering is a technique widely used in recommendation systems to fill in estimates of user preference based on the preferences of other similar annotators. In spoken dialogue systems, collaborative filtering has been used to model user utterances or user satisfaction [Yang et al., 2010, Higashinaka et al., 2009]. However, these previous studies use collaborative filtering only to evaluate the performance of the dialogue system or to predict user utterances. In contrast, we use collaborative filtering to estimate user preference to select the certain response for user.

We calculate expected satisfaction for the next system utterance based on predicted user satisfaction of the previous utterances. We do this by using collaborative filtering to compare the current user's predicted satisfaction with previous utterances with the tendencies of each annotator in the training data. Specifically, we estimate satisfaction data $\mathbf{s}_{est} = \left\{ s_{est,1}, \ldots, s_{est,|\mathbf{L_e}|} \right\}$ where each value represents the current user's satisfaction with a particular dialogue response in the list $\mathbf{L_e} = \left\{ \langle q_1, r_{1,1} \rangle, \langle q_1, r_{1,2} \rangle, \ldots \langle q_v, r_{v,w_v} \rangle \right\}$ that enumerates all the query-response pairs in example database $\mathbf{e}$. At first, these are filled by 0, which is the middle of the range of the normalized satisfaction score. Whenever a tri-turn passes, and the user makes a feedback utterance m, the system uses the satisfaction prediction model $\mathrm{R}(m)$ of Section 4.3.1 to predict the user's satisfaction to the system response. In the t-th tri-turn, user satisfaction data $\mathbf{s}_{est,t} = \left\{ s_{est,1}, \ldots, s_{est,|\mathbf{L_e}|} \right\}$ and user utterance q' are given, and the system selects as a response the n-th example in L_e, and finally the user replies a feedback utterance m_t. The system then estimates the user satisfaction for the example using the satisfaction prediction model $\mathrm{R}(m_t)$, and updates the n-the element of the user satisfaction data for the next $(t+1)$-th tri-turn:

$$\mathbf{s}_{est,(t+1)} = \left\{ s_{est,1}, \ldots, s_{est,n-1}, \mathrm{R}(m_t), s_{est,n+1}, \ldots, s_{est,|\mathbf{L_e}|} \right\} \tag{18}$$

The value of s_{est} corresponding to this system response is then updated to be equal to this predicted value.

Once the s_{est} calculated, the system compares the current user's predicted satisfaction with each response $\mathbf{s}_{est}$ and annotated data $\mathbf{s}_u = \left\{ s_{u,1}, \ldots, s_{u,|\mathbf{L_e}|} \right\}$ for each annotator $u \in U$ who participated in the satisfaction annotation described

in Section 4.2. Finally, the system estimates the satisfaction of each response by multiplying the cosine similarity between $\mathbf{s}_{est}$ and $\mathbf{s}_u$ with the annotator's satisfaction with the response $s_{u,\langle q,r\rangle}$ where $u \in U$, $r \in \mathbf{r}$ and the average satisfaction of all users is $\overline{s}_{\langle q,r\rangle}$:

$$\mathrm{C}_{\mathrm{adapt}}(q,r) = \overline{s}_{\langle q,r\rangle} + \sum_{u \in U} (s_{u,\langle q,r\rangle} - \overline{s}_{\langle q,r\rangle}) \cos(\mathbf{s}_{est}, \mathbf{s}_u). \tag{19}$$

In this formula, we regard the cosine similarity between the two satisfaction vectors s_{est} and s_u as the reliability that the present user is similar to an annotator u in the training data.

These proposed methods of response selection for the multi-response example database are an extension of EBDM, and does not inhibit a fundamental process of EBDM. It means that proposed methods secure the quality and merits that are obtained by working the fundamental process of EBDM as lower bounds. Furthermore, proposed models to calculate expected satisfaction help the response selection module to select a more appropriate response. Our proposed model, which tries to track user states for calculating an expected user satisfaction, may become the fundamental study to develop a dialogue management module like POMDP on the architecture of EBDM.

4.4 Experimental result

We evaluated the proposed method from two viewpoints: accuracy of satisfaction prediction, and effectiveness of response selection.

4.4.1 Evaluation for Predicting Satisfaction

In the evaluation for satisfaction prediction, we measured the Mean Squared Error (MSE) of predicted satisfaction for each tri-turn using 10-fold cross validation. We also show a baseline that always chooses the average satisfaction. We calculated the confidence interval of each evaluation measure using bootstrap resampling [Koehn, 2004] with significance level $p < 0.05$. In Figure 16, we show the accuracy of satisfaction prediction. From this result, we can see that when we used the proposed prediction model, error decreased significantly to 0.69 compared to 1.00

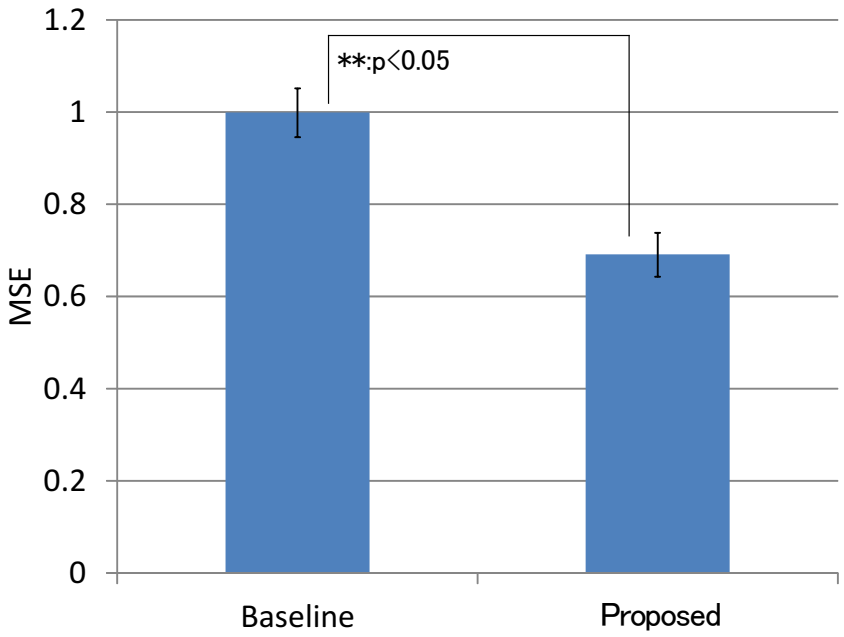

Figure 16. Evaluation for satisfaction prediction

of the baseline.

To analyze the effectiveness of features, we show ablation tests where we skip each variety of feature in Figure 17. From this result, we can see that the surface features of words are most effective. Features of word classes and the sentiment lexicon are not as important, but do provide some benefit.

4.4.2 Evaluation for Response Selection

In the response selection evaluation, we took 8 subjects who evaluate the responses provided by each response selection model. The subjects view replies selected by each response selection model for each input and assign satisfaction values for each reply. Thus, each subject gave a satisfaction score for 168 selected responses for 42 queries with 4 methods (RANDOM, MAXDB, MAXR, ADAPTIVE). The subjects also selected a response to which they want to reply and makes a feedback utterance for the selected response.

We compared the two baseline systems using random selection RANDOM, and the database of the most proficient writer MAXDB, with the proposed static re-

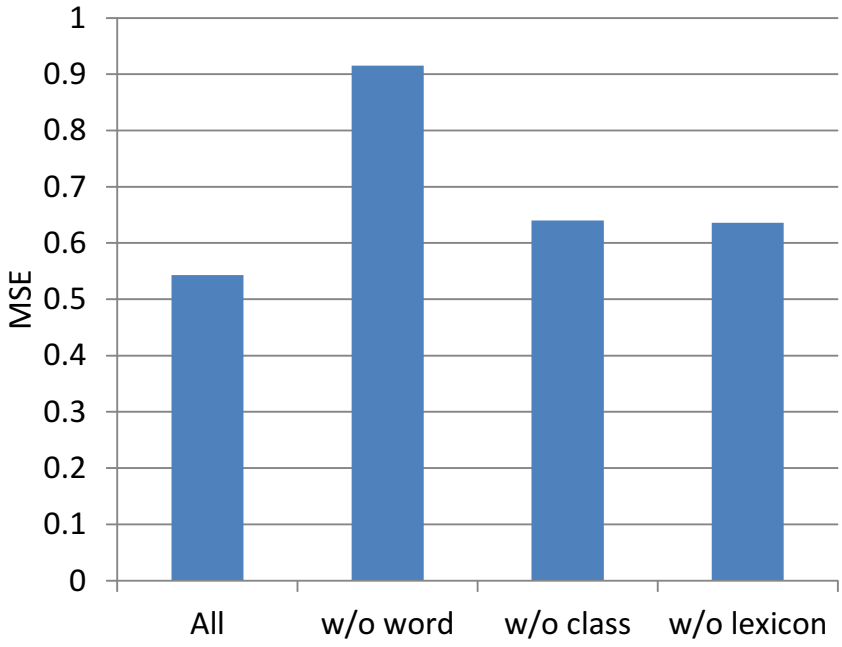

Figure 17. Ablation tests for satisfaction prediction

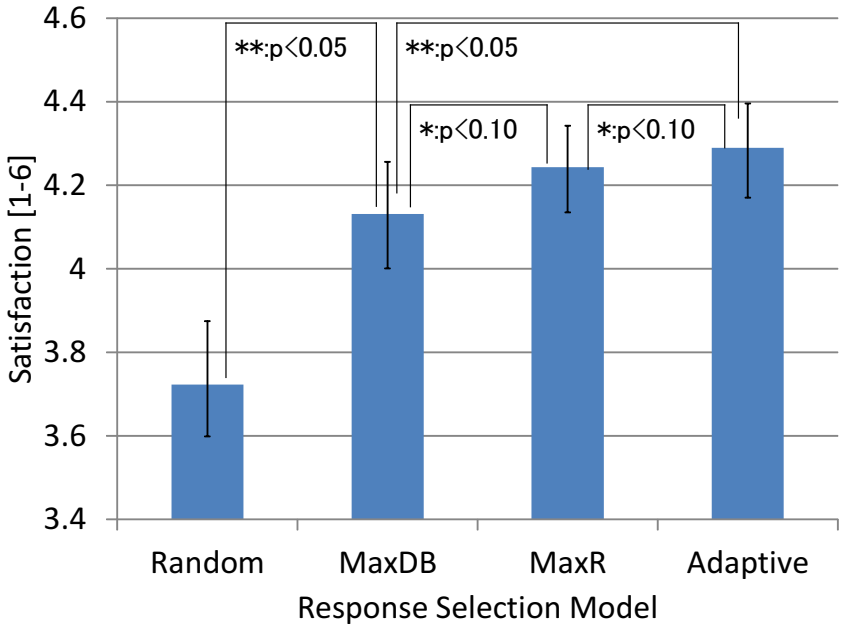

Figure 18. Evaluation for response selection

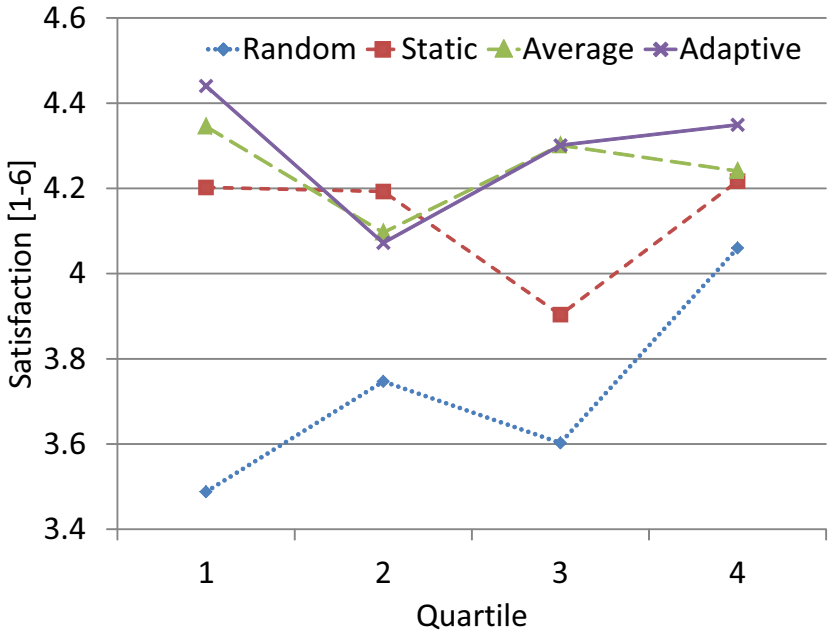

Figure 19. Satisfactions by quartile of the dialogue

sponse selection method MAXR , and the adaptive method ADAPTIVE, these are described in Section 4.3.2.

Error bars are obtained with bootstrap resampling, and we perform a pairwise bootstrap to measure significance of differences between each model ($p < 0.05$). In Figure 18, we show the evaluation for response selection.

First, focusing on the difference between RANDOM and MAXDB, we can see that we obtain a significant improvement by going from an example database in which quality or consistency of the response is not considered to having an example database with the highest average satisfaction. This demonstrates the validity of our premise that not all responses are created equal, and it is necessary to consider the quality and the expected satisfaction of the response in EBDM systems.

Second, focusing on the difference between MAXDB and MAXR, we also obtain a slight improvement. This demonstrates the utility of considering multiple responses for each utterance.

Finally, focusing on the difference between MAXDB and ADAPTIVE, we can see a significant improvement with the highest average satisfaction to having adaptive

response from all example databases. In addition, focusing on the difference between MAXR and ADAPTIVE, we can see a marginal significant improvement. These results indicate that performing adaptive response selection can increase in response quality.

In comparison with MAXR and ADAPTIVE, the EBDM problem, which always gives the same response, is still remained on MAXR, however, ADAPTIVE selects the different response depending on the history of dialogue. Furthermore, Figure 12 shows that the low correlation of satisfactions affects the upper limit of MAXR, and it inhibits to increase the user satisfaction by using MAXR. From these things, ADAPTIVE has merits to lead to further increase user satisfaction, and is improved architecture of EBDM to solve a problem of response monotonicity.

In Table 17, we show an example of responses selected by each model. In the 6-th turn of user A, MAXR and ADAPTIVE got the best satisfaction score from the user. These two system responses cause the interaction more kindly in comparison with other two system responses, and it is thought the reason that makes user satisfactory. Similarly, in the 37-th turn of user B, MAXR and ADAPTIVE selected the same system response, which was kindly interaction, and got a highest evaluation for a user. On the other hand, like the 29-th turn of user C, we were often able to observe the situation where the user wished the system did not strongly perform an interaction. From these results, user satisfaction is considered enough as well as the appropriateness for the system responses.

Finally, in Figure 19, we show average satisfactions for 4 quartiles of the dialogue (each period is approximately 10 tri-turns). From this result, we can see that ADAPTIVE is the same as MAXR in the 2nd and 3rd quartile's satisfactions, but in the final quartile, ADAPTIVE improves satisfaction, possibly indicating that the model has adapted to the user somewhat by the end of the dialogue.

Table 17. Examples of response selection by each model (translated from Japanese)

Turn	User	User Utterance	System Responses	Sat.	Model
6	A	I take a shower.	Certainly.	3	RANDOM
			When you take a shower, It seem to catch cold.	4	MAXDB
			Have a nice shower, please warm.	6	MAXR
			I prepare a change of clothes and a towel.	6	ADAPTIVE
37	B	Let me sleep a little more.	You said woke you up.	2	RANDOM
			It's fine today.	5	MAXDB
			I wake you up again at five minutes later.	6	MAXR, ADAPTIVE
29	C	Ah...	What did you happen?	4	RANDOM
			What did you say?	4	MAXDB
			Huh?	6	MAXR, ADAPTIVE
8	D	What shall we eat today?	Stew.	4	RANDOM
			Today is cold, let's make stew.	5	MAXDB, ADAPTIVE
			I think hamburger steak is better.	4	MAXR

4.5 Summary

In this chapter, we proposed methods constructing example-based dialogue system based on examples that pair one query and multiple responses, and adaptive response selection. In multi-response example database construction, we proposed the structure of example that has response candidates corresponding to a query. In response selection, we proposed two selection methods. The first is the STATIC method that considers the maximization of average satisfaction score, and the other is the ADAPTIVE method that uses collaborative filtering over explicit user feedback utterances. In an evaluation, we found that both proposed methods were effective, with adaptive response selection resulting in the highest satisfaction scores.

While the experimental results showed that the adaptive method is able to successfully select better response utterances, there are still a number of future challenges related to refining the example database and response selection model. The main potential for improvement lies in constructing response selection model acquired from larger training data. In collaborative filtering, the utility of performing collaborative filtering is largely influenced by whether a user similar to the current user can be found in the data. Therefore, it is important that there are a large number of diverse users in the database. Despite the fact that the database we used in this research was relatively small (5 annotators), we were still able to achieve an improvement in accuracy, but it is likely that larger databases could lead to further improvements in accuracy. We also plan to build the training data for response selection using un-annotated dialogue corpora. Specifically, when a dialogue is carried out by a new user, it may be beneficial to add the predicted satisfaction data as training data for collaborative filtering.

The proposed method has not only an improvement but also an innovation of engineering with EBDM architecture. The original EBDM architecture considers the only similarity between a user utterance and query utterances in the example database, it does not consider a response utterance when a conversational agent responds. In contrast, the proposed method considers both of query and response utterances to increase the response quality. This new EBDM architecture by considering both of query and response utterances gives a clue for future studies of EBDM that try to increase the performance of conversational agents.

Chapter 5

Response selection based on entrainment analysis

5.1 Introduction

In this chapter, we describe a response selection method based on entrainment analysis. This response selection method tries to select a response synchronized to user and conversation. First, we analyze entrainment on lexical and dialogue act level, and we construct an evaluation model for appropriateness of entrainment. Response selection method chooses a response based on this evaluation model to consider appropriateness of entrainment.

5.2 Entrainment in dialogue

Entrainment is a conversational phenomenon in which dialogue participants synchronize to each other with regards to various factors: lexical choice [Brennan and Clark, 1996], syntax [Reitter and Moore, 2007, Ward and Litman, 2007], style [Niederhoffer and Pennebaker, 2002, Danescu-Niculescu-Mizil et al., 2011], acoustic prosody [Natale, 1975, Coulston et al., 2002, Ward and Litman, 2007, Kawahara et al., 2015], pronunciation [Pardo, 2006] and turn-taking [Campbell and Scherer, 2010, Beňuš et al., 2014]. Previous studies have reported that entrainment is correlated with dialogue success, naturalness, and engagement.

However, there is much that is still unclear with regard to how entrainment affects the overall flow of the dialogue. For example, can entrainment also be observed in choice of dialog acts? Is entrainment on the lexical level more prevalent for utterances of particular dialogue acts? Does the level of entrainment increase as dialogue progresses?

If the answer to these questions is affirmative, it will be necessary to model entrainment not only on the lexical level, but also on the higher level of dialog flow. In addition, it will be necessary to adapt any entrainment features of conversational agents to be sensitive to dialogue acts or dialogue progression. Modeling

such entrainment phenomena appropriately has the potential to increase the naturalness of the conversation and open new avenues in human-machine interaction.

5.3 Analysis of the effect of entrainment

As mentioned in the introduction, entrainment has been shown to occur at almost every level of human communication [Levitan, 2013], including both human-human and human-system conversation.

In human-human conversation, Kawahara et al. showed the synchrony of backchannels to the preceding utterances in attentive listening, and they investigated the relationship between morphological patterns of backchannels and the syntactic complexities of preceding utterances [Kawahara et al., 2015]. Levitan et al. showed the entrainment of latency in turn-taking [Levitan et al., 2015].

In human-system conversation, Campbell et al. tried to predict user's turn-taking behavior by considering entrainment [Campbell and Scherer, 2010]. Fandrianto et al. modeled a dialogue strategy to increase the accuracy of speech recognition by using entrainment intentionally [Fandrianto and Eskenazi, 2012]. Levitan et al. unified these two studies [Levitan, 2013].

One of the most important questions about entrainment with respect to dialogue systems is its association with dialogue quality. Nenkova et al. proposed a score to evaluate the lexical entrainment in highly frequent words, and they found that the score has high correlation with task success and engagement [Nenkova et al., 2008]. This indicates that lexical entrainment has an important role in dialogue. In addition, it suggests that entrainment of lexical choice is probably affected by more detailed dialogue information, such as dialogue act.

5.3.1 Scoring of entrainment

The entrainment score that was proposed by Nenkova et al. is calculated by word counts in a corpus, and comparing between dialogue participants [Nenkova et al., 2008]. Specifically, we calculate a 1-gram language model probability $\mathrm{P}_{\mathrm{S}_1}(w)$ and $\mathrm{P}_{\mathrm{S}_2}(w)$ based on the word frequencies of speakers S_1 and S_2, and calculate the

entrainment score of word class V, $\mathrm{En}(V)$ as:

$$\mathrm{En}(V) = -\sum_{w \in V} \left| \mathrm{P}_{\mathrm{S}_1}(w) - \mathrm{P}_{\mathrm{S}_2}(w) \right|. \tag{20}$$

These entrainment scores have a range from -2 to 0, where higher means stronger entrainment.

In detail, we can express this formula with word count $\mathrm{C}_{S_1}(w)$ and $\mathrm{C}_{S_2}(w)$, and all of words W as,

$$\mathrm{En}(V) = -\sum_{w \in V} \left| \frac{\mathrm{C}_{S_1}(w)}{\sum_{w_i \in W} \mathrm{C}_{S_1}(w_i)} - \frac{\mathrm{C}_{S_2}(w)}{\sum_{w_i \in W} \mathrm{C}_{S_2}(w_i)} \right|. \tag{21}$$

[Nenkova et al., 2008] used following word classes as V.

25MFC: 25 Most frequent words in the corpus. The idea of using only frequent words is based on the fact that we would like to avoid the score being affected by the actual content of the utterance, and focus more on the way things are said. This word class was highly and significantly correlated with task success in the previous study. We mainly used this word class in this section.

25MFD: 25 Most frequent words in the dialogue. This word class was correlated with task success, like 25MFC.

ACW: Affirmative cue words [Gravano et al., 2012]. This word class includes *alright, gotcha, huh, mm-hm, okay, right, uh-huh, yeah, yep, yes,* and *yup.* This class was correlated with turn-taking.

FP: Filled pauses. This word class includes *uh, um,* and *mm.* It was correlated with overlaps.

ACW and FP were pre-defined, but 25MFC and 25MFD are calculated from corpora considering frequency ($V \in W$).

In order to use these measures to confirm whether entrainment is occurring between dialogue partners, these scores can be compared between the actual conversation partner, and an arbitrary other speaker in the database. If entrainment

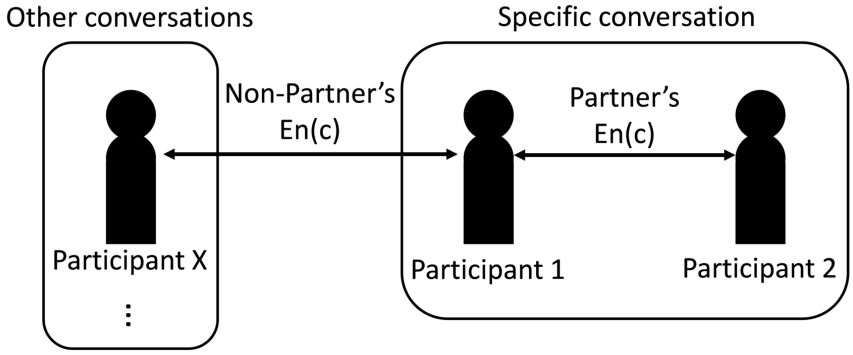

Figure 20. How to compare scores between the partner and non-partners

Table 18. The entrainment score of 25MFC

	Partner	Non-Partner
En(25MFC)	-0.211	-0.248

is actually occurring, then the score will be higher for the conversation partner than the score for the non-partner. Figure 22 shows an example of pairs used for calculation of these scores.

First, to confirm the results for previous study, we calculated the entrainment score of 25MFC using the Switchboard Corpus (Table 18). We can see that there is a difference of the entrainment score between "partner" who is talking the speaker and "non-partner" who is not talking with the speaker, as reported in previous study.

Our first contribution is an extension to the entrainment score that allows us to more accurate clarify the hypotheses that we stated in the introduction. This is necessary because the entrainment score given in Eqn. (22) does not consider the total size and variance of data to be calculated, and can be heavily influenced by data sparsity. This result in the score being biased when we compare target phenomena with different vocabulary sizes or data sizes.

For example, when considering the amount of entrainment that occurred for two different speakers, the entrainment score will tend to be higher for the more verbose speaker, regardless of the amount of entrainment that actually occurred. In addition, if we are comparing entrainment for two different sets of target phenomena, such as words and dialogue acts, the entrainment score will tend

to be higher for the phenomenon that has a smaller vocabulary and thus less sparsity (in this case, dialogue acts). Thus, we propose a new "entrainment rate" measurement that uses the rank in entrainment score, and language model smoothing to alleviate the effects of sparsity.

First, instead of using the entrainment score itself, we opt to use the relative position of the entrainment score of the partner compared to other non-partner speakers in the corpus. The entrainment score rank ratio is calculated according to the following procedure:

1. Calculate the entrainment score of the dialogue partner $\mathrm{En_p}(V)$. Also calculate entrainment scores of all non-partners in the corpus $\mathrm{En_{np_{1,\ldots,N}}}(V)$.

2. Compare the partner's entrainment score and all non-partners' entrainment scores.
$$\mathrm{Win}(\mathrm{En_p}(V), \mathrm{En_{np_i}}(V)) = \begin{cases} 1 & (\mathrm{En_p}(V) > \mathrm{En_{np_i}}(V)) \\ 0.5 & (\mathrm{En_p}(V) = \mathrm{En_{np_i}}(V)) \\ 0 & (\mathrm{En_p}(V) < \mathrm{En_{np_i}}(V)) \end{cases}$$

3. Calculate the ratio with which the partner's entrainment score exceeds that of the non-partners.
$$\mathrm{Ratio}(V) = \frac{1}{|N|} \sum_{i \in N} \mathrm{Win}(\mathrm{En_p}(V), \mathrm{En_{np_i}}(V))$$

Because this score is the ratio that dialogue with the partner takes a higher entrainment score than other combinations with non-partners, it is not sensitive to the actual value of the entrainment score, but only the relative value compared to non-partners. This makes it more feasible to compare between phenomena with different vocabulary sizes, such as lexical choice and dialogue act choice. While the entrainment score for dialogue acts may be systematically higher due to its smaller vocabulary size, the relative score compared to non-partners can be expected to be approximately equal if the effect of entrainment is the same between the two classes.

While the previous ranking score has the potential to alleviate problems due to comparing different types of phenomena, it does not help with problems caused by comparing data sets with different numbers of data points. The reason for

this is that the traditional entrainment score [Nenkova et al., 2008] used 1-gram probabilities, the accuracy of which is dependent on the amount of data used to calculate the probabilities. Thus for smaller data sets, these probabilities are not well trained, and show a lower similarity when compared with those of other speakers in the corpus. In order to create a method more robust to these size differences, we introduce a method that smooths these probabilities to reduce differences between distributions of different data sizes.

Specifically, the definition of a unigram distribution of a portion of the corpus (split by speaker s, dialogue act d, part of dialogue p) using maximum likelihood estimation is,

$$\mathrm{P}_{\mathrm{ML,s}}(w|d,p) = \frac{\mathrm{C}_s(w_{d,p})}{\sum_{w_{d,p} \in W_{d,p}} \mathrm{C}_s(w_{d,p})}. \tag{22}$$

When the size of data for speaker s is small, there will not be enough data to properly estimate this probability. To cope with this problem, we additively smooth the probabilities by introducing a smoothing factor α and large background language model $\mathrm{P}_{\mathrm{ML}}(w)$ that was trained using all of the available data:

$$\mathrm{P}_{\mathrm{DS,s}}(w|d,p) = \frac{\mathrm{C}_s(w_{d,p}) + \alpha \mathrm{P}_{\mathrm{ML}}(w)}{\sum_{w_{i,d,p} \in W_{d,p}} \mathrm{C}_s(w_{i,d,p}) + \alpha}. \tag{23}$$

This additive smoothing is equivalent to introducing a Dirichlet distribution conditioned on $\mathrm{P}_{\mathrm{ML}}(w)$ as a prior probability for the small language model distribution of $P_{DS,s}(w|d,p)$ [MacKay and Peto, 1995]. We determine the hyperparameter α by defining a Dirichlet process [Teh et al., 2006] prior, and maximizing the likelihood using Newton's method[6].

To verify that this method is effective, we calculated averages and variances of the standard entrainment score and the entrainment score using this proposed smoothing technique (Table 19). From the results, we can see that the entrainment rate for partners is slightly higher with smoothing, demonstrating that the smoothed scores are as effective, or slightly more effective in identifying the actual conversational partner. In addition, the difference between variances of entrainment scores has decreased, showing that smoothing has reduced the amount of

[6]The scripts for this and other calculations will be public at the link below:
`https://github.com/masahiro-mi/entrainment`

Table 19. The entrainment score variance with/without smoothing

		Partner		Non-Partner	
	Rate	Ave.	Var.	Ave.	Var.
w/o smoothing	0.671	-0.211	0.00537	-0.248	0.00181
w/ smoothing	0.706	-0.0983	0.00108	-0.123	0.000778

Table 20. The entrainment score of dialogue acts

	Partner	Non-Partner	Rank
En(D)	**-0.568****	-0.715	0.675

* $p < 0.10$, ** $p < 0.05$

fluctuation in scores. This indicates that the smoothing works effectively to reduce the negative influence of population size when we compare distributions that have different population sizes. Because of this, for the analysis in the rest of the section we use this smoothed entrainment score.

5.3.2 Entrainment of dialogue acts

First, we analyze the entrainment of dialogue acts based on the method of Section 5.3.1. We hypothesize that we can observe the entrainment of dialogue acts like other previously observed factors. To examine this hypothesis, we calculated the entrainment score of dialogue acts and compared between partner and non-partners. To measure the significance of these results, we calculated p-value of entrainment scores between partner and non-partner with the t-test.

Table 20 shows that there is a significant difference ($p < 0.05$) of entrainment score between partner and non-partner, with partners scoring significantly higher than non-partners. This result shows that the entrainment of dialogue acts can be observed in human-human conversation, and suggests that there may be a necessity to consider entrainment of dialogue act selection in human-machine interaction.

5.3.3 Lexical Entrainment given dialogue acts

Next, we analyze the entrainment of lexical choice given the 42 types of dialogue acts based on the method of Section 5.3.1. We can assume that the dialogue act

affects the entrainment of lexicons, which indicates that entrainment scores are different depending on the type of the given dialogue act.

In addition, we calculate entrainment score rate and Cohen's d [Cohen, 1988] to evaluate the effect size. Cohen's d is standardized mean difference between two groups, and can calculate the amount that a particular factor effects a value while considering each group's variance. If these groups have a large difference, Cohen's d will be larger, with values less than 0.2 being considered small, values around 0.5 being medium, and values larger than 0.8 being considered large.

We show the result in Table 21, and emphasize scores that are over 0.5 in Cohen's d, and over 0.55 in rate.

We can first notice an increase of the entrainment score is more prominent given some dialogue acts, Specifically, increasing of the entrainment score given following dialogue acts: Conventional closing, Conventional opening, Statement opinion, Statement non opinion, Acknowledge (Backchannel), Agree/Accept, Segment (multi-utterance), Appreciation, Abandoned or Turn-Exit, Uninterpretable, Yes answers, Non verbal, Hedge, Wh-Question, Backchannel in question form, Rhetorical-Questions, Response Acknowledgement, Repeat-phrase, Quotation, Collaborative Completion, Hold before answer/agreement, Summarize/reformulate, and Signal-non-understanding is obtained. Entrainment is particularly prevalent for acts that have little actual informational content, such as greeting, backchannel, agree, answer, and repeating.

In addition, we focus on why Conventional Opening and Conventional Closing were increased in the entrainment score. This is because that Conventional Opening and Conventional closing tend to be greetings ("hi", "hello") or farewells ("bye", "see you"), which tend to entrain more strongly than other words.

On other hand, dialogue acts that express one's opinion such as Apology, Action-directive, Negative non-no answers, as well as some questions do not increase entrainment scores.

5.3.4 Change in entrainment through dialogue

In addition, we analyzed the increase of entrainment based on the method of Section 5.3.1. We calculated entrainment scores of the earlier and later parts. "Earlier" is the entrainment score between utterances in the earlier part of dia-

Table 21. The entrainment score of lexicons given a dialogue act

	Partner	Non-Partner	Cohen's d	Rate
Conventional-closing	**-0.0391****	-0.185	**1.50**	**0.703**
Acknowledge (Backchannel)	**-0.201****	-0.252	**0.527**	**0.659**
Statement-non-opinion	**-0.0930****	-0.113	0.434	**0.672**
Statement-opinion	**-0.154****	-0.192	0.418	**0.634**
Conventional-opening	**-0.0112****	-0.0370	0.406	0.542
Uninterpretable	**-0.203****	-0.232	0.382	**0.618**
Agree/Accept	**-0.279****	-0.325	0.367	**0.592**
Appreciation	**-0.282****	-0.331	0.322	**0.564**
Yes answers	**-0.320****	-0.375	0.274	**0.555**
Non-verbal	**-0.104****	-0.124	0.259	**0.557**
Abandoned or Turn-Exit	**-0.203****	-0.228	0.244	**0.592**
Hedge	**-0.170****	-0.191	0.132	0.532
Wh-Question	**-0.147****	-0.160	0.122	0.530
Backchannel in question form	**-0.134****	-0.152	0.118	0.528
No answers	**-0.199****	-0.220	0.118	0.523
Rhetorical-Questions	**-0.0644****	-0.0754	0.102	0.522
Response Acknowledgement	**-0.207****	-0.227	0.100	0.521
Repeat-phrase	**-0.115****	-0.128	0.0962	0.522
Other	-0.160	**-0.150****	0.0772	0.476
Quotation	**-0.0817****	-0.0905	0.0749	0.517
Collaborative Completion	**-0.0867****	-0.0929	0.0616	0.514
Yes-No-Question	**-0.223***	-0.227	0.0490	0.512
Hold before answer/agreement	**-0.104****	-0.112	0.0488	0.511
Summarize/reformulate	**-0.109****	-0.114	0.0380	0.512
Signal-non-understanding	**-0.0377****	-0.0404	0.0377	0.507
Tag Question	**-0.0148****	-0.017	0.0356	0.504
Declarative Yes-No-Question	-0.134*	-0.138	0.0348	0.512
Other answers	-0.0584*	-0.0620	0.0313	0.507
Maybe/Accept-part	-0.0204	-0.0221	0.0247	0.503
Self-talk	-0.0189	-0.0205	0.0235	0.503
Thanking	-0.0180	-0.0195	0.0227	0.502
Reject	-0.0670	-0.0696	0.0209	0.504
Negative non-no answers	-0.0600	-0.0581	0.0181	0.497
Open-Question	-0.0877	-0.0894	0.0166	0.504
Affirmative non-yes answers	-0.134	-0.136	0.0161	0.504
Downplayer	-0.0238	-0.0247	0.0111	0.501
Declarative Wh-Question	-0.0147	-0.0152	0.00797	0.501
Action-directive	-0.0935	-0.0944	0.00748	0.502
Dispreferred answers	-0.0514	-0.0522	0.00716	0.502
Apology	-0.0183	-0.0179	0.00667	0.500
3rd-party-talk	-0.00969	-0.00955	0.00369	0.500
Offers, Options Commits	-0.0204	-0.0205	0.00222	0.500
Or-Clause	-0.0502	-0.0502	0.000816	0.500

N(Number of target speaker) = 2310, * $p < 0.10$, ** $p < 0.05$

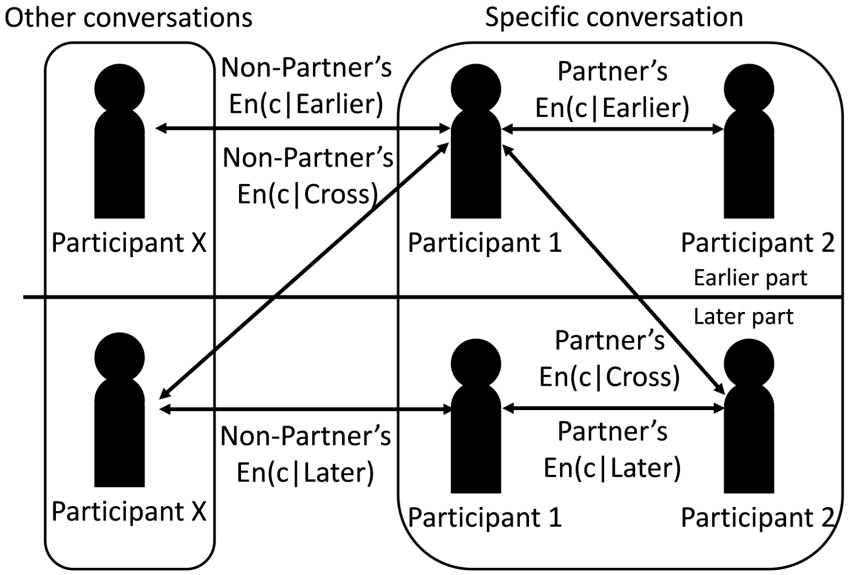

Figure 21. How we compare between earlier and later parts

Table 22. The entrainment score for combinations of part

	Partner	Non-Partner	Rate
En(25MFC|Earlier)	**-0.106****	-0.126	0.658
En(25MFC|Cross)	**-0.106****	-0.127	0.666
En(25MFC|Later)	**-0.104****	-0.126	0.674

* $p < 0.10$, ** $p < 0.05$

logue, and "Later" is the entrainment score between utterances in the later part. We hypothesize that "Later" will have a higher entrainment score than "Earlier," as it is possible that dialogue participants will demonstrate more entrainment as they talk for longer and grow more comfortable with each other.

In addition, we calculate "Cross," the entrainment score between the earlier and the later parts of dialogue. We calculated this because we can also hypothesize that the effect of entrainment is delayed, and words spoken in the earlier part of the conversation may appear in the later part of the partner's utterances. Figure 21 shows the pairs used for the calculation. We show the result in Table 22.

From these results, we can see that there is a significant difference of entrainment score between partner and non-partner in all of the parts. This indicates

Table 23. The p-values for partner's entrainment score between each part

		p-value
En(25MFC|Earlier)	En(25MFC|Later)	0.222
En(25MFC|Earlier)	En(25MFC|Cross)	0.238
En(25MFC|Later)	En(25MFC|Cross)	0.00425

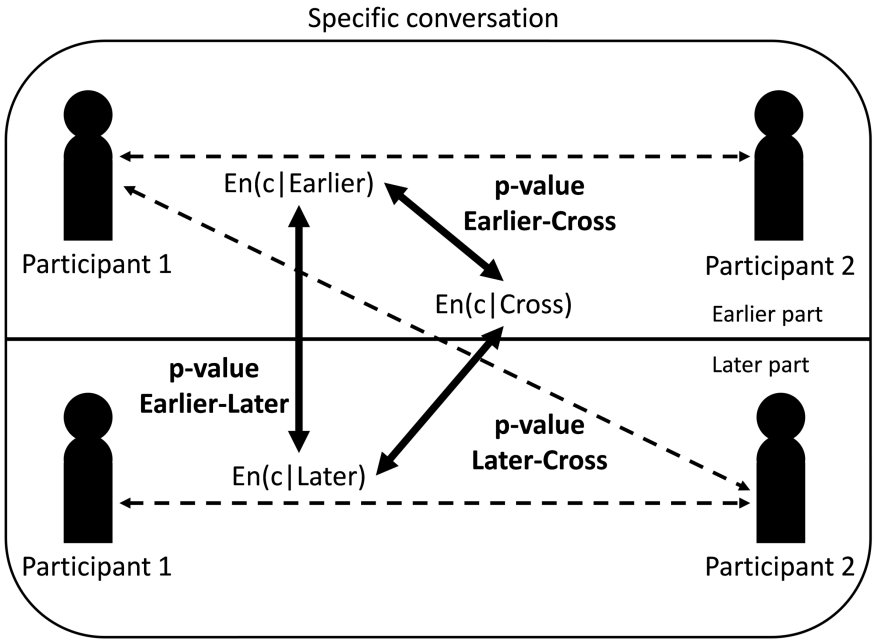

Figure 22. How to calculate p-values between each part in partner

that lexical entrainment can already be observed in the earlier part of dialogue.

In addition, we calculated p-values with the two-sided t test for partner entrainment scores between each part. Figure 21 shows an example of pairs used for calculation of p-values. We compare partner's entrainment scores between early, later, and cross, to indicate how the entrainment score changes in the partner through the dialogue. In fact, we compare three combinations of partner's entrainment scores, such as En(c|Earlier) and En(c|Later), En(c|Earlier) and En(c|Cross), and En(c|Later) and En(c|Cross). Table 23 shows that p-values of entrainment scores between each part in the partner. We find that the value of the entrainment score of the later part increased slightly over the entrainment score of the earlier part, but the increase was not significant. These results show

that if there is a difference in entrainment between earlier and later parts of the conversation, the difference is slight.

5.3.5 Summary of analysis

In this section, we explain in details of their varieties of entrainment that we examined. We focused on the entrainment with respect to dialogue acts and dialogue progression, and analyzed three phenomena: the entrainment of dialogue acts, the entrainment of lexical choice given dialogue acts, and the change in entrainment as the dialogue progresses.

From these results, we confirmed that the entrainment of dialogue acts was observed in conversations. Within dialogue systems, this result indicates the potential of entrainment given dialogue acts to contribute for the modeling of dialogue strategy, which allows the system to have a closer relationship with the partner. We also found that lexical entrainment has a different tendency depending on the type of dialogue act of the utterance. This result indicates the potential of contribution to models of language generation, which can consider entrainment of each dialogue act. We analyzed the differences of entrainment depending on the part of the dialogue.

5.4 Response selection based on dialogue act dependent entrainment

Previous studies indicate that lexical entrainment has an important role in dialogue to achieve high naturalness and engagement. Our analysis also indicates that the behavior depended on dialogue acts is important. They suggest that considering the appropriateness of entrainment given dialogue acts on response selection will increase conversational agent performances. In this section, we explain about the framework of response selection based on entrainment that try to entrain with users in a similar manner.

First, we show the overview of the proposed framework in Figure 23. The basic architecture is based on the EBDM, however, we consider not only similarity of query utterances but also the appropriateness of entrainment of responses to their queries. Specifically, this framework selects a response r_j as dialogue exam-

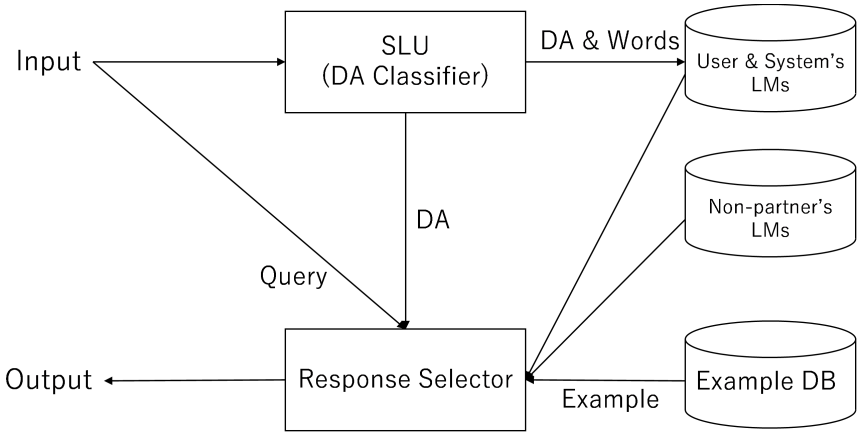

Figure 23. overview of proposed framework

ple $\langle q_j, r_j \rangle$ from example database $\mathbf{e}$, with maximizing of two scores: similarity $\mathrm{Sim}(q', q)$ with user utterance q' and example query q_j, and appropriateness of entrainment $\mathrm{Entr}(r_j)$ that is based on entrainment analysis.

$$\hat{r} = \underset{\langle q_j, r_j \rangle \in \mathbf{e}}{\mathbf{argmax}} (\lambda \mathrm{Sim}(q', q_j) + (1 - \lambda)\mathrm{Entr}(r_j)). \tag{24}$$

The score of appropriateness of entrainment $\mathrm{Entr}(r_j)$ is the measure based on Section 5.3.1. This measure will be 0 when conversational agent responds with the same utterance to the participant's response. However, each dialogue act has different appropriateness of entrainment is different, and it means that responding with the same utterance of the participant's response is not the best strategy. Therefore, we should predict the appropriateness of entrainment $\mathrm{Entr}(r_j)$ by considering dialogue acts, to select the best response from example database. Specifically, we calculate the appropriateness of entrainment $\mathrm{Entr}(r_j)$ by using conditional language model probabilities $P_{user}(w|d)$ and $P_{system}(w|d)$ given a dialogue act d. In addition, we calculate conditional language model probabilities $P_{non-partner_i}(w|d)$ $(i \in N)$ that were trained depending on speakers who do not participate the dialogue. We compare these language model probabilities, and calculate entrainment score ratio to calculate appropriateness of entrainment.

In our response selection method, we calculate conditional language model probabilities $P_{system+r_j}(w|d)$ for each response candidates r_j to calculate the en-

trainment score ratio $R(V|d, P_{user}, P_{system+r_j})$ as follow:

$$\mathrm{En}_{\mathrm{participants}}(V|d) = -\sum_{w \in V} \left| \mathrm{P}_{\mathrm{user}}(w|d) - \mathrm{P}_{\mathrm{system+r_j}}(w|d) \right| \tag{25}$$

$$\mathrm{En}_{\mathrm{non-partner_i}}(V|d) = -\sum_{w \in V} \left| \mathrm{P}_{\mathrm{non-partner_i}}(w|d) - \mathrm{P}_{\mathrm{system+r_j}}(w|d) \right| \tag{26}$$

$$\mathrm{R}(V|d, P_{user}, P_{system+r_j}) = \frac{1}{N} \sum_{i \in N} \begin{cases} 1 & (\mathrm{En}_{\mathrm{participants}}(V|d) > \mathrm{En}_{\mathrm{non-partner_i}}(V|d)) \\ 0.5 & (\mathrm{En}_{\mathrm{participants}}(V|d) = \mathrm{En}_{\mathrm{non-partner_i}}(V|d)) \\ 0 & (\mathrm{En}_{\mathrm{participants}}(V|d) < \mathrm{En}_{\mathrm{non-partner_i}}(V|d)) \end{cases} \tag{27}$$

This equation of $\mathrm{Entr}(r_j)$ is the appropriateness of entrainment score ratio, in other words, difference between an entrainment score ratio $R(V|d, P_{user}, P_{system+r_j})$ and an ideal entrainment score ratio $R_{ideal}(V|d)$ with normalization. The ideal entrainment score ratio $R_{ideal}(V|d)$ has been provided from the result of entrainment analysis, which used training data.This framework selects a response that is synchronized to user and improves naturalness.

$$\mathrm{Entr}(r_j) = 1 - \left| \frac{\mathrm{R}(V|d, P_{user}, P_{system+r_j}) - \mathrm{R}_{\mathrm{ideal}}(V|d)}{\max\{1 - \mathrm{R}_{\mathrm{ideal}}(V|d), \mathrm{R}_{\mathrm{ideal}}(V|d)\}} \right| \tag{28}$$

5.5 Experimental result of response selection

We evaluated the proposed method with the Switchboard DA corpus as same as the analysis of section 5.3, and divided the corpus into 9:1 to train language models and to test the response selection method.

In the test of response selection, we calculated the similarity between the true response utterance r, which obtained from test data, and the response candidate r_j, which obtained from training data. $\mathrm{Sim}(r, r_j)$ is assumed as the quality of response r_j, and the performance of the response selection is calculated by mean square error (MSE) between the response selection criterion $(\lambda \mathrm{Cos}(q', q_j) + (1 - \lambda)\mathrm{Entr}(r_j))$ and the similarity of responses $\mathrm{Cos}(r, r_j)$. Therefore, we calculate

MSEs to evaluate a proposed method as follows:

$$MSE = \sum_{\langle q_j, r_j \rangle \in \mathbf{e}'} ((\lambda \mathrm{Sim}(q', q_j) + (1 - \lambda)\mathrm{Entr}(r_j)) - \mathrm{Sim}(r, r_j))^2 \qquad (29)$$

These response candidates of 20-best that have higher similarity scores of queries $\mathrm{Sim}(q', q_j)$ in example database $\mathbf{e}$ are selected as $\mathbf{e}'$. In addition, we collected candidates $\mathbf{e}$, which have the same DA to the true response utterance. DA is related to constraints of the appropriateness of entrainment score ratio $\mathrm{Entr}(r_j)$ calculation.

We show the result of MSEs and entrainment score ratios in Table 24. We show results of $\lambda = 0$ (it does not consider entrainment in response selection), $\lambda = 1$ (it considers only entrainment appropriateness), and $\lambda = 0.5$ (it considers query similarity and entrainment appropriateness equally).

From Table 24, we obtain the decreasing average MSE of 0.08 than baseline average MSE of 0.10 that is set $\lambda = 1.0$. An appropriateness of entrainment score ratio $\mathrm{Entr}(r_j)$ works to decrease MSE, and it means that a synchronized response is better than a response that is not considered entrainment. Specifically, DAs of Acknowledge (Backchannel), Non-verbal, and Hedge, which analyzed as strong entrainment in a conversation, have low λ, and it signifies the importance of appropriateness of entrainment. In contrast, DAs of 3rd-party-talk, Apology, and Downplay have high λ, it means these responses are not entrained.

In this experimental evaluation, the performance will be the upper bound because it is the optimal parameter because we calculate the best λ from the training data. Estimating these parameters is a remaining future. However, this result of experimental evaluation investigated that the performance of response selection increases with considering the appropriateness of entrainment.

One simple method to synchronize for a user is always repeating content words of user utterance like ELIZA. This repeating method makes it possible to synchronize and gets a high entrainment score ratio. However, our results of analysis and objective evaluations show that using high entrainment candidate anytime is not appropriate. For example, in DAs of Apology, Downplay, 3rd-party-talk, and more, the result of response selection shows that it is not good to synchronize to a user at all.

5.6 Summary

In this chapter, we focused on the entrainment with respect to dialogue acts and dialogue progression, and analyzed three phenomena: the entrainment of dialogue acts, the entrainment of lexical choice given dialogue acts, and the change in entrainment as the dialogue progresses.

According to these results of analysis, we built a response selection method that can control the entrainment of responses. Our proposed method select a response that is synchronized to the user according to DA and language models.

An experimental evaluation shows us that the proposed method is effective, which achieves lower MSE if the method considers the appropriateness of entrainment. Entrainment score ratio of selected responses shows us that these responses were synchronized to the user appropriately.

These analyses and the response selection are based on data-driven methods that have no constraint of a specific task and corpus. It makes it possible to analyze and to respond by any corpus without a limitation of tasks.

Table 24. Lambda and MSE given a dialogue act

	Baseline				Proposed					
	$\lambda = 1.0$		$\lambda = 0.0$		$\lambda = 0.5$		Set λ appropriately			
DA	MSE	$Entr_r$	MSE	$Entr_r$	MSE	$Entr_r$	λ	MSE	$Entr_r$	Ratio
Acknowledge (Backchannel)	0.23	0.59	0.15	0.60	**0.09**	0.60	0.00	**0.15**	0.60	0.61
Non-verbal	0.12	0.47	0.13	0.50	**0.05**	0.50	0.00	0.13	0.50	0.56
Yes answers	0.27	0.49	0.12	0.52	**0.10**	0.52	0.00	**0.12**	0.52	0.55
Hedge	0.29	0.53	0.12	0.56	**0.09**	0.56	0.00	**0.12**	0.56	0.53
No answers	0.16	0.46	0.21	0.50	**0.07**	0.50	0.00	0.21	0.50	0.52
Response Acknowledgement	0.24	0.49	0.17	0.52	**0.06**	0.52	0.00	**0.17**	0.52	0.52
Backchannel in question form	0.27	0.49	0.18	0.52	**0.10**	0.52	0.05	**0.16**	0.52	0.53
Hold before answer/agreement	0.18	0.55	0.25	0.48	**0.07**	0.48	0.18	**0.15**	0.48	0.51
Open-Question	0.18	0.55	0.23	0.51	**0.05**	0.51	0.25	**0.10**	0.51	0.51
Agree/Accept	0.13	0.56	0.18	0.58	**0.06**	0.58	0.26	**0.10**	0.58	0.59
Signal-non-understanding	0.14	0.55	0.36	0.53	**0.07**	0.53	0.28	0.16	0.53	0.51
Other	0.18	0.48	0.27	0.48	**0.06**	0.48	0.29	**0.10**	0.48	0.48
Conventional-closing	0.16	0.74	0.16	0.78	**0.04**	0.77	0.30	**0.06**	0.77	0.68
Reject	0.15	0.54	0.43	0.50	**0.10**	0.50	0.30	0.18	0.50	0.50
Appreciation	0.12	0.54	0.19	0.58	**0.05**	0.58	0.32	**0.08**	0.58	0.56
Dispreferred answers	0.08	0.42	0.55	0.48	0.11	0.47	0.36	0.20	0.48	0.50
Negative non-no answers	0.15	0.46	0.38	0.48	**0.04**	0.48	0.37	0.09	0.48	0.50
Conventional-opening	0.08	0.71	0.24	0.70	0.08	0.70	0.38	0.10	0.70	0.54
Affirmative non-yes answers	0.12	0.48	0.38	0.52	**0.07**	0.52	0.40	**0.10**	0.52	0.50
Wh-Question	0.10	0.47	0.16	0.50	**0.03**	0.50	0.41	**0.04**	0.50	0.53
Abandoned or Turn-Exit	0.08	0.54	0.21	0.57	**0.06**	0.57	0.44	**0.07**	0.57	0.59
Repeat-phrase	0.09	0.47	0.37	0.51	**0.07**	0.51	0.45	0.09	0.51	0.52
Yes-No-Question	0.05	0.51	0.20	0.53	0.05	0.53	0.53	**0.04**	0.53	0.51
Declarative Yes-No-Question	0.08	0.49	0.40	0.52	**0.07**	0.52	0.53	**0.06**	0.52	0.50
Action-directive	0.08	0.50	0.28	0.53	**0.05**	0.53	0.53	**0.04**	0.53	0.51
Other answers	0.08	0.56	0.59	0.54	0.16	0.54	0.53	0.14	0.54	0.51
Statement-non-opinion	0.06	0.59	0.26	0.61	0.06	0.61	0.54	0.06	0.61	0.67
Statement-opinion	0.06	0.54	0.25	0.60	0.06	0.60	0.55	0.05	0.60	0.65
Quotation	0.02	0.47	0.87	0.47	0.22	0.47	0.59	0.16	0.47	0.50
Or-Clause	0.06	0.57	0.33	0.57	0.06	0.56	0.59	**0.04**	0.56	0.52
Summarize/reformulate	0.06	0.51	0.34	0.52	0.06	0.53	0.60	**0.04**	0.53	0.51
Rhetorical-Questions	0.05	0.55	0.46	0.51	0.09	0.51	0.62	0.05	0.51	0.52
Collaborative Completion	0.04	0.50	0.50	0.48	0.11	0.48	0.65	0.05	0.48	0.51
Uninterpretable	0.03	0.54	0.24	0.57	0.05	0.57	0.67	0.03	0.57	0.59
Maybe/Accept-part	0.05	0.41	1.20	0.50	0.42	0.50	0.92	0.08	0.50	0.50
Thanking	0.11	0.50	1.52	0.50	0.58	0.50	0.94	0.14	0.50	0.50
Self-talk	0.09	0.48	1.42	0.49	0.52	0.50	0.95	0.11	0.48	0.50
Offers, Options, Commits	0.03	0.50	1.15	0.50	0.35	0.50	0.97	0.04	0.50	0.50
Declarative Wh-Question	0.15	0.49	1.67	0.49	0.68	0.49	1.00	0.15	0.49	0.50
Downplayer	0.11	0.52	1.52	0.50	0.59	0.49	1.00	0.11	0.52	0.50
Tag-Question	0.09	0.64	1.22	0.50	0.46	0.50	1.00	0.09	0.64	0.50
3rd-party-talk	0.18	0.50	1.89	0.50	0.79	0.50	1.00	0.18	0.50	0.50
Apology	0.23	0.49	2.08	0.50	0.91	0.50	1.00	0.23	0.49	0.50
Ave	0.10	0.57	0.23	0.59	**0.07**	0.59	0.40	**0.08**	0.59	0.62

Chapter 6
Conclusion

6.1 Summary of this study

This thesis proposed an adaptive conversational agent that considers user preference on EBDM architecture. The user preference depends on agent individuality, response quality, response tendency, and entrainment. Therefore, we have to consider each factor to develop an adaptive conversational agent.

First, we focused on the linguistic individuality control of agent responses. The proposed method transforms an utterance that has specific target individuality from the original utterance based on the statistical machine translation framework and translation models. In experimental evaluation, we conducted subjective evaluations in both speaker groups of camera sales clerks and Twitter characters, and it was shown that the proposed methods improved the individuality significantly in both groups of camera sales clerks and Twitter characters. These results are consistent with other linguistic findings.

Second, we proposed a satisfaction prediction method, which evaluates example database in advance of using a conversational agent, which predicts user expected satisfaction during construction. The system can filter out non-satisfactory examples to achieve building high-quality example database before hand the conversation. We confirmed that our proposed model works well on the example database in Japanese daily life domain and English drama and movie script domain. These two experimental results of a satisfaction prediction show that the proposed method makes it possible to construct the high-quality example database.

Third, we proposed an adaptive response selection that considers user preferences. Standard EBDM architecture has no adaptability on response selection module, thus we could improve a response selection module to be able to adapt the response to user preferences. Our multi-response example database and adaptive response selection module increased user satisfaction based collaborative filtering for user feedback. In the experimental evaluation, the adaptive method increased user satisfactions more than the baseline method. Therefore, the adaptive re-

sponse selection method can select the better response for user preference under the limitation of EBDM.

Finally, we proposed a response selection based on entrainment analysis. We already know that the entrainment has an important role to make rapport. We analyze the entrainment with respect to dialogue acts, and we confirmed that the lexical entrainment has a different tendency depending on the type of dialogue act of the utterance. From this analysis, we propose the approach that tries to select a response considering the appropriateness of lexical entrainment given dialogue acts in a conversation. The appropriateness of entrainment is calculated by using the conversational agent's language model, the user's language model, and language models that trained by other conversations. In the experimental evaluation, the proposed method decrease MSE between the response selection criterion and the similarity of responses. Therefore, the proposed method selects more appropriate response for the user to synchronize depending on the dialogue act. It allows a conversational agent to synchronize to the user.

The proposed methods make it possible for the conversation agent to give users satisfactory responses in our individual and synchronized manner.

6.2 Remaining problems and future directions

We discuss the improvement of example database construction and response selection on EBDM, aiming to build an adaptive conversational agent. Many other works still remain to improve the conversational agent. For example, one of the most serious problems of EBDM is that architecture can not consider the context of a conversation. This problem prevents the conversational agent to talk with a user in deep context and makes difficult to establish rapport with conversational agent and user. Therefore, it is necessary to improve not only the modules but also the fundamental architecture of EBDM.

For the remaining problems on our research, it is necessary to consider more variations of user preference, for example, voice, dialogue strategies, and user personal information. We have a problem of individuality mismatch between acoustic and linguistic preferences on the current architecture. We only targeted the linguistic individuality transformation of speaking style on text (linguistic) stage. We need to convert the text to speech to build a conversational agent on

speech to speech. In other words, we need to clarify the relationship between linguistic individuality and acoustic individuality.

A balance of variation and quality on constructing example database is important. In the method of satisfaction prediction, we filtered examples only with satisfactions, however, the variation of example responses is also important for the response selection to consider user preferences. To achieve the best performance of the response selection, we have to filter the examples that consider not only satisfaction but also variation to leave more candidates in the response selection.

The adaptive response selection has to consider more variations of user preference to achieve higher adaptability. Our response selection method considers only satisfaction scores that are predicted from user feedbacks, however, considering other effective factors to estimate user satisfaction and preferences is necessary. For example, acoustic features, personal information, and facial information may affect the performance of prediction.

We have to evaluate a method of response selection based on entrainment analysis by the human. While the proposed response selection method works to make rapport, we need to handle a wider variety of entrainment. Especially, entrainment on acoustic factors may affect the rapport with human as speech to speech conversation.

Finally, we have to integrate the proposed methods. In current methods, we distinguish a method according to the purpose. For example, we use individuality transforming to give the first impression for a user, and use entrainment response selection to make the user more friendly after giving the first impression. However, to obtain the best performance, we have to handle these methods by a cooperative control. Specifically, we have to handle the tasks of synchronization and individuality simultaneously, because these modules have language models that work for the agent response decision. In future works, we try to develop an overall dialogue model to handle these proposed methods.

Appendix

A. Paraphrasing Database: Japanese

A.1 Introduction

Paraphrases are alternative ways of conveying the same meaning, and are useful in a number of NLP applications such as machine translation and question answering [Callison-Burch et al., 2006, Hermjakob et al., 2002]. In this section, we concern ourselves with building paraphrase resources for Japanese. In Japanese, methods and resources have been proposed for paraphrasing for a number of categories of paraphrases or domains [Ohtake and Yamamoto, 2001, Nakagawa and Masuda, 2004, De Saeger et al., 2009, Hashimoto et al., 2011]. However, most of these resources focus on a particular phenomenon in Japanese, and there are still no broad-coverage and freely available resources.

In previous research on paraphrasing, methods that use bilingual corpora have proven successful [Bannard and Callison-Burch, 2005, Ganitkevitch et al., 2013]. In these methods, paraphrases for one language (e.g., English) are acquired by treating another language (e.g., French) as an intermediate meaning representation, as described more completely in Section A.2. In this section, we describe a paraphrasing resource that we constructed for Japanese using a similar method, with Japanese as our target language and English serving as the intermediate meaning representation. In contrast, most previous work has focused on using bilingual corpora for language pairs such as English-French, English-Spanish and other Germanic and Romance languages in which the word order and grammar are similar, as shown in the example in Figure 24.

In contrast, there is a large divergence in both the word order, and the grammatical structure between Japanese and English. We describe in Section A.3 how we use a syntactic preprocessing method, Head Finalization [Isozaki et al., 2010], to help compensate for this difference.

As a target for this method, we collect Japanese-English bilingual data that is either in the public domain, or available under the Creative Commons license, as described in Section A.4 and use it to create a broad-coverage and freely available

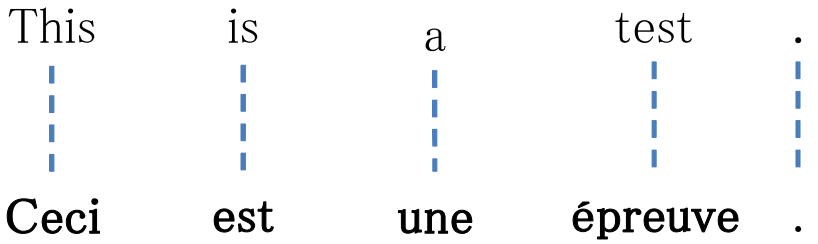

Figure 24. Example of alignment for a language pair with similar word order and grammar (e.g., English-French).

Japanese paraphrase database.[7] Finally, we perform an analysis of the database in Section A.5, and find that it achieves a standard of accuracy similar to that of previously reported paraphrasing resources.

A.2 Extracting paraphrases

We extract paraphrases using Bannard and Callison-Burch's bilingual pivoting method [Bannard and Callison-Burch, 2005]. This method is a general-purpose paraphrase extraction method, with the intuition that two English strings e_1 and e_2 that translate to the same foreign string f can often be assumed to have the same meaning.

In this work, instead of English, we extract paraphrases for Japanese, over Japanese-English bilingual parallel corpora. In Japanese-English, we can thus pivot over e and extract $\langle j_1, j_2 \rangle$ as a pair of paraphrases, as illustrated in Figure 25. We estimate the conditional paraphrase probability $P(j_2|j_1)$ by marginalizing over all shared English translations e:

$$P(j_2|j_1) = \sum_e P(j_2|e)P(e|j_1) \tag{30}$$

To calculate these pivoted pairs and probabilities, we need to calculate the conditional probabilities $P(j_2|e)$ and $P(e|j_1)$. This is done by first extracting

[7] http://ahclab.naist.jp/resource/jppdb/

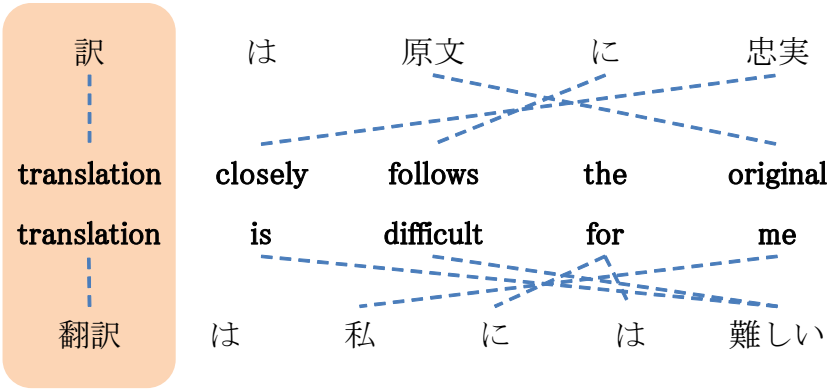

Figure 25. Phrase-based paraphrases are extracted via bilingual pivoting.

phrase pairs from a Japanese-English corpus according to the standard phrase extraction algorithm of phrase-based statistical machine translation [Koehn et al., 2003]. For example, translation probability $P(e|j)$ is calculated according to maximum likelihood estimation based on the counts of each phrase pair $c(e, j)$ and its constituent phrases:

$$P(e|j) = \frac{c(e, j)}{\sum_e c(e, j)} \tag{31}$$

This method has been shown to accurately extract a diverse set of paraphrases in past research [Bannard and Callison-Burch, 2005].

A.3 Syntactic Preprocessing

In order to use the previously described method, it is necessary to acquire phrase alignments in parallel corpus as pivots between English phrases. In general, these automatic alignments are produced in an unsupervised manner with the GIZA++ toolkit [Och and Ney, 2003]. However, for languages with greatly different syntax and word order, standard alignment with GIZA++ has worse performance in comparison to languages with more similar syntax and word order.

In this section, we help ameliorate this problem using recent syntactic preprocessing approaches to statistical machine translation. Specifically, we use the Head Finalization (HF; [Isozaki et al., 2010]) syntactic preprocessing method to change the English sentence to a syntactic structure similar to Japanese before

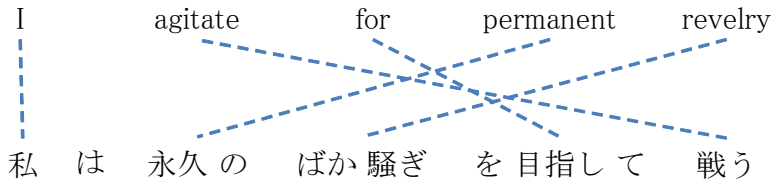

Figure 26. Example of alignment in standard English-Japanese.

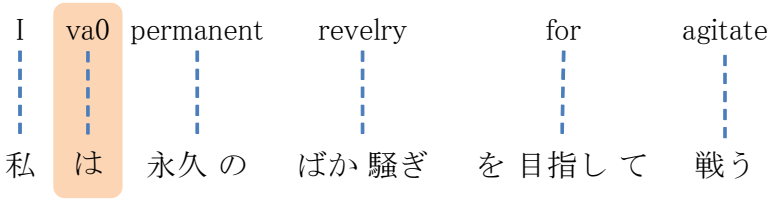

Figure 27. Example of alignment in head-finalized English-Japanese.

running alignment and phrase extraction.

In English-Japanese translation, one of the most serious problems is the difference of grammar and word ordering between the two languages. Isozaki et al. proposed the HF preprocessing method, which helps reduce this syntactic divergence and increase the accuracy of statistical machine translation results [Isozaki et al., 2010].

The main element of the method lies in reordering English into a similar order to Japanese. This is done by using a syntactic parser to parse the English sentence, then moving the head of all phrases to the end of the phrase, transforming English into head-final order, similar to Japanese. In addition, the HF algorithm considers the fact that Japanese contains explicit case markers after the subject and object which are not present in English, and inserts pseudo-words that correspond to these particles into the head-finalized English. As these will be aligned into Japanese particles, it can be expected that performing this processing will also improve the accuracy of acquiring paraphrases for these particles. We show examples of alignment in standard English-Japanese and head-finalized English-Japanese in Figure 26 and 27.

Table 25. The details of corpus

	Sentences	Words (en)	Words (ja)	Licenses
Tanaka [Tanaka, 2001]	150k	1.4M	2.1M	Public Domain
AOZORA [Utiyama and Takahashi, 2003]	108k	1.6M	2.5M	Public Domain
Common Crawl [Smith et al., 2013]	821k	13.8M	22.0M	Public Domain
WWWJDIC [Breen, 2014]	373k	866k	373k	CC BY-SA 3.0[4]
Kyoto Wiki [Utiyama and Takahashi, 2011]	440k	11.5M	11.8M	CC BY-SA 3.0[4]
Total	1.9M	29.2M	38.3M	CC BY-SA 3.0[4]

Table 26. The details of the phrase table

Phrases	67.1M
Alignment	GIZA++ [Och and Ney, 2003]
Tokenization (en)	Stanford Parser [Socher et al., 2013]
Tokenization (ja)	Kytea [Neubig et al., 2011]
Max phrase length	7 words

A.4 PPDB : Japanese

Based on this data, we extracted Japanese paraphrases according to the proposed method. We extract alignments from a 1.9M sentence Japanese-English parallel corpus, the details of which are shown in Tables 25 and 26. For alignment, we use GIZA++, with the English side being pre-processed with HF as mentioned in the previous section. For tokenization, we use the Stanford Parser[8] [Socher et al., 2013] for English and KyTea[9] [Neubig et al., 2011] for Japanese.

In paraphrasing, we chose paraphrases where the conditional probability of the target is less than 1% to reduce the number of extracted paraphrases with low probability. The total number of extracted Japanese paraphrases were 10.5M pairs. We analyze and evaluate the paraphrases, the detailed results of which are below.

Table 27. Examples of paraphrases with their rough English gloss

Seed	Paraphrases
メンバー	メンバ, 一族, 一員, 員, 会員, 加盟, 会員の
member	*member, family, a member (of), member, membership, member (join), member's*
魏志 倭人 伝 に	魏志 倭人 伝 の 記述 に, 魏志 倭人 に, 魏志 倭人 伝 に は
in the Gishi-wajin-den	*According to a description in the Gishi-wajin-den, in the Gishi-wajin, in Gishi-wajin-den*
論争	の 論争, 争議, 紛争, 争い, 討論, 議論
argue	*'s dispute, dispute, conflict, controversy, discussion, argue*
突如	とつぜん, 急 に, 不意 に, 突然
suddenly	*all of a sudden, hastily, abruptly, sudden*
で	に より, に よって, に よる, に, の
by	*because of, depending on, according to, to, 's*

[8] http://nlp.stanford.edu/software/lex-parser.shtml

[9] http://www.phontron.com/kytea/

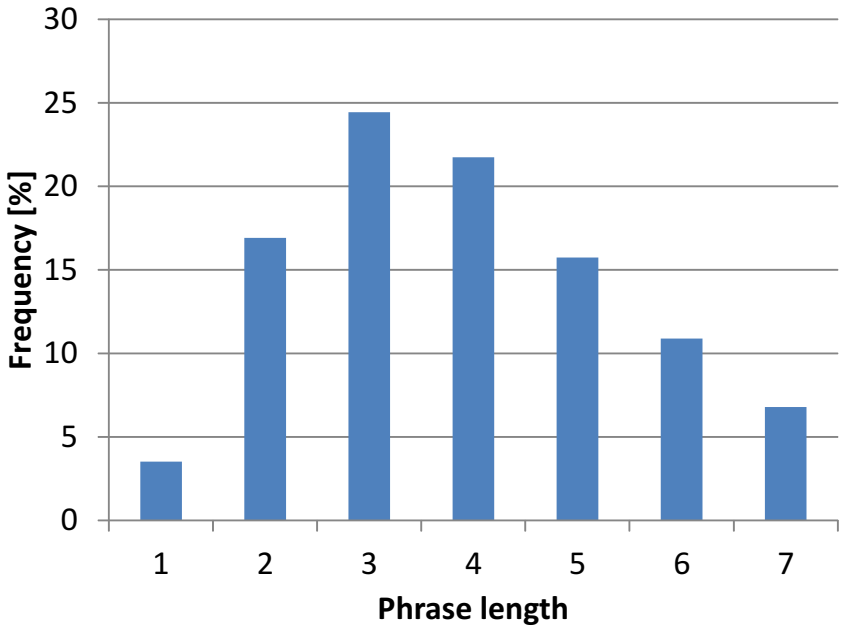

Figure 28. Histogram of every phrase length in the acquired paraphrases.

A.5 Analysis

In the analysis, we evaluated mean phrase length of paraphrases, number of paraphrases provided for single words (as opposed to phrases), number of paraphrases of particles, and a histogram of the phrase lengths of paraphrases.

The mean phrase length of paraphrases is 3.42 words. The number of paraphrases provided for words is 60.8k, and the number of paraphrases for phrases of length two or more is 986k. We were able to acquire 134 paraphrases of Japanese particles in this method. We show examples of the acquired paraphrases in Table 21.

We show a histogram of every phrase length of paraphrases in Figure 28.

A.6 Evaluation

We evaluate the paraphrases under the same conditions as previous work [Bannard and Callison-Burch, 2005], with the details as follows. We substituted candidate paraphrases into 24 sentences which contained the original phrase, creating a total of 85 unique sentences through substitution. We had 3 native

Table 28. Evaluation of the acquired paraphrases

	Correct rate [%]
Meaning	84.7
Grammar	55.3
Both	45.8

Japanese speakers produce judgments as to whether the new sentences preserved the meaning of the original phrase and as to whether they remained grammatical. Paraphrases that were judged to preserve both meaning and grammaticality were considered to be correct, and examples which failed on either judgment were considered to be incorrect. We show results of the judgement in Table 28.

The accuracy of paraphrases is 45.8%, almost same with previous work [Bannard and Callison-Burch, 2005]. The inter-annotator agreement for these judgements was measured at $\kappa = 0.60$ [Fleiss, 1971], which is conventionally interpreted as "moderate" agreement [Landis and Koch, 1977].

Ignoring the constraint that the new sentences remain grammatically correct, these paraphrases were judged to have the correct meaning 84.7% of the time. This indicates that the paraphrases are semantically correct, but may vary in their syntactic categories or contexts. These tendencies are similar to Bannard and Callison-Burch [Bannard and Callison-Burch, 2005].

A.7 Related Works

With regards to related resources for paraphrasing created using bilingual data, there are English paraphrase data extracted by the same method [Ganitkevitch et al., 2013]. In their work, they use English-French, English-Spanish and other European language bilingual parallel corpora. Finally, 16.7M paraphrases were extracted from 1G sentences of parallel data. Fujita et al. [Fujita et al., 2012] also extract 28M Japanese paraphrases from 3.2M sentence pairs of Japanese-English patent translation data, although these paraphrases are limited to the patent domain and not publicly available as a resource.

Considering work on Japanese paraphrasing not limited to those extracted from parallel data, there is Japanese honorifics paraphrase data [Ohtake and Yamamoto, 2001], action word paraphrases [Nakagawa and Masuda, 2004] and

others. The Japanese honorifics paraphrase data offers 130k paraphrases gathered from 50k sentences of monolingual data covering honorifics. In the action word paraphrase data, 1.1k paraphrases covering action words such as verbs are included.

In these related works, various types of paraphrases are suggested, but there are few resources freely available. In addition, there are many paraphrase resources that have some kind of theme, but there are few large-scale and general paraphrase resources. In comparison, our paraphrase data is large, general, and freely available.

A.7.1 Conclusion

In the end, we were able to acquire 1.2M paraphrases from 1.9M sentences of our bilingual parallel corpus. Our paraphrase data is larger than some previous works that created Japanese paraphrases [Ohtake and Yamamoto, 2001, Nakagawa and Masuda, 2004]. And the proposed method was able to acquire paraphrase data that is large, covers several domains, and is high quality. We hope that our paraphrase data will be able to contribute to future studies that require paraphrases in the Japanese language.

[4]CC BY-SA 3.0 : Creative Commons Attribution-ShareAlike 3.0 Unported License

References

[Banchs, 2012] Rafael E. Banchs. Movie-DiC: a movie dialogue corpus for research and development. In *Proc. of ACL*, pages 203–207, 2012.

[Banchs and Li, 2012] Rafael E. Banchs and Haizhou Li. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proc. of ACL*, pages 37–42, 2012.

[Bannard and Callison-Burch, 2005] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL*, pages 597–604, 2005.

[Barzilay and Lee, 2003] Regina Barzilay and Lillian Lee. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proc. of NAACL HLT*, pages 16–23, 2003. doi: 10.3115/1073445.1073448. URL `http://dx.doi.org/10.3115/1073445.1073448`.

[Basak et al., 2007] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10):203–224, 2007.

[Beňuš et al., 2014] Štefan Beňuš, Agustín Gravano, Rivka Levitan, Sarah Ita Levitan, Laura Willson, and Julia Hirschberg. Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, 71:3–14, 2014.

[Bessho et al., 2012] Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. Dialog system using real-time crowdsourcing and twitter large-scale corpus. In *Proc. of SIGDIAL*, pages 227–231, 2012.

[Bickmore and Cassell, 2000] Timothy Bickmore and Justine Cassell. how about this weather? ” social dialogue with embodied conversational agents. In *Proc. of AAAI Fall Symposium on Socially Intelligent Agents*, pages 4–8, 2000.

[Bickmore et al., 2011] Timothy Bickmore, Laura Pfeifer, and Daniel Schulman. Relational agents improve engagement and learning in science museum visitors. In *International Workshop on Intelligent Virtual Agents*, pages 55–67. Springer, 2011.

[Bird et al., 2008] Steven Bird, Ewan Klein, Edward Loper, and Jason Baldridge. Multidisciplinary instruction with the natural language toolkit. In *Proc. of TeachCL*, pages 62–70, 2008.

[Bond et al., 2009] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Enhancing the Japanese wordnet. In *Proc. of ALR*, pages 1–8, 2009.

[Breen, 2014] Jim Breen. WWWJDIC Japanese dictionary server user guide. `http://www.csse.monash.edu.au/~jwb/wwwjdicinf.html`, 2014.

[Brennan and Clark, 1996] Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482, 1996.

[Brill and Moore, 2000] Eric Brill and Robert C Moore. An improved error model for noisy channel spelling correction. In *Proc. of ACL*, pages 286–293, 2000.

[Brown et al., 1990] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

[Brown et al., 1993] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[Callison-Burch et al., 2006] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proc. of NAACL*, pages 17–24, 2006. URL `http://www.aclweb.org/anthology/N/N06/N06-1003`.

[Campbell and Scherer, 2010] Nick Campbell and Stefan Scherer. Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In *INTERSPEECH*, pages 2546–2549, 2010.

[Cassell et al., 1999] Justine Cassell, Timothy Bickmore, Mark Billinghurst, Lee Campbell, Kenny Chang, Hannes Vilhjálmsson, and Hao Yan. Embodiment

in conversational interfaces: Rea. In *Proc. of SIGCHI conference on Human Factors in Computing Systems*, pages 520–527. ACM, 1999.

[Chung and Pennebaker, 2007] Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, pages 343–359, 2007.

[Cohen, 1988] Jacob Cohen. Statistical power analysis for the behavioral sciences. 2nd edn. hillsdale, new jersey: L, 1988.

[Coulston et al., 2002] Rachel Coulston, Sharon Oviatt, and Courtney Darves. Amplitude convergence in children's conversational speech with animated personas. In *Proc. of ICSLP*, volume 4, pages 2689–2692, 2002.

[Dagan et al., 1999] Ido Dagan, Lillian Lee, and Fernando CN Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3): 43–69, 1999.

[Danescu-Niculescu-Mizil et al., 2011] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM, 2011.

[De Saeger et al., 2009] Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. Large scale relation acquisition using class dependent patterns. In *Proc. of IEEE International Conference on Data Mining*, pages 764–769. IEEE, 2009.

[Elzer et al., 1994] Stephanie Elzer, Jennifer Chu-Carroll, and Sandra Carberry. Recognizing and utilizing user preferences in collaborative consultation dialogues. In *Proc. of Fourth International Conference on User Modeling*, volume 19, page 24, 1994.

[Engelbrech et al., 2009] Klaus-Peter Engelbrech, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. Modeling user satisfaction with hidden markov model. In *Proc. of SIGDIAL*, pages 170–177, 2009.

[Engelbrecht and Möller, 2010] Klaus-Peter Engelbrecht and Sebastian Möller. A user model to predict user satisfaction with spoken dialog systems. In *Proc. of IWSDS*, pages 150–155, 2010.

[Fandrianto and Eskenazi, 2012] Andrew Fandrianto and Maxine Eskenazi. Prosodic entrainment in an information-driven dialog system. In *Proc. of INTERSPEECH*, pages 342–345, 2012.

[Fleiss, 1971] Joseph.L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[Fujita et al., 2012] Atsushi Fujita, Pierre Isabelle, and Roland Kuhn. Enlarging paraphrase collections through generalization and instantiation. In *Proc. of EMNLP-CoNLL*, EMNLP-CoNLL '12, pages 631–642, 2012. URL `http://dl.acm.org/citation.cfm?id=2390948.2391019`.

[Ganitkevitch et al., 2013] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proc. of NAACL HLT*, pages 758–764, 2013. URL `http://cs.jhu.edu/~ccb/publications/ppdb.pdf`.

[Gosling et al., 2003] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

[Gravano et al., 2012] Agustín Gravano, Julia Hirschberg, and Štefan Beňuš. Affirmative cue words in task-oriented dialogue. *COLING*, 38(1):1–39, 2012.

[Grice, 1975] H Paul Grice. Logic and conversation. *1975*, pages 41–58, 1975.

[Hajdinjak and Mihelič, 2006] Melita Hajdinjak and France Mihelič. The PARADISE evaluation framework: Issues and findings. *Computational Linguistics*, 32(2): 263–272, 2006.

[Hara et al., 2010] Sunao Hara, Norihide Kitaoka, and Kazuya Takeda. Estimation method of user satisfaction using n-gram-based dialog history model for spoken dialog system. In *Proc. of LREC*, pages 78–83, 2010.

[Hashimoto et al., 2011] Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi. Extracting paraphrases from definition sentences on the web. In *Proc. of ACL : Human Language Technologies*, pages 1087–1097, 2011. URL `http://www.aclweb.org/anthology/P11-1109`.

[Hermjakob et al., 2002] Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu. Natural language based reformulation resource and wide exploitation for question answering. In *Proc. of TREC*, volume 90, page 91, 2002.

[Higashinaka et al., 2009] Ryuichiro Higashinaka, Noriaki Kawamae, Kohji Dohsaka, and Hideki Isozaki. Using collaborative filtering to predict user utterances in dialogue. In *Proc. of IWSDS*, 2009.

[Higashinaka et al., 2010] Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. Modeling user satisfaction transitions in dialogues from overall ratings. In *Proc. of SIGDIAL*, pages 18–27, 2010.

[Hiraoka et al., 2014] Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Construction and analysis of a persuasive dialogue corpus. In *Proc. of IWSDS*, 2014.

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

[Inui and Fujita, 2004] Kentaro Inui and Atsushi Fujita. A survey on paraphrase generation and recognition. *Journal of Natural Language Processing*, 11(5):151–198, 2004.

[Isard et al., 2006] Amy Isard, Carsten Brockmann, and Jon Oberlander. Individuality and alignment in generated dialogues. In *Proc. of INLG*, pages 25–32, 2006. ISBN 1-932432-72-8. URL `http://dl.acm.org/citation.cfm?id=1706269.1706277`.

[Isozaki et al., 2010] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. Head finalization: A simple reordering rule for SOV languages. In *Proc. of MATR*, pages 244–251, 2010. URL `http://www.aclweb.org/anthology/W10-1736`.

[Kawahara et al., 2015] Tatsuya Kawahara, Takashi Yamaguchi, Miki Uesato, Koichiro Yoshino, and Katsuya Takanashi. Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening. In *APSIPA*, pages 392–395. IEEE, 2015.

[Kim et al., 2010] Kyungduk Kim, Cheongjae Lee, Donghyeon Lee, Junhwi Choi, Sangkeun Jung, and Gary Geunbae Lee. Modeling confirmations for example-based dialog management. In *Proc. of SLT*, pages 324–329, 2010.

[Koehn, 2004] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395, 2004.

[Koehn, 2009] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[Koehn et al., 2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proc. of NAACL HLT*, volume 1 of *The 2013 NAACL : HLT*, pages 48–54, 2003. doi: 10.3115/1073445.1073462. URL `http://dx.doi.org/10.3115/1073445.1073462`.

[Landis and Koch, 1977] J. Richard. Landis and G. Gary Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[Lee et al., 2009] Cheongjae Lee, Sungjin Lee, Sangkeun Jung, Kyungduk Kim, Donghyeon Lee, and Gary Geunbae Lee. Correlation-based query relaxation for example-based dialog modeling. In *Proc. of ASRU*, pages 474–478, 2009.

[Levitan, 2013] Rivka Levitan. Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *Proc. of HLT-NAACL*, pages 84–90, 2013.

[Levitan et al., 2015] Rivka Levitan, Stefan Benus, Agust¥in Gravano, and Julia Hirschberg. Entrainment and turn-taking in human-human dialogue. In *Proc. of AAAI*, pages 44–51, 2015.

[MacKay and Peto, 1995] David JC MacKay and Linda C Bauman Peto. A hierarchical dirichlet language model. *Natural language engineering*, 1(03):289–308, 1995.

[Mairesse and Walker, 2011] François Mairesse and Marilyn A Walker. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488, 2011.

[Matsuyama et al., 2014] Yoichi Matsuyama, Alexandros Papangelis, Ran Zhao, Justine Cassell, et al. 2 者会話におけるラポール形成・維持・崩壊の計算モデル. *SIG-SLUD*, 4(02):13–18, 2014.

[Meguro et al., 2010] Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. Controlling listening-oriented dialogue using partially observable markov decision processes. In *Proc. of COLING*, pages 761–769. Association for Computational Linguistics, 2010.

[Metze et al., 2009] Florian Metze, Roman Englert, Udo Bub, Felix Burkhardt, and Joachim Stegmann. Getting closer: tailored human-computer speech dialog. *Universal Access in the Information Society*, 8(2):97–108, 2009. ISSN 1615-5289. doi: 10.1007/s10209-008-0133-0. URL `http://dx.doi.org/10.1007/s10209-008-0133-0`.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Miyazaki et al., 2016] Chiaki Miyazaki, Toru Hirano, and Ryuichiro Higashinaka Yoshihiro Matsuo. Towards an entertaining natural language generation system: Linguistic peculiarities of japanese fictional characters. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 319, 2016.

[Murao et al., 2003] Hiroya Murao, Nobuo Kawaguchi, Shigeki Matsubara, Yukiko Yamaguchi, and Yasuyoshi Inagaki. Example-based spoken dialogue system using WOZ system log. In *Proc. of SIGDIAL*, pages 140–148, 2003.

[Nakagawa and Masuda, 2004] Hiroshi Nakagawa and Hidetaka Masuda. Extracting paraphrases of japanese action word of sentence ending part from web and mobile news articles. In *Asia Information Retrieval Symposium*, pages 94–105. Springer, 2004.

[Natale, 1975] Michael Natale. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790, 1975.

[Navarro, 2001] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001. ISSN 0360-0300.

[Nenkova et al., 2008] Ani Nenkova, Agustín Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In *Proc. of ACL*, pages 169–172. Association for Computational Linguistics, 2008.

[Neubig et al., 2011] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proc. of 49th ACL HLT*, volume 2 of *HLT '11*, pages 529–533, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6. URL `http://dl.acm.org/citation.cfm?id=2002736.2002841`.

[Neubig et al., 2012] Graham Neubig, Yuya Akita, Shinsuke Mori, and Tatsuya Kawahara. A monotonic statistical machine translation approach to speaking style transformation. *Computer Speech & Language*, 26(5):349–370, 2012.

[Niederhoffer and Pennebaker, 2002] Kate G Niederhoffer and James W Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, 2002.

[Nio et al., 2012] Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, Mirna Adriani, and Satoshi Nakamura. Developing non-goal dialog system based on examples of drama television. In *Proc. of IWSDS*, pages 315–320, 2012.

[Nio et al., 2014a] Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Conversation dialog corpora from drama television and movie scripts. In *Proc. of O-COCOSDA*, pages 144–148, 2014a.

[Nio et al., 2014b] Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Improving the robustness of example-based dialog retrieval using recursive neural network paraphrase identification. In *Proc. of SLT*, pages 306–311, 2014b.

[Och and Ney, 2002] Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of ACL*, pages 295–302. Association for Computational Linguistics, 2002.

[Och and Ney, 2003] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March 2003. ISSN 0891-2017. doi: 10.1162/089120103321337421. URL `http://dx.doi.org/10.1162/089120103321337421`.

[Och and Ney, 2004] Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4): 417–449, 2004.

[Ogan et al., 2012] Amy Ogan, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. Rudeness and rapport: Insults and learning gains in peer tutoring. In *International Conference on Intelligent Tutoring Systems*, pages 11–21. Springer, 2012.

[Ohtake and Yamamoto, 2001] Kiyonori Ohtake and Kazuhide Yamamoto. Paraphrasing honorifics. In *Proc. of NLPRS Post-Conference Workshop*, pages 13–20, 2001.

[Pardo, 2006] Jennifer S Pardo. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393, 2006.

[Reitter and Moore, 2007] David Reitter and Johanna D Moore. Predicting success in dialogue. In *Proc. of ACL*, pages 808–815, 2007.

[Riesa et al., 2011] Jason Riesa, Ann Irvine, and Daniel Marcu. Feature-rich language-independent syntax-based alignment for statistical machine translation. In *Proc. of EMNLP*, pages 497–507, 2011.

[Schmitt et al., 2011] Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. Modeling and predicting quality in spoken human-computer interaction. In *Proc. of SIGDIAL*, pages 173–184, 2011.

[Smith et al., 2013] Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. In *Proc. of Annual Meeting of the Association for Computational Linguistics*, pages 1374–1383, 2013.

[Socher et al., 2013] Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. Parsing with compositional vector grammars. In *Pric. of ACL*, pages 455–465, 2013.

[Spencer-Oatey, 2005] Helen Spencer-Oatey. (im) politeness, face and perceptions of rapport: unpackaging their bases and interrelationships, 2005.

[Takamura et al., 2005] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proc. of ACL*, pages 133–140, 2005.

[Takeuchi et al., 2007] Shota Takeuchi, Tobias Cincarek, Hiromichi Kawanami, Hiroshi Saruwatari, and Kiyohiro Shikano. Construction and optimization of a question and answer database for a real-environment speech-oriented guidance system. In *Proc. of O-COCOSDA*, pages 149–154, 2007.

[Takezawa et al., 2002] Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of LREC*, pages 147–152, 2002.

[Tanaka, 2001] Yasuhito Tanaka. Compilation of a multilingual parallel corpus. *Proc. of 2001 Conference of the Pacific Association for Computational Linguistics*, pages 265–268, 2001.

[Teh et al., 2006] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

[Teshigawara and Kinsui, 2012] Mihoko Teshigawara and Satoshi Kinsui. Modern Japanese "Role Language"(Yakuwarigo): fictionalised orality in Japanese literature and popular culture. In *Sociolinguistic Studies Vol 5-1*, 2012.

[Tickle-Degnen and Rosenthal, 1990] Linda Tickle-Degnen and Robert Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293, 1990.

[Ultes and Minker, 2014] Stefan Ultes and Wolfgang Minker. Interaction quality estimation in spoken dialogue systems using hybrid-hmms. In *Proc. of SIGDIAL*, page 208–217, 2014. URL `http://www.aclweb.org/anthology/W14-4328`.

[Utiyama and Takahashi, 2003] Masao Utiyama and Mayumi Takahashi. English-Japanese translation alignment data. `http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/align/`, 2003.

[Utiyama and Takahashi, 2011] Masao Utiyama and Mayumi Takahashi. Japanese-English bilingual corpus of Wikipedia's Kyoto articles. `http://alaginrc.nict.go.jp/WikiCorpus/index_E.html`, 2011.

[Vardoulakis et al., 2012] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L Sidner, and Timothy Bickmore. Designing relational agents as long term social companions for older adults. In *Proc. of IVA*, pages 289–302. Springer, 2012.

[Walker et al., 1997] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Proc. of EACL*, pages 271–280, 1997.

[Ward and Litman, 2007] Arthur Ward and Diane Litman. Measuring convergence and priming in tutorial dialog. In *University of Pittsburgh*, 2007.

[Wärnestål et al., 2007] Pontus Wärnestål, Lars Degerstedt, and Arne Jönsson. Pcql: A formalism for human-like preference dialogues. *Knowledge and Reasoning in Practical Dialogue Systems*, page 46, 2007.

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

[Wu et al., 2016] Xianchao Wu, Kazushige Ito, Katsuya Iida, Kazuna Tsuboi, and Momo Klyen. りんな: 女子高生人工知能. In *Proc. of JNLP*, pages 306–309, 2016.

[Xu et al., 2012] Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for style. In *Proc. of ACL*, pages 2899–2914, 2012. URL `http://www.aclweb.org/anthology/C12-1177`.

[Yang et al., 2010] Zhaojun Yang, Baichuan Li, Yi Zhu, Irwin King, Gina-Anne Levow, and Helen M. Meng. Collaborative filtering model for user satisfaction prediction in spoken dialog system evaluation. In *Proc. of SLT*, pages 472–477, 2010.

[Yoshino and Kawahara, 2015] Koichiro Yoshino and Tatsuya Kawahara. Conversational system for information navigation based on pomdp with user focus tracking. *Computer Speech & Language*, 34(1):275–291, 2015.

[Young et al., 2013] Steve Young, Milica Gasic, Blaise Thomson, and Jason Williams. POMDP-based statistical spoken dialogue systems: a review. *Proc. of IEEE*, 101(5):1160–1179, 2013.

[Yu et al., 2016a] Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alex I Rudnicky. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 55, 2016a.

[Yu et al., 2016b] Zhou Yu, Ziyu Xu, Alan W Black, and Alex I Rudnicky. Strategy and policy learning for non-task-oriented conversational systems. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 404, 2016b.

Publication lists

Journals

1. 水上 雅博, Lasguido Nio, 木付 英士, 野村 敏男, Graham Neubig, 吉野 幸一郎, Sakriani Sakti, 戸田 智基, 中村 哲.
   “快適度推定に基づく用例ベース対話システム”
   人工知能学会論文誌, 31-1. 2016 年 1 月.

2. 東中 竜一郎, 船越 孝太郎, 荒木 雅弘, 塚原 裕史, 小林 優佳, 水上雅博
   “テキストチャットを用いた雑談対話コーパスの構築と対話破綻の分析”
   自然言語処理学会論文誌, Vol.23 No.1. 2016 年 11 月.

3. 杉山 享志朗, 水上 雅博, Graham Neubig, 吉野 幸一郎, 鈴木 優, 中村 哲.
   “言語横断質問応答に適した機械翻訳評価尺度の調査”
   言語処理学会論文誌, Vol.23 No.5. 2016 年 12 月.

Conference papers (peer reviewed)

1. Masahiro Mizukami, Koichiro Yoshino, Graham Neubig, David Traum, Satoshi Nakamura.
   “Analyzing the Effect of Entrainment on Dialogue Acts”
   17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. September 2016.

2. Yuiko Tsunomori, Graham Neubig, Takuya Hiraoka, Masahiro Mizukami, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura.
   “A Dialog System to Detect Deception”
   7th International Workshop on Spoken Dialog Systems (IWSDS). January 2016.

3. Masahiro Mizukami, Hideaki Kizuki, Toshio Nomura, Graham Neubig, Koichiro Yoshino, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura.
   ”Adaptive Selection from Multiple Reponse Candidates in Example-based Dialogue”
   2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). December 2015.

4. Kyoshiro Sugiyama, Masahiro Mizukami, Graham Neubig, Koichiro Yoshino, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura.
   "An Investigation of Machine Translation Evaluation Metrics in Cross-lingual Question Answering"
   10th Workshop on Statistical Machine Translation (WMT). September 2015.

5. Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi.
   "Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems"
   Conference on Empirical Methods in Natural Language Processing (EMNLP). September 2015.

6. Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Mizukami, Hiroshi Tsukahara, Yuka Kobayashi Masahiro Araki.
   "Analyzing dialogue breakdowns in chat-oriented dialogue systems"
   In Proc. Errare 2015. September 2015.

7. Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, Masahiro Mizukami.
   "Towards taxonomy of errors in chat-oriented dialogue systems"
   16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. September 2015.

8. Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura.
   "Linguistic Individuality Transformation for Spoken Language"
   6th International Workshop on Spoken Dialog Systems (IWSDS). January 2015.

9. Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura.
   "Building a Free, General-Domain Paraphrase Database for Japanese"
   The 17th Oriental COCOSDA Conference. September 2014.

10. Graham Neubig, Shinsuke Mori, Masahiro Mizukami.

“A Framework and Tool for Collaborative Extraction of Reliable Information”
Workshop on Language Processing and Crisis Information (LPCI). October 2013.

Conference papers (without review)

1. 水上 雅博, 吉野 幸一郎, Graham Neubig, 中村 哲.
“エントレインメント分析に基づく用例選択モデルの提案”
人工知能学会言語・音声理解と対話処理研究会 (SIG-SLUD). 東京. 2016 年 10 月. **若手奨励賞**.

2. 杉山 享志朗, 水上 雅博, 吉野 幸一郎, 田中 宏季, 鈴木 優, 中村 哲.
“対話履歴との矛盾を考慮した発話選択”
人工知能学会言語・音声理解と対話処理研究会 (SIG-SLUD). 東京. 2016 年 10 月.

3. 水上 雅博, Graham Neubig, 吉野 幸一郎, Sakriani Sakti, 鈴木 優, 中村 哲.
“快適度推定に基づく用例ベース対話システム”
言語処理学会第 22 回年次大会 (NLP2016). 宮城. 2016 年 3 月.

4. 石川 葉子, 平岡 拓也, 水上 雅博, 吉野 幸一郎, Graham Neubig, 中村 哲.
“対話状態推定のための外部知識ベースを利用した意味的素性の提案”
情報処理学会 第 109 回音声言語情報処理研究会 (SIG-SLP). 2015 年 12 月. **学生ポスター発表賞**.

5. 水上 雅博, 杉山 享志朗, Graham Neubig, 吉野 幸一郎, Sakriani Sakti, 中村 哲.
“RNN を用いた対話破綻検出器の構築”
人工知能学会言語・音声理解と対話処理研究会 (SIG-SLUD). 東京. 2015 年 10 月.

6. 水上 雅博, Lasguido Nio, Graham Neubig, 吉野 幸一郎, Sakriani Sakti, 戸田 智基, 中村 哲.
“快適度推定に基づく用例ベース対話システム”
人工知能学会言語・音声理解と対話処理研究会 (SIG-SLUD). 東京. 2015 年 10 月.

7. 水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“用例ベース対話システムにおける用例の評価値推定”
2015 年度人工知能学会全国大会 (JSAI2015). 北海道. 2015 年 5 月.

8. 杉山 享志朗, 水上 雅博, Graham Neubig, 吉野 幸一郎, Sakriani Sakti, 戸田 智基, 中村 哲.
“言語横断質問応答に適した機械翻訳評価尺度の検討”
情報処理学会 第 223 回自然言語処理研究会 (SIG-NL). 広島. 2015 年 9 月.

9. 水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“言語的個人性変換における言語モデルの適応と分析”
第 72 回 人工知能学会 音声・言語理解と対話処理研究会（SIG-SLUD）. 神奈川. 2014 年 12 月.

10. 水上 雅博, 木付 英士, 野村 敏男, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“対話システムにおける応答文選択法の検討”
日本音響学会 2014 年度秋季研究発表会 (ASJ). 北海道. 2014 年 9 月.

11. 水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“特徴的話者を対象とした言語的個人性変換”
情報処理学会 第 216 回自然言語処理研究会 (SIG-NL). 東京. 2014 年 5 月.

12. 水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“日本語言い換えデータベースの構築と言語的個人性変換への応用”
言語処理学会第 20 回年次大会 (NLP). 北海道. 2014 年 3 月.

13. 水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“話し言葉における言語情報の個人性変換手法の拡張と評価”
日本音響学会 2013 年秋季研究発表会 (ASJ). 豊橋. 2013 年 9 月.

14. 水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“話し言葉における言語情報の個人性変換における変換辞書拡張”
電子情報通信学会音声研究会 (SP). 千葉. 2013 年 9 月.

15. 水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“話し言葉の書き起こし文章の話者性の変換”
2013 年度人工知能学会全国大会 (JSAI). 富山. 2013 年 6 月.

16. 水上 雅博.
“雑談対話システムの実践（と裏話）”
NLP 若手の会 (YANS) 第 11 回シンポジウム. 和歌山. 2016 年 8 月

17. 水上 雅博, Lasguido Nio, Graham Neubig, 吉野 幸一郎, Sakriani Sakti, 戸田 智基, 中村 哲.
“快適度推定に基づく用例ベース対話システム”
NLP 若手の会 (YANS) 第 10 回シンポジウム. 石川. 2015 年 9 月.

18. 水上 雅博, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲.
“日本語言い換えデータベースの改善と評価”
第 9 回自然言語処理若手の会シンポジウム (YANS). 神奈川. 2014 年 9 月.