Advancing Dialogue Systems through Corpus Construction Focusing on User Internal States and External Knowledge

Takashi Kodama

February 2024

Abstract

Dialogue systems are computational systems designed to interact with humans using natural language. Their considerable practical utility has made their development a critical area of research in the realm of natural language processing. In recent years, end-to-end learning methods based on deep neural networks have attracted much attention in the field of dialogue systems. Such systems are trained to generate responses from dialogue contexts, leveraging a large-scale dataset of context-response pairs. While this method has facilitated a more adaptable generation of responses, it concurrently presents unresolved challenges.

One such challenge is the understanding of *user internal states*. The user internal states represent the user's underlying states behind their utterance. In this thesis, we primarily focus on the user's *knowledge*, *interest*, and *willingness*. Understanding these states can help dialogue systems identify user intentions and generate responses effectively. However, existing research in dialogue systems has not tracked these user internal states in detail.

The second challenge involves the utilization of *external knowledge*. External knowledge refers to information external to the dialogue system, obtainable through means like internet searches. By leveraging external knowledge, dialogue systems can generate more specific and informative responses. Prior research on external knowledge-grounded dialogue systems has primarily concentrated on the search for external knowledge and its incorporation into responses, often overlooking the underlying engagingness of the dialogue itself.

This thesis addresses the two challenges: user internal states and external knowledge. To tackle these challenges, we adopt an approach through the construction of dialogue corpora to improve dialogue systems. This corpus-driven

approach not only contributes to the advancement of dialogue systems via dataset utilization but also aids in the understanding of human behavior through corpus analysis.

First, we study dialogue management based on user internal states. We develop a rule-based dialogue system and collect dialogues between the system and humans. Then, we annotate all the user utterances with three user internal states: knowledge, interest, and willingness. Using the constructed corpus, we train classifiers to estimate each user internal state. Finally, we propose a system designed to modify responses based on the estimated internal states of the user. Our experimental results show that the proposed system can provide more natural responses than the baseline system.

Second, we tackle dialogue response generation based on external knowledge. We first construct a dialogue collection framework capable of constructing large-scale, human-to-human dialogue datasets in Japanese. Leveraging this framework, we construct a Japanese external knowledge-grounded dialogue dataset. Employing the constructed dataset, we present a strong baseline model that simultaneously selects external knowledge and generates responses based on it.

Third, we analyze the engagingness of external knowledge-grounded responses. We annotate every entity in external knowledge-grounded responses, labeling whether the entity is derived from external knowledge sources or from the speaker's inherent knowledge and opinions. Our analysis of this annotated corpus reveals that information derived from the speaker plays a significant role in enhancing the engagingness of the response.

Last, we study response generation based on both user internal states and external knowledge. We construct a movie recommendation dialogue dataset annotated at the entity level focusing on two user internal states: knowledge and interest. By employing movie enthusiasts as recommenders, we realize a more engaging dialogue collection. Furthermore, using this dataset, we propose a model that generates responses based on both user internal states and external knowledge utilizing Chain-of-Thought prompting.

Acknowledgments

研究室配属から今日に至るまでの 6 年間，本当に多くの方々にお世話になりました．周囲の方々の助けがなければ，この博士論文を完成させせることはできませんでした．ここに感謝の意を表します．

まずはじめに，6 年間親身になってご指導頂いた黒橋禎夫特定教授に感謝申し上げます．的確かつ洞察に満ちた助言によって，私の研究を常に適切な方向に導いて頂きました．修士課程が終わろうとしている時，まだ黒橋先生の元で学びたいと思い，博士課程に進学しましたが，その決断が間違いではなかったと確信しています．

河原達也教授と西田眞也教授には，論文審査を引き受けて頂き，貴重な助言を頂いたことに感謝申し上げます．

東中竜一郎教授（名古屋大学），光田航さんには，NTT 研究所でのインターンシップと，その後の国際学会や論文誌への論文投稿において多大なるご支援，ご指導をいただきました．研究に対して妥協なく取り組む姿は，私の研究に対する姿勢に大きな影響を与えて頂きました．

研究室の方々には，日々の研究活動で大変お世話になりました．河原大輔教授（早稲田大学），村脇有吾准教授，Chenhui Chu 特定准教授，Fei Cheng 特定助教には，研究やプレゼンテーションに関する多くの助言や，論文の添削をして頂きました．田中リベカ助教（東京工科大学）には，研究室配属当初，すべてにおいて未熟な私に研究の仕方を基礎から丁寧に教えて頂きました．私が博士課程に進学できるまで成長できたのは，田中リベカ助教のご指導のおかげです．岡久太郎助教（静岡大学），Yin-Jou Huang 特定助教，清丸寛一特定研究員には，研究に関する議論から論文の作文指導まで，幅広くきめ細やかなご指導を頂きました．私のぼんやりとした思索を的確かつ論理的に整理して頂いたおかげで，研究の方向性を明確にすることができました．Frederic Bergeron さんには研究で使用するシステムの実装をよく助けて頂きました．また，Qianying Liu さん，Haiyue Song さん，植田暢大

さん，大村和正さん，Zhuoyuan Mao さんといった，多くの優秀な同期に囲まれたことは，私にとって大きな刺激となりました．秘書の吉利菜帆さん，石田幸美さん，小杉照美さんには事務手続きを大変よく助けていただきました．そのおかげで研究に専念することができました．

また博士課程の間に研究助成をいただいた京都大学および日本学術振興会に感謝申し上げます.

最後に，今まで温かく私を見守ってくれた家族に感謝します.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Dialogue entails the exchange of information through communication in natural language (Nakano et al., 2015). Humans engage in numerous dialogues in daily life, sharing their thoughts and knowledge with each other. Dialogue plays a crucial role in human social life, contributing significantly to the formation of culture and the transmission of history.

Dialogue systems are computational systems designed to interact with humans using natural language. Dialogue systems can provide a user-friendly interface by employing natural language for the system's input and output. Owing to its characteristic of minimal user burden, the practical applicability of dialogue systems is highly significant, making their development a critical area of research within the field of natural language processing (NLP).

Text dialogue systems use text as system input and output, and have been studied for many years. Early text dialogue systems, like ELIZA (Weizenbaum, 1966), predominantly used rule-based pattern matching to generate responses to human utterances. These systems used a limited set of response patterns, limiting the flexibility of the dialogue response.

In recent years, end-to-end learning methods based on deep neural networks have attracted much attention in the field of dialogue systems. End-to-end dialogue systems are trained to generate responses from the dialogue context using a large number of context-response pairs. This learning method simplifies the construction of dialogue systems and allows for more flexible response generation.

However, dialogue systems still face challenges. One such challenge is the understanding of *user internal states*. The user internal states refer to the user's underlying states behind their utterance, with emotions being a prime example. In this thesis, we primarily focus on the user's *knowledge*, *interest*, and *willingness*. Below are the definitions of each internal state.

- **Knowledge**: Whether the user has knowledge of the topic.
- **Interest**: Whether the user has an interest in the topic.
- **Willingness**: Whether the user actively participates in the dialogue.

Understanding such states behind utterances can help dialogue systems identify user intentions and generate appropriate responses. However, existing research in dialogue systems has not tracked these user internal states in detail.

The second challenge involves the utilization of *external knowledge*. External knowledge refers to information external to the dialogue system, which can be acquired through means such as internet searches. By leveraging external knowledge, dialogue systems can generate more specific and informative responses. Prior research on external knowledge-grounded dialogue systems has focused on the retrieval of external knowledge and its incorporation into responses, often neglecting the underlying engagingness of the dialogue itself.

In this thesis, we address the two challenges: user internal states and external knowledge. To tackle these challenges, we adopt an approach through the construction of dialogue corpora to improve dialogue systems. With the advent of versatile generative models like the Transformer (Vaswani et al., 2017), the significance of dialogue corpora in dialogue system research has increased. This corpus-driven approach not only contributes to the improvement of the dialogue system via dataset utilization but also aids in the understanding of human behavior through corpus analysis.

For the remainder of this chapter, we first review the dialogue systems in Section 1.2. Then, we describe the challenges of current dialogue systems and our approaches in Section 1.3. Finally, we present the outline of this thesis in Section 1.4.

1.2 Review of Dialogue Systems

Dialogue systems require a variety of NLP techniques, from utterance understanding to response generation. Thus, the development of human-like dialogue systems has been considered one of the ultimate goals in the field of NLP and has attracted much attention. In this section, we will give a detailed review of dialogue systems.

1.2.1 Typology of Dialogue Systems

In this section, we present the representative types of dialogue systems from the perspectives of task, domain, and modality and describe the characteristics of each.

Task is an objective that a user seeks to accomplish. Users may or may not have an explicit task through dialogue. For instance, when the user says to the dialogue system, "Tell me the weather tomorrow," the user assigns the task of providing weather information to the system. On the other hand, when the user says, "It is hot today," the user is not assigning any task. Like the former scenario, a dialogue system designed to perform a task is referred to as a "task-oriented dialogue system." Conversely, like the latter scenario, a dialogue system that is not primarily engaged in any task is referred to as a "non-task-oriented dialogue system."

Domain represents the range of topics the dialogue system can handle. A weather information retrieval system handles only topics related to weather forecasts, and a restaurant reservation system handles only topics related to restaurants. Like these examples, dialogue systems that limit the topics they can handle are called "closed-domain dialogue systems." On the other hand, dialogue systems that do not limit the topics are called "open-domain dialogue systems."

Modality refers to the style of information exchange, encompassing aspects such as tone of voice, facial expressions, gestures, and eye gaze. Dialogue systems can be classified based on their input and output modalities. Dialogue systems that use text for both input and output are referred to as "text dialogue systems." In contrast, systems that employ voice for input and output are referred to as "spoken dialogue systems." Additionally, systems that incorporate multiple modalities are termed "multimodal dialogue systems."

1.2.2 Rule-based Dialogue Systems

Rule-based dialogue systems generate responses based on hand-crafted rules. The earliest rule-based non-task-oriented dialogue system is ELIZA (Weizenbaum, 1966), developed in 1966 by Weizenbaum. ELIZA mimics a psychotherapist and engages in chat with the user, a patient. This system selects responses from a collection of response templates by pattern matching. Example (1) shows an example of an interaction with ELIZA. Upon detecting the word "depressed" in the user's utterance, ELIZA responds with a predefined response template associated with "depressed" like "I am sorry to hear you are depressed."

(1) *USER*: Men are all alike.
ELIZA: In what way.
USER: They're always bugging us about something or other.
ELIZA: Can you think of a specific example.
USER: Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here.
USER: He says I'm depressed much of the time.
ELIZA: I am sorry to hear you are depressed.

Although ELIZA does not attempt to understand the user's utterances, well-designed response rules have successfully made users feel that the system is listening to their utterances.

One of the pioneering systems in rule-based task-oriented dialogue systems is SHRDLU (Winograd, 1972). SHRDLU is able to understand instructions given by users in natural language, manipulate blocks, and inquire about the state

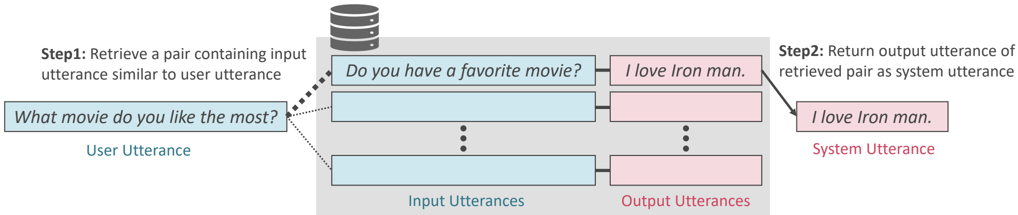


Figure 1.1: An overview of the retrieval-based dialogue system.

of these blocks. Analyzing the results of language analysis and the block states enables the understanding of user utterances. While SHRDLU demonstrates these behaviors by restricting the dialogue domain to a simple block world, extending these capabilities to the real world presents significant challenges.

These rule-based dialogue systems are able to generate accurate responses within highly limited domains. However, as the domain expands, the description of rules becomes more complex, leading to decreased maintainability. Furthermore, these systems have issues with flexibility in responding to user utterances.

1.2.3 Retrieval-based Dialogue Systems

Retrieval-based dialogue systems generate responses by retrieving similar utterances from a database of dialogue examples. Figure 1.1 shows an overview of retrieval-based dialogue systems. In the first step, a large number of input and output utterance pairs are collected and stored in a database. In the next step, the system searches for the most similar input utterance to the user utterance. Finally, the system returns the corresponding output utterance as its response.

To construct a database of input and output utterance pairs, Ritter et al. (2011) collect Twitter reply pairs, which are pairs of a user's tweet and a reply to that tweet. They collect 1.3 million reply pairs and use them to construct a dialogue system. Banchs and Li (2012) utilize movie scripts, which are freely available at The Internet Movie Script Data Collection.[1]

To retrieve utterance pairs, it has been common to employ information retrieval algorithms such as TF-IDF and BM25. Recently, neural network-based

[1] https://imsdb.com/

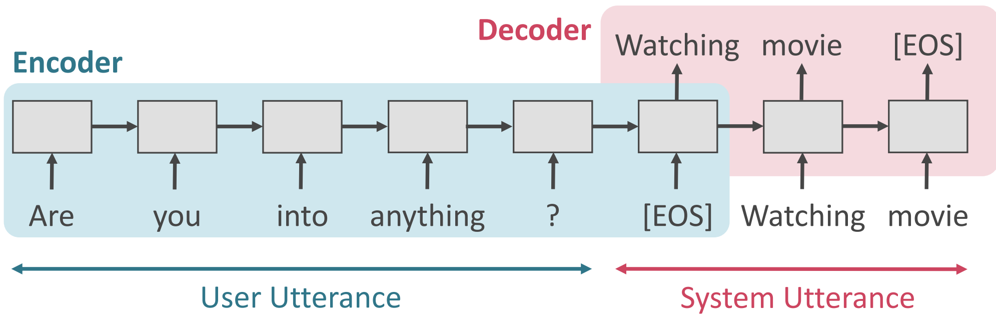


Figure 1.2: An overview of seq2seq dialogue system.

methods have been proposed to calculate the similarity between utterances (Lowe et al., 2015, 2017). Methods employing powerful pre-trained language models, particularly those like BERT (Devlin et al., 2019), have become increasingly prevalent (Lu et al., 2020; Han et al., 2021).

Retrieval-based systems can prepare responses for any user utterance, eliminating the need for costly hand-crafted rule descriptions. However, the responses are limited to utterances within the database, leading to a lack of flexibility.

1.2.4 Generation-based Dialogue Systems

Generation-based dialogue systems generate responses from scratch to user utterances without predefined rules or response candidates stored in a database. This generation-based method has rapidly developed along with recent advancements in deep neural networks and has become mainstream in current dialogue system research. In this section, we first discuss the initially developed model architecture-based approach, followed by the later corpus-based approach. Finally, we describe the rapidly evolving large language model-based approach.

Model Architecture-based Approach

The most fundamental generative model is the sequence-to-sequence (seq2seq) model (Sutskever et al., 2014). Originally proposed as a neural machine translation model, it has since been applied to dialogue systems. Figure 1.2 shows an overview of dialogue systems using the seq2seq model (Shang et al., 2015; Vinyals and Le, 2015). The seq2seq model comprises an encoder, which processes the input, and a decoder, which generates the output. Consider a user utterance such as "Are you into anything?" as input. The sequence of tokens in this input (i.e., "Are", "you", "into", "anything", "?", "[EOS]") is fed into the encoder and converted into a vector ([EOS] is a special token indicating the end of a sentence). The last state of the encoder represents the contextualized embedding of the input utterance. The decoder receives this vector and generates the system utterance "Watching Movie", word by word.

This seq2seq model is trained by end-to-end learning. Specifically, the model is trained to generate the output system utterance from the input user utterance using large-scale dialogue data. This training scheme demands a substantial amount of dialogue data and extensive computational resources. Nonetheless, the growing accessibility of the internet and advancements in computational technology have made this feasible.

This simple framework has dramatically simplified the construction of a complex dialogue system. On the other hand, the encoder compresses input sequences into a fixed-length vector; thus, handling longer input sequences has remained a challenge. Consequently, improvements from the perspective of model architecture continued for some time.

Hierarchical Recurrent Encoder-Decoder (HRED) (Serban et al., 2016) extends the seq2seq model to consider not only the immediate prior user utterance but also the multi-turn dialogue contexts. Figure 1.3 shows the HRED architecture. The HRED encoder consists of two encoders: an utterance-level encoder and a context-level encoder. The utterance-level encoder encodes each utterance and compiles it into a vector. The context-level encoder consolidates the outputs of the utterance-level encoder and outputs a vector representing the context of the dialogue.

Attention mechanism (Bahdanau et al., 2015) highlights important words in the dialogue context, helping the model focus on the most relevant parts to generate better responses. Specifically, this mechanism first computes weights that signify levels of attention. Then, the weighted sum of the vectors of each word in the input utterance is calculated based on these weights. During decoding, the

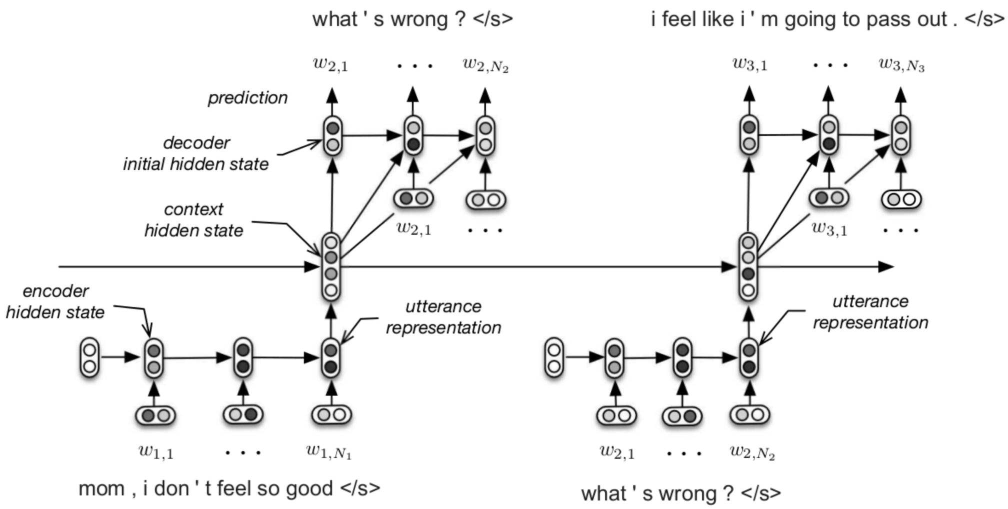


Figure 1.3: The HRED architecture. Adapted from Serban et al. (2016).

model generates words while referring to this weighted vector. Such a mechanism facilitates dynamic referencing of the dialogue context throughout the generation process. Consequently, this approach facilitates the generation of responses that are consistent with the dialogue contexts.

Transformer (Vaswani et al., 2017) refines the attention mechanism and has become a de facto standard model in today's NLP. Transformer employs self-attention mechanisms in both its encoder and decoder. The self-attention mechanism attends to which parts of the input sequence concern each other to capture relationships between words. This architecture can provide a robust understanding of context, and such Transformer-based dialogue systems have exhibited higher response quality than naive seq2seq models like HRED (Oluwatobi and Mueller, 2020).

Corpus-based Approach

With the maturation of versatile generative models like Transformer, the dialogue systems have come to be constructed through a combination of pre-training on large-scale data and fine-tuning on small-scale data (Wolf et al., 2019; Zhang et al., 2020). Pre-training involves training versatile generative models using large corpora. At this stage, the focus is not on learning task-specific features but rather on understanding language's fundamental structures and contexts. The large corpora primarily employed for this purpose include web documents like Wikipedia and BookCorpus (Zhu et al., 2015), as well as dialogue format corpora from sources such as Reddit, Twitter, and OpenSubtitles (Lison and Tiedemann, 2016; Lison et al., 2018).

Fine-tuning is a process that adapts a pre-trained model to a particular task using a smaller, task-specific corpus. Unlike the extensive data needed for pre-training, just a few thousand to tens of thousands of dialogue pairs are often enough. This amount of data can be collected manually. Recently, crowdsourcing has become a popular way to create such corpora for dialogue research. Here, we introduce three representative tasks and their corresponding datasets.

Persona refers to an individual's personality and background. For instance, utterances like "I am a student" or "I like cats" reflect the personality and background of the speaker. Unlike humans, dialogue systems lack a consistent persona, often resulting in contradictory utterances across multi-turn dialogues (Li et al., 2016). In Example (2), the system answers with a different location each time the user asks where the system lives.

(2) $USER$: Where do you live now?
$SYSTEM$: I live in Los Angeles.
$USER$: In which city do you live now?
$SYSTEM$: I live in Madrid.
$USER$: In which country do you live now?
$SYSTEM$: England, you?

To address the issue of consistency in these personas, Zhang et al. (2018) constructed a PersonaChat dataset. Figure 1.4 shows an example of PersonaChat. Each speaker (i.e., worker) is given a persona consisting of short sentences like "I like to ski." or "I am an artist." and instructed to engage in dialogue based on that persona.

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Figure 1.4: A dialogue example from PersonaChat (Zhang et al., 2018).

Empathy refers to understanding and relating to the feelings of others. EmpatheticDialogues (Rashkin et al., 2019) attempts to incorporate such human-like behavior into dialogue systems. Figure 1.5 shows a dialogue example from EmpatheticDialogues. In this dialogue collection, paired workers have different roles: speaker and listener. The speaker is given one emotion randomly selected from 32 emotions (e.g., surprised, excited, and proud) and asked to describe a situation that elicits this emotion in a few sentences. Subsequently, based on this situation, the speaker expresses that emotion in dialogue, and the listener responds to demonstrate empathy towards the speaker.

External Knowledge refers to information external to the dialogue system, accessible through the internet or database searches. This external knowledge assists dialogue systems in generating more informative responses. Wizard of Wikipedia (Dinan et al., 2019) is one of the most famous external knowledge-grounded dialogue benchmarks. Figure 1.6 shows an example of dialogue in Wizard of Wikipedia. In Wizard of Wikipedia, workers are divided into the roles of

Label: Proud
Situation: Speaker felt this when...
"I finally got that promotion at work! I have tried so hard for so long to get it!"
Conversation:
Speaker: I finally got promoted today at work!
Listener: Congrats! That's great!
Speaker: Thank you! I've been trying to get it for a while now!
Listener: That is quite an accomplishment and you should be proud!

Figure 1.5: A dialogue example from EmpatheticDialogues (Rashkin et al., 2019).

Topic:	Armadillo
Wizard:	I love animals and think armadillos are awesome with their leathery shell.
Apprentice:	I don't think I've ever seen an armadillo in real life!
Wizard:	I've seen them at the zoo. Armadillo means little armored one in Spanish.
Apprentice:	Are they native to a Spanish-speaking part of the world?
Knowledge:	Armadillos are New World placental mammals in the order Cingulata . . . The word "armadillo" means "little armoured one" in Spanish. . . . The nine-banded armadillo ("Dasypus novemcinctus"), or the nine-banded, long-nosed armadillo, is a medium-sized mammal found in North, Central, and South America.
Wizard:	Yes, they are most commonly found in North, Central, and South America

Figure 1.6: A dialogue example from Wizard of Wikipedia (Dinan et al., 2019).

"Wizard" and "Apprentice" to discuss a specific topic. The wizard worker, provided with Wikipedia articles as external knowledge, creates responses based on these articles.

These three *skills* introduced so far are essential for dialogue systems to behave in a human-like manner. However, each dataset is specialized for a single skill, making it challenging to learn multiple skills simultaneously. To address this, BlendedSkillTalk (Smith et al., 2020) is proposed, aiming to construct a dialogue system that integrates these skills. This dataset is created by preparing generative models specialized in each skill and guiding workers to collect dialogues that combine these three skills. The BlenderBot (Roller et al., 2021), trained on this BlendedSkillTalk, has been reported to generate human-level quality responses.

In summary, approaches using such high-quality, small-scale datasets created by manual efforts have succeeded considerably.

Large Language Model-based Approach

Large language models (LLMs) (Brown et al., 2020; OpenAI, 2023) are language models with billions of parameters. These models are trained on massive amounts of corpora, encompassing hundreds of billions to trillions of tokens. Such LLMs have demonstrated remarkable performance across various benchmarks in NLP, including translation tasks and natural language understanding, thereby exhibiting their extensive versatility. For example, few-shot learning (Brown et al., 2020) leverages the broad applicability of LLMs and solves a task by providing the model with a task description and a few example solutions. Furthermore, zero-shot learning (Brown et al., 2020), which does not require any example solutions, has also been proposed. These methodologies, eliminating the need for task-specific datasets, have garnered significant attention in the NLP field.

Dialogue systems have recently utilized LLM-based approaches (Kim et al., 2021; Thoppilan et al., 2022; Shuster et al., 2022). Blenderbot 3.0 (Shuster et al., 2022) is an open-domain chatting dialogue system that has been fine-tuned on various datasets, built on the Open Pre-trained Transformer (Zhang et al., 2022) with 175 billion parameters. Figure 1.7 shows the architecture of Blenderbot 3.0. This Blenderbot 3.0 has been designed to generate higher-quality responses by integrating various modules such as internet search and long-term memory. InstructGPT (Ouyang et al., 2022) is the fine-tuned version of GPT-3 (Brown et al., 2020) through reinforcement learning from human feedback (Christiano et al., 2017; Stiennon et al., 2020) to follow human instructions. ChatGPT[2] (GPT-3.5 and GPT-4 (OpenAI, 2023)), which is a successor to InstructGPT, demonstrates outstanding performance, capable of handling a wide range of dialogues from task-oriented to non-task-oriented dialogues.

In summary, generation-based dialogue systems have developed through end-to-end learning. Due to the improvements in model architecture, general-purpose generative models like Transformer have emerged. Additionally, specialized cor-

[2] `https://openai.com/blog/chatgpt`

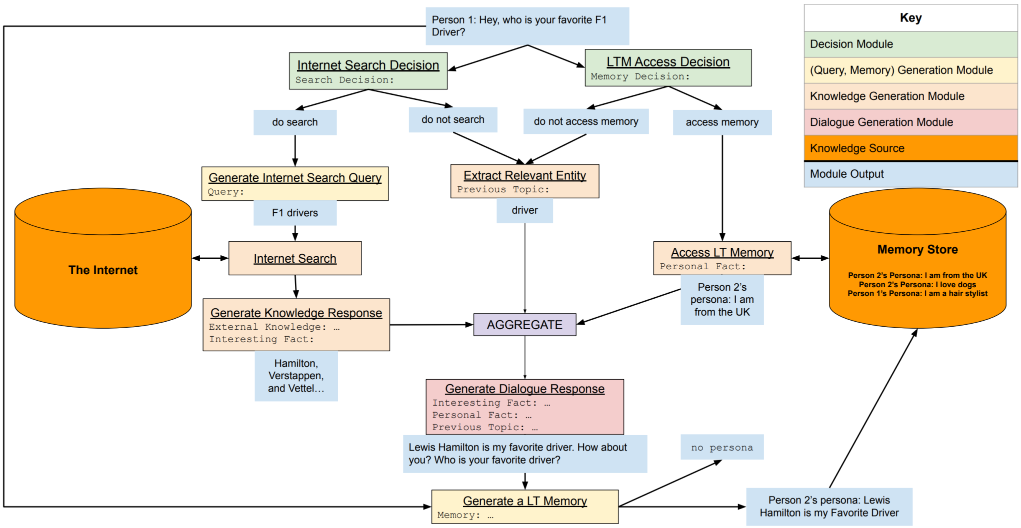


Figure 1.7: An overview of Blenderbot3.0 (Shuster et al., 2022).

pora that capture various aspects of human behavior have been created to produce more human-like responses. In recent years, the emergence of large language models has further enhanced the quality of generated responses.

1.3 Challenges and Aproarches

With the recent advancements in deep neural networks, generation-based dialogue systems are increasingly achieving human-like response quality. However, challenges remain. This thesis addresses the two major challenges in generation-based dialogue systems: user internal states and external knowledge. To address these challenges, we adopt a corpus-based approach, which collects small-scale but high-quality corpora tailored to specific tasks to enhance the dialogue systems. Dialogue corpora form the foundation of research and development in dialogue systems. They are not merely a source of training data but also provide new insights through their analysis.

Building on this foundation, this thesis examines recommendation dialogues, considering the human tendency to seek information tailored to their interests. In addition, our research specifically explores movie recommendation dialogues, chosen for their universal appeal and potential to stimulate engaging conversations. This section details the challenges of user internal states and external knowledge, and discusses a corpus-based approach to these issues.

1.3.1 User Internal States

User internal states refer to the user's underlying mental states behind their utterances. For dialogue systems, understanding the user internal state is instrumental in identifying the user's intentions. For instance, when a user says "That is fine," the user's intention behind this utterance can vary depending on whether the user is in a state of reassurance or frustration. In the former case, the user's utterance is likely to indicate approval or agreement. Here, the dialogue system should confirm the user's satisfaction and proceed to the next step. In contrast, in the latter case, the utterance might encompass dissatisfaction or sarcasm. The word "fine" in this context could actually imply that there is a problem. In such situations, the system needs to respond more cautiously, posing additional questions or suggestions to resolve the user's dissatisfaction. Thus, understanding the user internal states enables dialogue systems to generate more appropriate responses.

The user internal states during a dialogue have been extensively studied from various perspectives. *Emotions* (Poria et al., 2019; Zhou and Wang, 2018; Song et al., 2019), capturing the user's feelings of joy, sadness, anger, and more, provide a foundation for dialogue systems to respond more appropriately. *Intention* (Griol and Callejas, 2016) represents what the user seeks from the dialogue system and is often managed through labels of dialogic acts (Stolcke et al., 2000) such as "question" or "providing information." *Personality traits* (Meguro et al., 2009) like the Persona (Zhang et al., 2018) or the Big Five (Digman, 1990; Wu and Sakai, 2020; Guo et al., 2021) are essential elements for generating responses that align with the user's character and background. *Satisfaction* (Arimoto et al., 2019; Deng et al., 2022; Ye et al., 2023; Feng et al., 2023), beyond just understanding the user, is also employed in the automatic evaluation of dialogue responses (Bodigutla et al., 2020). Moreover, in counseling dialogues, it is crucial to understand the user mental state, such as *depression* (Yang et al., 2023) and *suicidal ideation* (Lamichhane, 2023).

In the context of user internal states focusing on the relationship between users and systems, *rapport* (Acosta, 2009; Müller et al., 2018) denotes the extent to which users believe in the system, which is acquired through mutual understanding fostered during dialogues. *Intimacy* (Kodama et al., 2021; Arimoto et al., 2023) indicates the level of closeness a user feels towards the system. When intimacy is high, it is expected that the system behaves akin to a family member or a close friend.

In task-oriented dialogue systems, it is effective to understand the user's *familiarity with the system* and the *urgency of the task* (Komatani et al., 2003). For users with high familiarity, dialogues can be made more efficient by omitting basic information, while for users with low familiarity, providing more detailed information is effective. Additionally, understanding the urgency of the task enables the determination of whether the user is seeking specific information quickly or requires a more detailed explanation.

Although there are various user internal states, as mentioned above, this thesis focuses on *knowledge*, *interest*, and *willingness*, which are considered to be particularly important in the context of movie recommendations. The user's knowledge state represents the amount of knowledge the user possesses on a specific topic. For instance, by examining the following utterances of User A and User B, it can be inferred that User A has knowledge of "Iron Man," whereas User B does not.

(3) User A: I love Iron Man.
User B: I am not familiar with Iron Man.

The user's interest state refers to whether the user is interested in a specific topic. For example, upon examining the utterances of User A and User B below, User A seems to exhibit an interest in "Iron Man," whereas User B seems to lack interest.

(4) User A: I want to see Iron Man.
User B: I have no desire to see Iron Man.

The user's willingness state refers to whether the user intends to continue the dialogue. For instance, by examining the utterances of User A and User B below, it appears that User A intends to continue the dialogue. In contrast, User B seems to lack such an intention.

(5) User A: I think it's an interesting idea. Can you tell me more about it?
User B: Oh, okay.

This willingness is closely related to interest, yet the two concepts are distinct. Interest represents an interest in a specific topic, whereas willingness represents an interest in the dialogue. Thus, some users exhibit a high willingness even without interest in a particular topic, and vice versa.

These user internal states have been actively utilized in multimodal dialogue systems that can handle non-linguistic features (Schuller et al., 2006; Miyazaki et al., 2013; Ishihara et al., 2018; Inoue et al., 2018; Komatani and Okada, 2021). However, the utilization of such features requires the technology to collect and process high-quality video and eye-tracking data in real-time, which imposes significant costs and computational resource demands.

In text dialogue systems, Inaba and Takahashi (2018) annotates the human-to-human chat dialogues with the degree of interest in predetermined 24 topics, such as travel, movies, and celebrities. Curiosity (Rodriguez et al., 2020) is a dialogue dataset annotated with users' prior knowledge about geographical topics. Specifically, before collecting dialogues, users are presented with 15 entities related to the topic, and they annotate in a binary manner whether they are familiar with each entity. These previous studies label the user internal states at the dialogue level. Considering the fact that the user internal states can change moment by moment within a dialogue, more fine-grained annotation is required.

In this thesis, we present a fine-grained annotation of user internal states at both the utterance and entity levels. Chapter 2 describes our initial step, where we develop a rule-based dialogue system for movie recommendations and collect dialogues between users and the system. In this phase, we annotate each user utterance in the dialogues, focusing on their levels of knowledge, interest, and willingness. Subsequently, in Chapter 5, we curate human-to-human dialogues in the same context of movie recommendations. Here, one dialogue participant acts as a recommender, and the other acts as a seeker. Based on insights from Chapter 2, we focus on knowledge and interest as important user internal states. We then provide a more granular annotation of the seeker's utterances, marking degrees of knowledge and interest at the entity level, a step beyond the utterance-level annotation. This approach enables us to track multiple user internal states within a single utterance, such as "I know the movie title, but I do not know the plot."

1.3.2 External Knowledge

External knowledge refers to information external to the dialogue system and is accessed through searches, such as on the internet or in databases. Utilizing external knowledge is crucial for dialogue systems due to the limited capacity of internal knowledge storage and management. For instance, a dialogue system with data only up to the year 2020 would be unable to generate appropriate responses to questions about the 2021 Tokyo Olympics. By acquiring external knowledge through internet searches and other means and reflecting it in responses, the system can provide more accurate and up-to-date responses.

In the realm of dialogue research, external knowledge-grounded dialogue response generation (Ghazvininejad et al., 2018) is a significant topic of interest, with numerous datasets having been proposed to facilitate this area of study (Zhou et al., 2018; Moghe et al., 2018; Dinan et al., 2019; Wu et al., 2019; Gopalakrishnan et al., 2019; Zhou et al., 2020). However, these works primarily focus on the selection of external knowledge or its incorporation into responses. Consequently, the underlying dialogues often lack engagingness (Wang et al., 2021).

In this thesis, we aim to curate more engaging external knowledge-grounded dialogues. In Chapters 3 and 4, we construct an external knowledge-grounded dialogue dataset and analyze engaging dialogues. The analysis, focusing on the information sources, reveals that external knowledge-grounded dialogues are engaging, especially when the recommenders utilize information they already possess about the topic. Based on this finding, in Chapter 5, we construct a dialogue dataset by hiring movie enthusiasts who possess knowledge about the topic (in this case, movies) as the recommenders, thereby collecting more engaging dialogues based

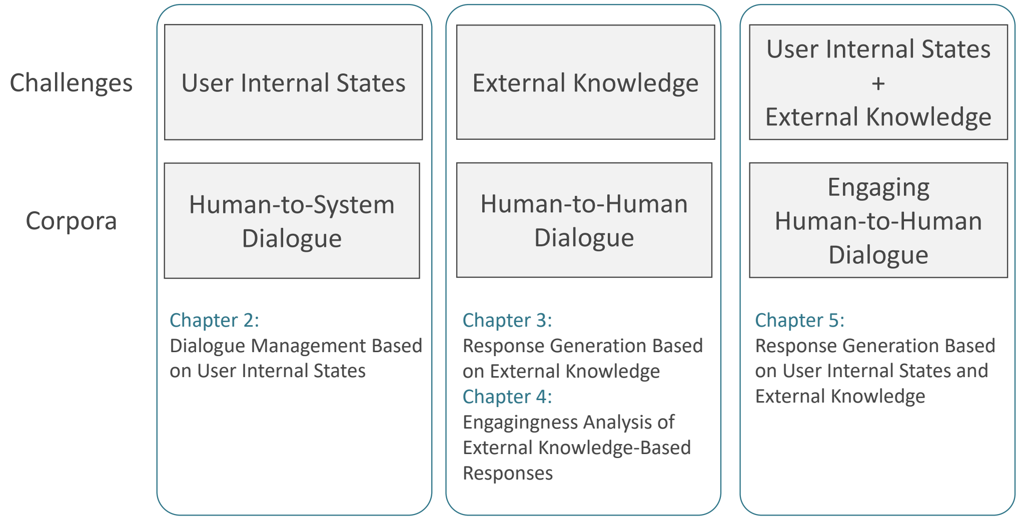


Figure 1.8: An overview of the thesis.

on external knowledge.

1.4 Outline of the Thesis

To conclude the introduction, we present the outline of the thesis. Figure 1.8 provides a visual overview of this thesis, illustrating the challenges to be addressed and the types of dialogue corpora constructed for these challenges. In this thesis, we focus on utilizing user internal states and external knowledge, researching these elements individually and their integrated utilization. To utilize these elements effectively, we adopt approaches through the construction of specialized dialogue corpora. At each stage, we aim to approximate dialogues in real-world situations, encompassing human-to-system dialogues, human-to-human dialogues, and engaging human-to-human dialogues.

In Chapter 2, we introduce our work on dialogue management based on user internal states. We develop a dialogue system using pre-created scenarios and rules and collect dialogues between the system and human participants. Then, we annotate all the user utterances with three user internal states: knowledge, interest, and willingness. Using the constructed corpus, we train classifiers to estimate each user internal state. Finally, by adding rules to alter responses according to the estimated user internal states, we develop a dialogue system capable of more natural responses and demonstrate the effectiveness of considering the three user internal states.

In Chapter 3, we introduce our work on dialogue response generation based on external knowledge. First, we develop a dialogue collection framework capable of constructing large-scale, human-to-human dialogues in Japanese. Utilizing this framework, we build a Japanese external knowledge-grounded dialogue dataset. Using this dataset, we present a strong baseline model that simultaneously selects external knowledge and generates responses based on it. This model utilizes the attention mechanisms that focus on the history and structure of external knowledge.

In Chapter 4, we introduce our work on analyzing the engagingness of external knowledge-grounded responses. In the dialogue data constructed in Chapter 3, we annotate every entity in the recommender's utterances, classifying whether the entity is derived from external knowledge or from the recommender's inherent knowledge and opinions. By analyzing the annotated corpus, we reveal that information derived from the recommender contributes to the engagingness of the response.

In Chapter 5, we introduce our work on response generation based on both user internal states and external knowledge. We construct a dialogue dataset annotated at the entity level for two user internal states: knowledge and interest. By requesting annotations from the dialogue participants themselves, we aim to reflect the participants' actual internal states. Furthermore, using this dataset, we propose a model that generates responses based on both user internal states and external knowledge utilizing Chain-of-Thought prompting (Wei et al., 2022).

In Chapter 6, we present the overall conclusion of this thesis and discuss the future prospects.

Chapter 2

Dialogue Management Based on User Internal States

2.1 Introduction

In human dialogues, individuals pay careful attention to their interlocutor's internal state (Chiba et al., 2014), including their level of understanding and emotional state. For example, if their interlocutor does not seem to understand their utterance, they add or rephrase their words to help their interlocutor understand. Recent dialogue systems utilizing deep neural networks are trained to generate plausible responses to input utterances based on large-scale dialogue-formatted data (Adiwardana et al., 2020; Smith et al., 2020). However, the same input utterances can have different underlying intents. Thus, dialogue systems should discern the interlocutor's intent appropriately and adjust their responses accordingly. In order to understand the interlocutor's intent, we deal with the interlocutor's internal state during dialogue and address the following two issues.

1. Modeling the interlocutor's internal state
2. Response changes based on the interlocutor's internal state

In general, real-life dialogues always have specific purposes, such as "conveying information through dialogue" or "influencing the interlocutor." Under this consistent purpose, we humans exchange information over multiple dialogue turns.

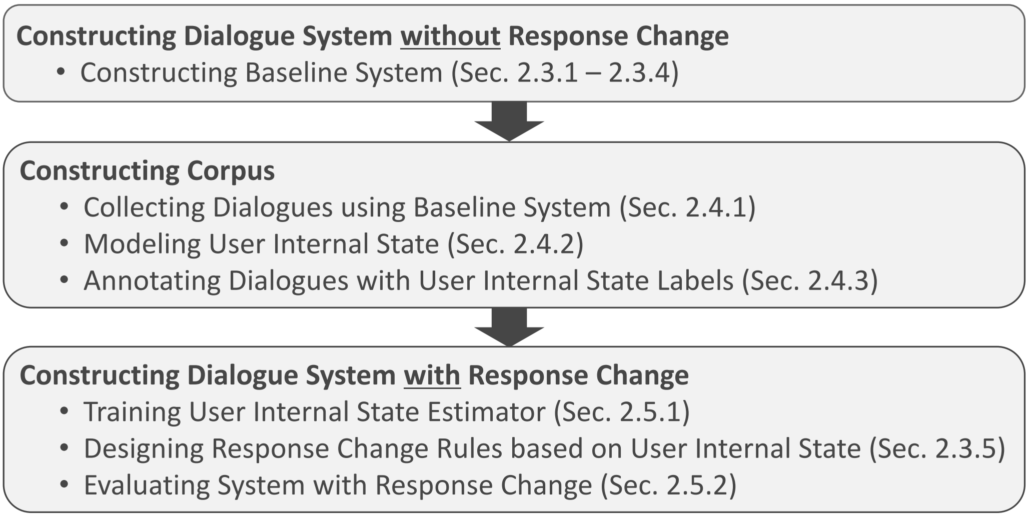


Figure 2.1: Flowchart of this study

Thus, when modeling the interlocutor's internal state, dialogue data with multiple turns and a consistent purpose throughout is required. However, analyzing and modeling the internal states in human-to-human dialogues is challenging due to the intertwining of intentions. On the other hand, while recent deep neural network-based dialogue systems can provide appropriate responses at a single-turn level, they face many challenges in multi-turn dialogues. In most cases, these dialogues lack a consistent purpose.

In this study, we construct a rule-based dialogue system in the domain of movie recommendations. We then collect multiple-turn dialogue data with a consistent purpose between the dialogue system and humans. The consistent purpose here is "movie recommendation." Figure 2.1 shows the overall flow of this study. First, we construct a baseline dialogue system and collect the dialogues between the dialogue system and humans. Based on the analysis of the collected dialogues, we model the interlocutor's internal state in dialogue along the following three distinct axes.

- **Knowledge**: Whether the interlocutor has knowledge of the topic.

- **Interest**: Whether the interlocutor has an interest in the topic.

- **Willingness**: Whether the interlocutor actively participates in the dialogue.

Understanding the interlocutor's knowledge and interest enables the dialogue system to provide appropriate information or change topics accordingly. Moreover, by considering the interlocutor's willingness for dialogues, the dialogue system can adapt its behavior, for instance, by playing a more passive listening role when the interlocutor has a high willingness.

We annotate our collected dialogues with the modeled interlocutor's internal states, and subsequently train models to estimate these states using the annotated data. As a result, we achieved a high estimation accuracy of approximately 80–85% for each internal state when allowing an error of ± 1 on a 7-point scale.

Furthermore, we have the dialogue system change its response according to the results of the trained model's estimation. Specifically, we devise rules for modifying responses depending on the presence or absence of knowledge, interest, and willingness. Both dialogue-level and utterance-level evaluations showed that the naturalness of responses improved when modified according to the estimated interlocutor's internal state.

Our contributions are two-fold:

- We constructed a text dialogue corpus of about 10,000 utterances with the interlocutor's internal states (knowledge, interest, and willingness) assigned to each utterance.

- We empirically demonstrated the effectiveness of automatically estimating the user internal state and modifying the system responses according to the estimation results.

2.2 Related Work

For dialogue systems, it is crucial not only to comprehend the literal meaning of an utterance but also to understand the internal states of the speaker who generates the utterance. Here, we refer to the internal state of the speaker behind their utterance as "user internal states." In this section, we introduce prior studies on user internal states, including knowledge, interest, and willingness.

Emotion, one of the user internal states, has been actively used in dialogue research. There has been a lot of research on estimating emotions from utterances and generating utterances based on emotions. Poria et al. (2019) estimate emotions (e.g., anger, disgust) using verbal and non-verbal information from speeches. In utterance generation based on emotions, Zhou and Wang (2018) take emoticons in tweets as emotion annotations and propose a method to generate emotional utterances from a large number of tweets. Song et al. (2019) have proposed a method of reflecting specific emotions in utterances, focusing on the explicit and implicit expressions of emotions.

Persona (Li et al., 2016; Zhang et al., 2018) is a speaker's personal background, such as age and gender, as well as the way of speaking based on that background. Various modeling methods have been proposed to realize consistent personalities, such as IDs (Li et al., 2016), personal attributes (Qian et al., 2018), and short personal profile texts (Zhang et al., 2018). Emotions and personas are mainly used to make dialogue responses more informative, but the purpose of our study is to understand the user internal states.

Research has also been conducted with a focus on users' knowledge, interests, and willingness. Miyazaki et al. (2013) investigate effective features for estimating callers' levels of knowledge in call center dialogues and propose a method to estimate their levels of knowledge. Schuller et al. (2006) estimate users' interest level from spoken dialogues. Inaba and Takahashi (2018) estimate the interest level in 24 topics in text chit-chat dialogues between humans. Ishihara et al. (2018) use multi-modal information to estimate the users' willingness in interview dialogues. Inoue et al. (2018) have attempted to estimate engagement using multiple non-verbal behaviors from dialogues between humans and the android robot ERICA (Inoue et al., 2016).

While these prior works individually estimate knowledge, interest, and willingness, our study concurrently addresses these three user internal states. Moreover, we estimate these states at each turn of the dialogue, assuming these states fluctuate dynamically. Subsequently, we construct a dialogue system that can interpret

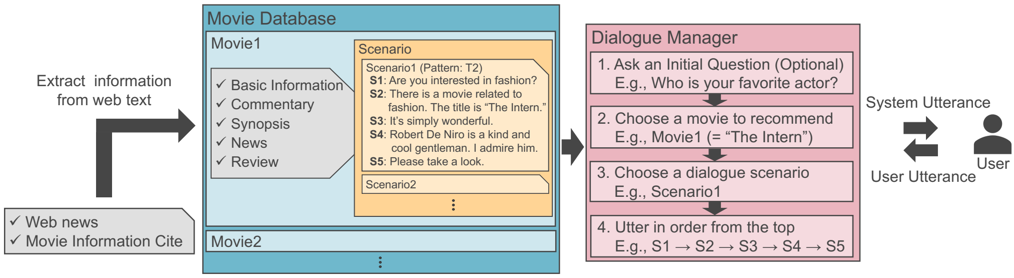


Figure 2.2: Overview of our movie recommendation dialogue system.

the user internal states and respond naturally by changing the response appropriately according to the estimation results.

2.3 Configuration of Dialogue System

In order to observe and analyze the interlocutor's internal state, we need a dialogue system capable of engaging in multi-turn dialogues with a consistent purpose. Consequently, we develop a rule-based dialogue system designed for recommending movies.

Figure 2.2 shows the overview of our dialogue system. Our system comprises two primary components: *Movie Database* and *Dialogue Manager*. Movie Database stores the information for each movie. Movie Database also prepares several templates of system utterances for each movie in advance. We refer to these templates as "scenarios." Dialogue Manager progresses the dialogue based on the scenarios and provides brief answers to the user's question by referring to the movie database as needed. Example (1) shows an example of a dialogue between our system and a user. S and U denote system utterances and user utterances, and the numbers next to them indicate the number of turns in the dialogue.

(1) $S1$: ファッションって興味ありますか？ (Are you interested in fashion?)
$U1$: 興味ありますよ (I'm interested in it.)
$S2$: ファッションに関係のある映画があるんです.「マイ・インターン」というタイトルです. (There is a movie related to fashion. The title is

"The Intern.")

*U*2: タイトルは聞いたことあるような気がします．(I think I've heard the title before.)

*S*3: 素晴らしいの一言ですよ．(It's simply wonderful.)

*U*3: そうなんですね．(I see.)

*S*4: ロバート・デニーロは優しくカッコいい紳士，憧れますよ．(Robert De Niro is a kind and cool gentleman. I admire him.)

*U*4: それはいいですね (It's nice!)

*S*5: 是非見てみてください (Please take a look.)

2.3.1 Dialogue Strategy

Our system employs the following three dialogue strategies to facilitate more natural movie recommendations.

(1) System Initiative The system takes the initiative in the dialogue based on the pre-defined scenario while responding to user utterances.

(2) Indirect Topic Introduction Our system avoids directly mentioning the movie title, which is the central subject from the beginning of the dialogue. Instead, it initiates dialogue from related subjects. We use the following three patterns for indirect topic introduction. The underlined parts depend on the movie.

P1: Recent Entertainment News

ex., 俳優の染谷将太と女優の菊地凛子夫妻に第2子が誕生したことが11日までにわかったと話題です (It's a hot topic that the second child has been born to actor Shota Sometani and actress Rinko Kikuchi.)

P2: Movie Theme

ex., 飛行機って興味ありますか？ (Are you interested in airplanes?)

P3: Movie Information (e.g., director's name)

ex., 細田守監督をご存知ですか? (Do you know director Mamoru Hosoda?)

(3) Selling Points We remind the user that they want to watch the recommended movie by repeatedly uttering the movie's selling points.

2.3.2 Movie Database

The movie database stores the basic information (such as title, cast, staff, and genre), plot, news, and reviews for 213 movies. We obtained the basic information, plot, and reviews from the movie information website.[1] As for reviews, 300 top-rated reviews for each movie were collected. News information was obtained from entertainment and sports news websites. This news information is utilized for the creation of the scenario pattern **P1** as discussed in Section 2.3.1.

2.3.3 Dialogue Scenario

Based on the aforementioned dialogue strategies, we prepare one or more scenarios for each movie. Each scenario consists of five system utterances ($S1$ to $S5$). This section describes how to create each utterance.

Utterance $S1$ The system starts with the utterance $S1$ containing a movie-related topic. This topic is based on the indirect topic introduction patterns **P1** to **P3** as discussed in Section 2.3.1.

In the pattern **P1**, the dialogue begins with a topic from recent entertainment news and recommends movies in which the characters from that news have appeared. If the first sentence of a news article contains the cast's name, the utterance is made like "It is the hot topic that ⟨the first sentence of a news article⟩."

In the pattern **P2**, the dialogue is initiated with an utterance like "Are you interested in ⟨movie theme⟩?" using the movie's theme. The movie's theme is the noun with the highest tf-idf score from the movie's plots after removing the names of the cast and staff. However, in order to prevent the use of generic words as themes, we discard the scenario of pattern **P2** if the highest tf-idf value falls below 0.35.

The pattern **P3** incorporates the names of two primary cast members and the director. The dialogue begins with utterances such as "Do you know direc-

[1] `https://movies.yahoo.co.jp/`

tor/actor/actress ⟨person name⟩?" The system then recommends movies associated with that person.

Utterance $S2$ The system presents the movie title associated with $S1$ as "The title is ⟨movie title⟩."

Utterances $S3$ **and** $S4$ The system then presents two selling points of the movie. We select the selling points from the movie reviews.

We initially split the review text into individual sentences. Given the varied tone across these sentences, we employ the Juman++ dictionary[2] to harmonize their endings into a uniformly polite style, characterized by the use of "です, *desu*" or "ます, *masu*." Additionally, we append the suffix "よ, *yo*" to impart an informative tone.

Next, we select sentences praising the movie. Specifically, we first manually select 50 positive words related to movies, such as "masterpiece" and "interesting." We calculate the cosine similarity between the sum of word vectors for these positive words and that for each review sentence. To avoid selecting short sentences, we multiply this similarity score by the sentence length to rank the sentences. Based on this ranking, we select the top 100 sentences for each movie. We then randomly choose two sentences from these top 100 for each scenario. In other words, two different sentences (i.e., selling points) may be used for each scenario, even for the same movie. It is noted that we use the word2vec (Mikolov et al., 2013) model, trained on roughly 9.8 billion sentences from web texts, to obtain these word vectors.

Utterance $S5$ Lastly, the system encourages the user to watch the recommended movie. This final utterance is randomly selected from the following five utterances.

- Please take a look.
- Please watch it.

[2] `https://github.com/ku-nlp/jumandic-grammar`

- Would you consider watching it if you're interested?
- It's an interesting movie, so I sincerely hope you'll watch it.
- I believe you'll certainly enjoy it.

2.3.4 Dialogue Manager

The dialogue manager selects a recommended movie and a scenario and then produces utterances based on that scenario.

There are two methods for selecting the recommended movie: the random method and the initial question method. The distribution between these methods is 80% for the random method and 20% for the initial question method. The former method chooses a movie at random. The latter asks a user's movie preference and then chooses a movie according to the user's answer. This initial question asking for user preferences is randomly selected from the following list:

- Who is your favorite actor?
- Who is your favorite actress?
- Who is your favorite director?
- What is your favorite movie genre?
- Do you prefer Japanese movies or foreign movies?

If a user mentions their favorite actress, the system identifies a movie featuring that actress. It then picks a scenario from a set of predefined scenarios and produces utterances based on that scenario. The system uses regular expressions to identify personal names in user responses. When the user says something like "None in particular," the recommended movie is determined randomly.

When users ask simple questions about a recommended movie, like the cast or director's name, or its genre, the system checks the movie database and inserts the answer before the original response. To determine if a user utterance is a question, we utilize the COTOHA API.[3] This API can identify whether a sentence is declarative, interrogative, or imperative. In our study, we treat sentences

[3] `https://api.ce-cotoha.com/contents/index.html`

User Internal State Type	Response Change Summary
I. No knowledge of a person	Add brief profile about the person
II. No knowledge of a movie	Add the movie information (the release year)
III. Knowledge of a movie	Change the end of the utterance to a consenting tone
IV. Deep knowledge of a movie	Change the utterance to one that assumes the user has watched the movie
V. No interest in news	Add supplemental description
VI. No interest in a movie theme	Change the utterance to a question that asks the user's preference
VII. No interest in a person	Change the utterance to a question that asks the user's preference
VIII. No willingness	Change the utterance to a modest tone

Table 2.1: Summary of response change per user internal state type.

classified as interrogative by the API as questions. We then identify the specific topic of the question, such as the cast's name, using regular expression matching.

2.3.5 Response Change based on User Internal State

In this section, we describe the rules for changing the system's response based on the results of the user internal state estimator. The user internal state estimators are created by the following three procedures:

1. Collect dialogues using the movie recommendation dialogue system described in Sections 2.3.1 to 2.3.4. (See Section 2.4.1 for details.)

2. Annotate the collected dialogues with the user's level of knowledge, interest, and willingness. (See Section 2.4.3 for details.)

3. Train the user internal state estimators using the annotated dialogues. (See Section 2.5.1 for details.)

Our movie recommendation dialogue system follows a predefined scenario, which specifies system utterances in advance. Therefore, we can guess the specific target of each user internal state (such as "what the user has knowledge about" and "what the user is not interested in") from the previous system utterances. We predefine appropriate responses according to the degree of each user internal state and the specific focus of each state. Table 2.1 shows the summary of response change per user internal state type. We aim to make system responses more natural by estimating the user internal state and accordingly modifying the response.

Knowledge-based Response Change

There are four types of knowledge-based response change: I. No knowledge of a person, Ⅱ. No knowledge of a movie, Ⅲ. Knowledge of a movie, and Ⅳ. Deep knowledge of a movie.

I. No knowledge of a person If the user lacks knowledge in $U1$ in the pattern **P3**, it is considered that the user does not know the person (i.e., casts or director). In this case, the system adds a brief profile of that person before $S2$. This profile is sourced from the first sentence of the person's Wikipedia article, obtained using the MediaWiki API.[4] Example (2) shows a dialogue example. The subscripts *before* and *after* denote the case where the response is not changed and where the response is changed, respectively. We underline the main difference in S_{after}. The angle bracket in the example shows the user internal state, which is the basis for judging the response change.

(2) $S1$: ジョージ・ルーカスさんをご存知ですか？ (Do you know George Lucas?)
$U1$: 良く分かりません。 (I'm not sure.) ⟨**No knowledge**⟩
$S2_{before}$: ジョージ・ルーカスさんが監督をされている映画があります。「スター・ウォーズ」というタイトルです。 (There is a movie directed by George Lucas. The title is "Star Wars.")
$S2_{after}$: <u>ジョージ・ルーカスさんはアメリカの映画監督、映画プロデューサー、脚本家</u>です。ジョージ・ルーカスさんが監督をされている映画があります。「スター・ウォーズ」というタイトルです。(<u>George Lucas</u>

[4] https://www.mediawiki.org/wiki/API:Main_page/ja

is an American film director, producer, and screenwriter. There is a movie directed by George Lucas. The title is "Star Wars.")

II. No knowledge of a movie If the user lacks knowledge in $U2$, it is assumed that the user does not know the recommended movie. In this case, we insert the released year before $S3$, as in Example (3).

(3) $S2$: 「本能寺ホテル」という映画です (It's called "Honnōji Hotel.")
$U2$: 知らないです。(I don't know that movie.) ⟨**No knowledge**⟩
$S3_{before}$: 最後の京都の現在の町と過去が交差する景色は、人の営みと思いは綿々と繋がっていること、そして過去の人々のお陰で現在の平和があることを示唆していますよ (The last scene in Kyoto, where the present city and the past city intersect, suggests that people's activities and thoughts are connected to each other and that the peace we have today is only possible because of the people of the past.)
$S3_{after}$: 2017 年に公開された作品です。最後の京都の現在の町と過去が交差する景色は、人の営みと思いは綿々と繋がっていること、そして過去の人々のお陰で現在の平和があることを示唆していますよ (This movie was released in 2017. The last scene in Kyoto, where the present city and the past city intersect, suggests that people's activities and thoughts are connected to each other and that the peace we have today is only possible because of the people of the past.)

III. Knowledge of a movie If the user possesses knowledge in $U2$ or $U3$, the system changes the end of the next utterance ($S3$ or $S4$) to a consenting tone (See Example (4)). The response changes in the case of $S3$ and $S4$ are treated separately.

(4) $S2$: 「僕だけがいない街」という映画です (The title is "Erased.")
$U2$: はい、ＤＶＤで鑑賞しました (Yes, I watched it on DVD.) ⟨**Knowledge**⟩

$S3_{before}$: 本作は物語のベースに暖かいメッセージがこめられていて感動しますよ (This work has a warm message at its core, and it's touching.)

$S3_{after}$: 本作は物語のベースに暖かいメッセージがこめられていて感動しますよね (This work has a warm message at its core, and it's touching, isn't it?)

IV. Deep knowledge of a movie As shown in Example (5), if the user has knowledge in all of $U2$, $U3$, and $U4$, we assume that the user has already watched the recommended movie. Then, the system changes $S5$ to the utterance that assumes the user has watched the movie. This utterance is randomly selected from the following:

- また見てみてはいかがですか (You may want to watch this movie again.)
- 是非もう一度見てみてください (Please watch it again.)

(5) $S1$: じゃあ薬師丸ひろ子さんが出演している作品を紹介しますね (Then I'll introduce the movie in which Hiroko Yakushimaru appears.)

$U1$: よろしくお願いします (Thank you.)

$S2$: 「コーヒーが冷めないうちに」というタイトルです (The title is "Cafe Funiculi Funicula.")

$U2$: よく知っていますね (You know it well.) ⟨**Knowledge**⟩

$S3$: 設定のせいもあるかもしれないけど、やっぱり、本物の役者さんって言うのは、凄いのだと、改めて感じますよね (While it might be attributed to the setting, one is compelled to acknowledge the remarkable talent of genuine actors.)

$U3$: 年季を感じますね (One indeed senses their years of experience.) ⟨**Knowledge**⟩

$S4$: 数ちゃんと新谷君の心が近づいていって結ばれる軸となるエピソードに、ほっと心が温まりますよね (The episode where Kazu-chan and Shintani-kun get closer and succeeds in love is very heartwarming, isn't it?)

$U4$: 私もです (I think so too.) ⟨**Knowledge**⟩

$S5_{before}$: 面白い作品なのでぜひ見て頂きたいです (It's an interesting movie, so I sincerely hope you'll watch it.)

$S5_{after}$: 是非もう一度見てみてください (Please watch it again.)

Interest-based Response Change

We prepare three types of interest-based response change: V. No interest in news, Ⅵ. No interest in a movie theme, and Ⅶ. No interest in a person.

V. No interest in news If the user has no interest in $U1$ in the pattern **P1**, it is assumed that the user is not interested in the presented news. However, the user may still be interested in the presented movie. Thus, the system continues to recommend the same movie. In order to reduce the effect of ignoring the user's utterance, the system randomly selects and inserts one of the following utterances before $S2$ (See Example (6)).

- 結構評判みたいですよ。(It seems to be quite well-known.)
- 結構話題になっているみたいなのですが，(It seems to be quite a hot topic, but ...)

(6) $S1$: 俳優の木村拓哉と歌手の工藤静香の長女・Ｃｏｃｏｍｉが26日、自身のインスタグラムを更新し、久々に外出したことをつづったと話題となっています (It's a hot topic that Cocomi, the eldest daughter of actor Takuya Kimura and singer Shizuka Kudo, reported that she had been out after a long time on her Instagram.)

$U1$: そうなん (I see.) ⟨**No interest**⟩

$S2_{before}$: 木村拓哉さん、「無限の住人」という映画に出演されています (Takuya Kimura is starring in the movie "Blade of the Immortal.")

$S2_{after}$: 結構評判みたいですよ。その木村拓哉さん、「無限の住人」という映画に出演されています (It seems to be quite well-known. Takuya Kimura is starring in the movie "Blade of the Immortal.")

VI. No interest in a movie theme If the user has no interest in $U1$ in the pattern **P2**, the user is probably not interested in the movie's theme. In this case, the system changes the recommended movie. As shown in Example (7), the system asks an initial question to understand the user's preferences. This initial question is randomly selected from a list of candidates described in Section 2.3.4.

(7) $S1$: タイムトラベルって興味ありますか？ (Are you interested in time travel?)
$U1$: いえ、あまり興味ありません (No, I'm not very interested.) ⟨**No interest**⟩
$S2_{before}$: タイムトラベルに関係のある映画があります。「アバウト・タイム 愛おしい時間について」というタイトルです (There is a movie related to time travel. The title is "About Time.")
$S2_{after}$: そうなんですね。では好きな映画監督は誰ですか？ (I see. Then, who is your favorite movie director?)

VII. No interest in a person If the user has no interest in $U1$ in the pattern **P3**, the user is probably not interested in the person. The system then changes the recommended movie by asking an initial question. In this case, if the system starts the dialogue by mentioning the name of the actress (actor/director), it then inquires about the user's favorite actress (actor/director). We show an example in Example (8).

(8) $S1$: サンドラ・ブロックさんをご存知ですか？ (Do you know Sandra Bullock?)
$U1$: 知っていますが、あまり興味はないです。 (I know, but I'm not so interested in her.) ⟨**No interest**⟩
$S2_{before}$: サンドラ・ブロックさんが声優をしている映画があります。タイトルは「ミニオンズ」です。 (There is a movie in which Sandra Bullock does a character voice. The title is "Minions.")
$S2_{after}$: そうなんですね。では好きな女優は誰ですか？ (I see. Then, who is your favorite actress?)

Willingness-based Response Change

There is one type of willingness-based response change: VIII. No willingness.

Ⅷ. No willingness If the user does not have willingness in $U4$, it is believed that the system's recommendations have been less effective. Example (9) shows an example. In this case, the system changes $S5$ to a modest tone. This system utterance is randomly chosen from the following list:

- 騙されたと思って見てみてはいかがですか (Just take my word for it and try to watch.)
- 見てみたら意外と面白いかもしれませんよ (You may find it unexpectedly interesting.)

(9)
- $S1$: ヘイリー・アトウェルさんをご存知ですか？ (Do you know Hayley Atwell?)
- $U1$: しらないな。(I don't know.)
- $S2$: ヘイリー・アトウェルさんはイングランド生まれの女優です。ヘイリー・アトウェルさんが出演されている映画があります。タイトルは「プーと大人になった僕」です。 (Hayley Atwell is an English actress. There is a movie in which Hayley Atwell does a character voice. The title is "Christopher Robin.")
- $U2$: そうなんですね。(I see.)
- $S3$: はい、プーだけじゃなくて、イーヨーやピグレット、他のキャラたちも、声に魅力があるし、ユアンマクレガーのイギリス人らしいイントネーションも、原語ならではの雰囲気ですよ (Yes, not only Pooh, but Eeyore, Piglet, and the other characters have charm in their voices, and Ewan McGregor's British intonation is a unique atmosphere of the original language!)
- $U3$: へー (Hmm.)
- $S4$: クリストファーロビンに邪魔者扱いされ、悲しそうな顔から、最後は絆を取り戻せて幸せそうな表情に変わるプーさんがキュートすぎてたまりませんよ (Pooh is so cute that his face changes from sadness when Christopher Robin treats him like an obstacle to happiness at the end when he regains his bond with Christopher Robin.)
- $U4$: そうですか。(I see.) ⟨**No willingness**⟩

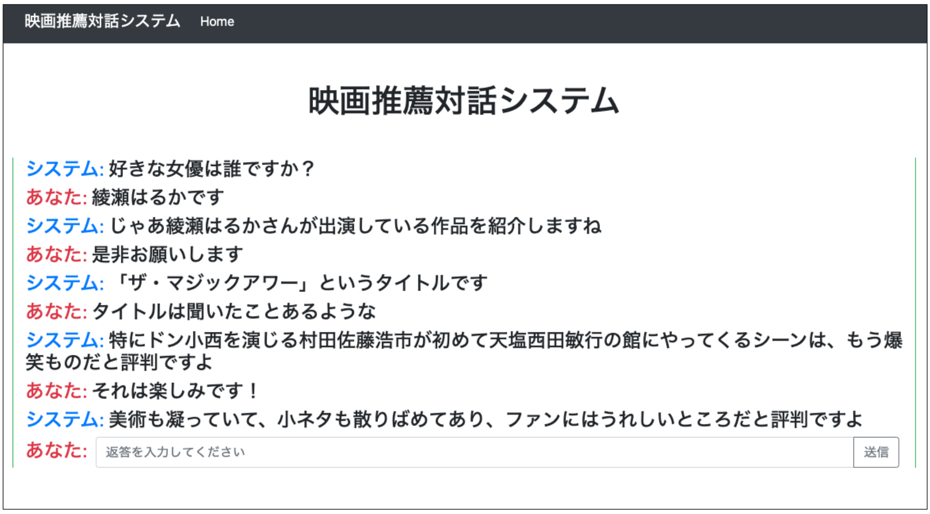


Figure 2.3: The screenshot of the dialogue system.

$S5_{before}$: 是非見てみてください (Please take a look.)
$S5_{after}$: 見てみたら意外と面白いかもしれませんよ (You may find it unexpectedly interesting.)

2.4 Corpus Construction

First, we collect dialogues between the baseline system, which does not change responses based on the user internal state, and users. Subsequently, we model the user internal state based on the collected dialogues. Then, we annotate the collected dialogues with the modeled user internal state.

2.4.1 Dialogue Collection for Corpus Construction

We collect dialogues via crowdsourcing[5] using the baseline system. Workers are informed in advance that they talk with the dialogue system designed to recommend movies from a selection of 213 movies.

Figure 2.3 shows the screenshot of the dialogue system. Workers access the dialogue system's website and participate in dialogues. After the dialogue, workers

[5] http://crowdsourcing.yahoo.co.jp/

Dialogue	K	I	W
$S1$: お姫さまって興味ありますか？ (Are you interested in princesses?)	-	-	-
$U1$: 興味あります (I'm interested in it.)	2	3	3
$S2$: お姫さまに関係のある映画があります。「今夜、ロマンス劇場で」というタイトルです (There is a movie related to princesses. Its title is "Color Me True.")	-	-	-
$U2$: どんなストーリーですか？ (What is it about?)	-3	3	3
$S3$: 内容は詳しく言えませんが、泣ける映画です。リアリズムを徹底的に求めるのではなく、綾瀬はるかのセリフと音楽を楽しむのが良いと思いますよ (I can't tell you the details, but it's a moving movie. I think it's better to enjoy Haruka Ayase's lines and music instead of thoroughly seeking realism.)	-	-	-
$U3$: 音楽興味ありですね (I'm interested in music.)	0	3	3
$S4$: 悲劇ではないけど切ない結末だなと思いきや、白黒からフルカラーへの反転で大団円ですよ (It's not a tragedy, but I thought it had a melancholic ending. However, it turns into a grand finale with the shift from black and white to full color.)	-	-	-
$U4$: 素敵な展開ですね (That's a nice development.)	-2	3	3
$S5$: 是非見てみてください (Please take a look.)	-	-	-

Table 2.2: An example of the dialogue corpus. K, I, and W represent knowledge, interest, and willingness of the user internal state, respectively.

are required to answer the following questionnaire on a five-point Likert scale.

(1) **Recommendation Success**: Would you like to watch the recommended movie?
{5: want to watch, 4: somewhat want to watch, 3: neutral, 2: somewhat do not want to watch, 1: do not want to watch}

(2) **Naturalness**: Was the flow of the dialogue natural?
{5: natural, 4: somewhat natural, 3: neutral, 2: somewhat unnatural, 1: unnatural}

Table 2.2 shows an example of the collected dialogue, and Table 2.3 shows the statistics of the collected dialogues. We collected 1,060 dialogues in total.

# dialogues	1,060
# scenarios	836
# workers	432
Avg. # utterances per dialogue	10.6
# utterances (R / S)	6,154 / 5,094
# unique utterances (R / S)	4,840 / 2,485
# morphemes (R / S)	163,347 / 20,279
# unique morphemes (R / S)	5,123 / 1,786

Table 2.3: Statistics of the collected dialogues. The number of workers is calculated based on the Yahoo! Crowdsourcing member ID. R and S denote the recommender and seeker, respectively. For morpheme segmentation, we use Juman++ (Morita et al., 2015; Tolmachev et al., 2018).

Our dialogue system encompasses a substantial number of scenarios, totaling 836, which has enabled the collection of a relatively diverse set of dialogues.

Figure 2.4 shows the results of the questionnaire. For recommendation success, 51.7% (16.0% + 35.7%) of the dialogues were rated 5 or 4, while for naturalness, 60.3% (23.0% + 37.3%) received these ratings. These results suggest that while the system effectively fulfills the objective of movie recommendation, it is also capable of facilitating dialogues with a reasonable degree of naturalness.

2.4.2 Modeling User Internal State

We analyze the user internal states based on collected dialogues to enhance the naturalness of dialogues. As a result, we propose that dialogue systems should understand users' "knowledge," "interest," and "willingness" and respond appropriately according to each state for more natural dialogues. Prior studies have recognized the significance of these aspects separately (Miyazaki et al., 2013; Inaba and Takahashi, 2018), but integrating them could offer a more comprehensive insight into the user's intentions. This integrated approach can express a richer representation of the user internal state, not just in movie recommendations but across various domains.

Each user internal state in this study is defined as follows.

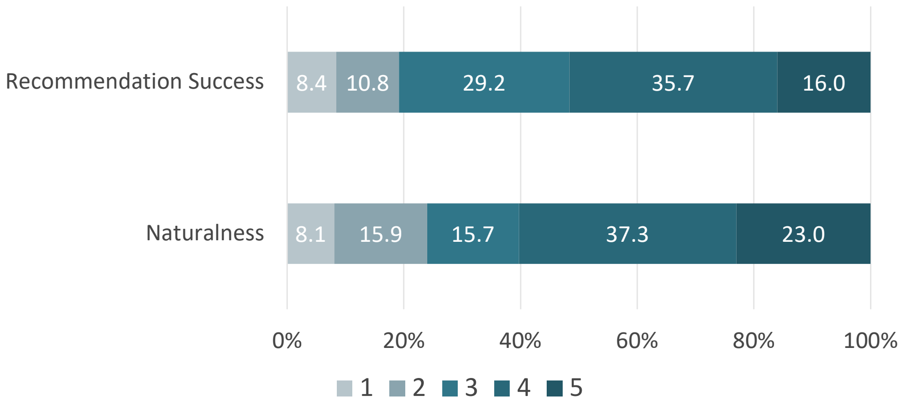


Figure 2.4: The results of the questionnaire for the collected dialogues.

Score	Knowledge (%)	Interest (%)	Willingness (%)
3	13.4 (684)	20.2 (1,030)	20.4 (1,039)
2	15.3 (781)	22.2 (1,130)	19.7 (1,006)
1	15.6 (793)	18.2 (929)	17.8 (906)
0	14.4 (735)	13.3 (680)	14.1 (716)
-1	15.8 (807)	11.4 (579)	12.0 (609)
-2	14.0 (711)	8.7 (443)	9.5 (486)
-3	11.4 (583)	5.9 (303)	6.5 (332)

Table 2.4: The distribution of the user internal state annotation. The number in parentheses represents the number of utterances.

- **Knowledge**: Whether the user has knowledge of the topic.
- **Interest**: Whether the user has an interest in the topic.
- **Willingness**: Whether the user actively participates in the dialogue.

We assume that the user internal state dynamically changes in dialogue. Thus, we track that state every time a user utters in this study.

2.4.3 Annotation of User Internal State

Using crowdsourcing, we annotate all of the user utterances in the collected dialogues with the user internal state. The workers annotate the user internal states (i.e., knowledge, interest, and willingness) with the target user utterance referring to the dialogue context. Each internal state has three levels: high (scored as 1), neutral (scored as 0), and low (scored as -1).

458 workers participated in the annotation process. Each utterance was annotated by three workers, and their scores were summed. Instead of using majority voting to aggregate scores, we opted for a summative approach. This method was chosen to maintain the diversity of viewpoints among the workers. As a result, the annotated score for each utterance ranges from -3 to 3.

Table 2.2 shows an example of the annotated dialogue corpus, and Table 2.4 shows the score distribution for each internal state. The scores for interest and willingness tended to be high, whereas the score for knowledge was distributed almost uniformly.

We acknowledge that the annotation of user internal states might be influenced by the annotators' subjectivity. To ensure the reliability of the annotations, we measure the agreement among annotators.[6] For measuring agreement, we use Krippendorff's α (Krippendorff, 2004). Krippendorff's alpha is a versatile metric that quantifies the degree of agreement among two or more annotators. It can be adapted to different types of scales (Stevens, 1946) by adjusting the score distances (i.e., the degree of disagreement). In this study, we adopt the difference function for the ordinal scale.

The *All* column of Table 2.5 shows the agreements among annotators in all data (called *All*). The α values for any internal states were around 0.40. In sociology, an α value above 0.80 is typically considered indicative of reliable data. In the field of natural language processing, Komatani et al. (2018) have annotated the multimodal dialogue data with participants' interest and have found that the α is around 0.50 using the interval scale as the scale type. While this result can

[6] We investigate the annotation reliability for each annotator in Appendix A.1. The results show that there are very few annotators with low reliability. Therefore, we use all the annotations as they are in this study.

User Internal State	*All*	*Filtered*
Knowledge	0.41	0.67
Interest	0.40	0.59
Willingness	0.35	0.63

Table 2.5: The agreement among annotators for each internal state. *All* and *Filtered* denote the agreement among annotators for all data and filtered data, respectively.

not be directly compared due to the use of different scales, their annotation is conducted by experts and is considered to be of higher quality than those by crowdsourcing, which allows for the participation of an unspecified number of people. Chollet et al. (2016) conduct the annotation through crowdsourcing for the presentation skills of native and non-native speakers, reporting that an α of approximately 0.40 is acceptable for subjective evaluation in natural language processing tasks. Therefore, it is reasonable that our data's α, whose data is annotated by crowdsourcing, is about 0.40.

Moreover, we prepare another set of data, which is filtered by excluding utterances that contain both 1 and -1 scores. We call this data *Filtered*. The benefit of using the *Filtered* dataset is that it consists of high-quality data with a high agreement among annotators. However, a downside is that it omits more complex cases where the annotators disagreed, potentially limiting the system's ability to handle such scenarios in real-world applications. We experimentally compare the accuracy of estimators trained by each of *All* and *Filtered* to verify the effect of filtering. For reference, the inter-annotator agreement for the *Filtered* data is presented in the *Filtered* column of Table 2.5.

2.5 Experiment

We construct a dialogue system that can change its response according to the estimation results of the user internal state. First, we train the user internal state estimator using the annotated dialogue corpus. We then incorporate the estimators into the dialogue system so that the system can change its response according

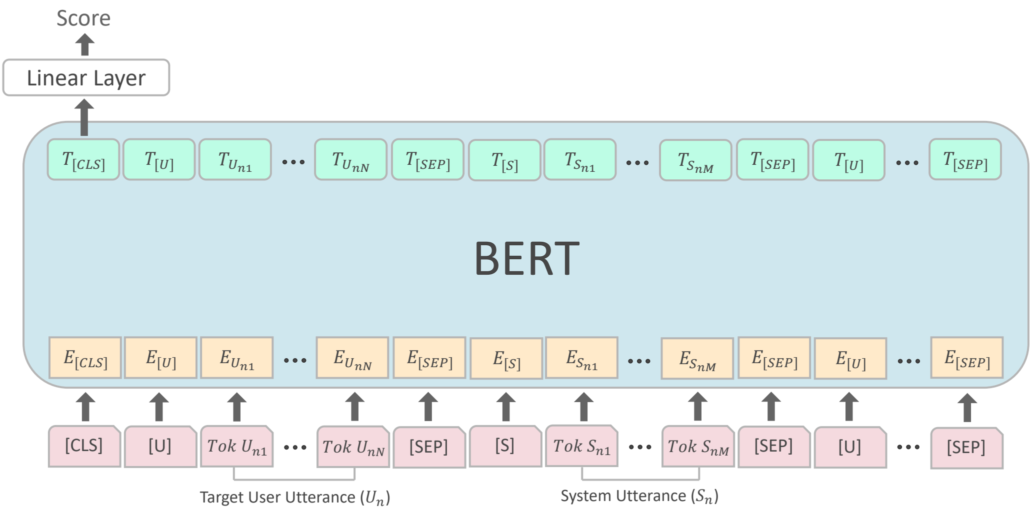


Figure 2.5: Overview of user internal state estimator. [S] and [U] denote the separation special tokens for system and user utterances, respectively.

to the user internal state. Finally, we collect dialogues using the constructed dialogue system with these response changes and evaluate its effectiveness.

2.5.1 User Internal State Estimation

Training of User Internal State Estimator

We fine-tune BERT (Devlin et al., 2019) model to estimate the user internal states. BERT is based on the Transformer (Vaswani et al., 2017) architecture and has achieved outstanding performance in various natural language processing tasks by pre-training on a large-scale raw corpus and then fine-tuning on downstream tasks. We use the NICT BERT Japanese pre-trained model (with BPE (Sennrich et al., 2016))[7] in this study. This BERT model has been pre-trained for 1.1 million steps after conducting morphological and subword segmentation on the entire text of the Japanese Wikipedia.

Figure 2.5 shows an overview of the user internal state estimator. The model is fed with the target user's utterance and the dialogue context, and it outputs

[7] https://alaginrc.nict.go.jp/nict-bert/index.html

User Internal State	*All*	*Filtered*
Knowledge	5,094 (4,082/511/501)	4,073 (3,266/410/397)
Interest	5,094 (4,082/511/501)	4,292 (3,424/432/436)
Willingness	5,094 (4,082/511/501)	3,926 (3,134/396/396)

Table 2.6: Number of utterances in *All* and *Filtered* data. The numbers in parentheses represent the number of utterances in the training, development, and test data, respectively.

the user's internal state estimation score (from -3 to 3) for the target user's utterance. Specifically, the [CLS] token is inserted at the beginning of the target user's utterance, and the dialogue context is input in reverse chronological order, up to a maximum of 512 tokens. The separation token [S] is inserted before system utterances and [U] before user utterances. We also insert a [SEP] token after all utterances. The final output is a real-valued scalar, derived from linearly transforming the vector associated with the [CLS] token. In post-processing, any scores above 3 are adjusted down to 3, and those below -3 are brought up to -3. We use mean squared error as the loss function, aiming to minimize the difference between the estimated and actual scores. We train three estimators for each user internal state (i.e., knowledge, interest, and willingness).

Experimental Settings

We randomly split 1,060 dialogues into three sets for different purposes: 80% for training, 10% for development, and 10% for testing. Table 2.6 shows the number of utterances in each data. The *Filtered* data is approximately 80% of the *All* data in volume.

We train and test the models with the following four conditions:

- ***AA***: Trained and tested on the *All* data.
- ***AF***: Trained on the *All* data and tested on the *Filtered* data.
- ***FA***: Trained on the *Filtered* data and tested on the *All* data.
- ***FF***: Trained and tested on the *Filtered* data.

To assess the performance of our estimators, we use the following evaluation metrics:

- **Strict Accuracy (Strict Acc)**: The percentage of cases where the difference between the estimated and actual scores lies within a narrow margin of ± 0.5. It is particularly suited for precise evaluations on a 7-class classification.

- **Loose Accuracy (Loose Acc)**: The percentage of cases where the difference between the estimated and actual scores lies within a wider margin of ± 1.5. This allows for slightly larger errors than Strict Acc.

- **Pearson Correlation Coefficient (Pearson)**: The Pearson's correlation coefficient between the estimated and actual scores.

- **Spearman Rank Correlation coefficient (Spearman)**: The Spearman's correlation coefficient between the estimated and actual scores.

We conduct hyperparameter tuning in accordance with the settings of Devlin et al. (2019) under the following conditions. We choose the best estimator in Strict Acc on each development set. Note that the dropout probability is fixed at 0.1 throughout the experiments.

- **Batch size**: 16, 32

- **Learning rate (AdamW (Loshchilov and Hutter, 2019))**: $5e{-5}$, $3e{-5}$, $2e{-5}$

- **Number of epochs**: 2, 3, 4

Result

Table 2.7 shows the accuracies of user internal state estimators. In the results of AA, Strict Accs are about 30% and Loose Accs are about 70% to 80% for all user internal states. As indicated in Table 2.4, the majority baselines for knowledge, interest, and willingness are 15.6%, 22.2%, and 20.4%, respectively. These results suggest that our estimators can estimate each state with reasonably

User Internal State	*AA*	*AF*	*FA*	*FF*
Knowledge	29.1 / 73.7	31.5 / 76.1	30.9 / 74.5	36.3 / 80.6
Interest	32.9 / 82.0	36.0 / 85.8	31.5 / 79.0	33.5 / 83.0
Willingness	28.3 / 72.5	32.1 / 78.5	30.7 / 75.8	36.4 / 84.8

Table 2.7: The accuracies for user internal state estimation. The numbers in the cells represent the Strict Acc / Loose Acc.

User Internal State	*AA*	*AF*	*FA*	*FF*
Knowledge	0.738 / 0.730	0.794 / 0.769	0.735 / 0.718	0.796 / 0.762
Interest	0.770 / 0.734	0.807 / 0.763	0.758 / 0.718	0.796 / 0.747
Willingness	0.701 / 0.669	0.776 / 0.715	0.731 / 0.688	0.816 / 0.756

Table 2.8: The correlation coefficients for user internal state estimation. The numbers in the cells represent the Pearson / Spearman.

high accuracy. In *AF*, we can observe the improvement in both Strict Acc and Loose Acc compared to *AA*. This is probably because *Filtered* data has a higher annotation agreement, meaning the data is easy to recognize the presence or absence of each user internal state.

Next, we analyze the models trained on the *Filtered* data. Compared to *AA* and *AF* models respectively, the *FA* and *FF* models exhibit an improvement in Strict Acc and Loose Acc for knowledge and willingness while a slight decrease for interest. These results indicate that the filtering of the training data is effective in improving the accuracy of the internal state estimation. The effectiveness of filtering on interest was likely minor because the estimation accuracy of interest was inherently higher than that of knowledge and willingness. Another reason for the lower accuracy in interest estimation may be that the annotator agreement rate of interest after filtering was lower than that of knowledge and willingness, as shown in Table 2.5.

Table 2.8 shows the correlation coefficients for user internal state estimation. We can observe that both correlation coefficients are around 0.7 to 0.8, indicating that our estimators can estimate each user internal state with high accuracy.

	Knowledge			Interest			Willingness		
	w/ Q (success)	w/ Q (failure)	w/o Q	w/ Q (success)	w/ Q (failure)	w/o Q	w/ Q (success)	w/ Q (failure)	w/o Q
Topic Introduction	0.93 [8] (0.00, 1.43)	0.89 [55] (-0.35, 4.67)	0.92 [20] (-0.30, 4.96)	0.76 [11] (2.09, 0.89)	1.19 [60] (-0.32, 4.02)	0.87 [25] (0.04, 4.71)	0.78 [12] (1.92, 1.72)	0.94 [54] (-0.06, 4.51)	0.96 [20] (-0.20, 4.80)
Title Presentation	0.72 [13] (1.08, 5.91)	0.97 [45] (-0.40, 4.61)	0.88 [23] (0.43, 6.08)	0.79 [12] (2.00, 0.91)	0.98 [56] (-0.07, 3.67)	1.07 [26] (0.31, 2.22)	0.78 [13] (1.77, 2.53)	1.11 [50] (-0.16, 5.44)	1.11 [20] (0.35, 2.87)
Selling Point I	0.69 [11] (0.82, 5.56)	1.19 [49] (-0.41, 4.33)	1.15 [23] (-0.60, 3.73)	0.44 [13] (2.23, 1.36)	1.10 [52] (0.58, 4.05)	0.83 [22] (0.64, 2.34)	0.56 [10] (2.50, 0.28)	0.97 [46] (0.26, 5.31)	0.95 [19] (0.26, 4.54)
Selling Point II	0.71 [11] (1.27, 3.22)	1.35 [50] (-0.34, 4.02)	1.44 [20] (-0.65, 4.33)	0.73 [10] (2.50, 0.50)	0.94 [55] (0.71, 4.51)	0.81 [24] (0.79, 2.87)	0.60 [10] (2.50, 0.50)	1.01 [53] (0.53, 4.95)	0.92 [23] (0.74, 4.11)

Figure 2.6: The mean absolute errors between the estimated internal state scores and the actual scores. The numbers within square brackets denote the number of samples. Additionally, the numbers within parentheses in each cell represent the mean and unbiased variance of the actual scores for those samples, respectively. The color gradient within the table intensifies in correlation with the magnitude of the mean absolute error, visually emphasizing larger errors with a stronger color intensity.

Analysis

Our system, which utilizes pre-defined system utterances, allows us to analyze the estimation error for each scenario turn. The approach for choosing the recommended movie varies based on whether an initial question is used or not. Consequently, we categorize the discussion into three distinct patterns:

w/ Q (success) : When the system uses the initial question and decides the recommended movie according to the user's preference.

w/ Q (failure) : When the system uses the initial question but randomly decides the recommended movie. This might happen, for example, when the movie database does not have a recommended movie that matches the user's preference.

w/o Q : When the system does not use the initial question and randomly decides the recommended movie.

Figure 2.6 shows the mean absolute errors between the estimated internal state scores and the actual scores per dialogue turn in each scenario. For instance, in the "Title Presentation" line, the target utterance is the user utterance immediately after the system presents the movie title. The color gradient within the table intensifies in correlation with the magnitude of the mean absolute error, visually emphasizing larger errors with a stronger color intensity.

In the case of “w/ Q (success),” the errors for knowledge, interest, and willingness remain consistently low throughout the dialogue. Regarding interest and willingness, the average actual score is around two points, suggesting that “w/ Q (success)” leads to the acquisition of high interest and willingness. Moreover, the average actual score for knowledge is nearly one point higher compared to cases “w/ Q (failure)” and “w/o Q.” This may be because users have some movie knowledge of their favorite actors or directors. Consequently, since the average scores for these internal states are higher, it becomes easier for the estimators to make accurate predictions, resulting in smaller mean absolute errors.

On the other hand, the estimation error for knowledge in the cases of “w/ Q (failure)” and “w/o Q” increased as the dialogue progressed. Example (10) shows a dialogue example in this case. For the last user utterance, $U4$, the correct score was -2, whereas the estimated score was 2.22. For utterances such as $U4$, where the user internal state is not explicitly clear, the interpretation of the dialogue context becomes crucial. From the dialogue context in Example (10), we can observe that the user has not watched the recommended movie, as indicated in $U2$, and that the system utterance $S4$ immediately preceding $U4$ describes a scene in the recommended movie. In summary, it can be inferred that the user probably does not know about the movie scene they have not seen. However, making such an inference requires fine-grained tracking of user internal states, which leaves room for future research.

(10) $S1$: 俳優の染谷将太と女優の菊地凛子夫妻に第２子が誕生したことが１１日までにわかったと話題です (It’s a hot topic that the second child has been born to actor Shota Sometani and actress Rinko Kikuchi.)

$U1$: おめでたいです (Congratulations on that.)

$S2$: 染谷将太さん、「バクマン。」という映画に出演されていましたね (Mr. Shota Sometani appeared in a movie titled “Bakuman,,” did he not?)

$U2$: バクマン。は観ていませんが確か漫画原作ですよね。(I haven’t watched

"Bakuman," but I believe it's based on a manga.)

*S*3: 話しも素直に友情、青春、ライバル関係、生い立ち等も分かりやすく展開され好感が持てますよ (The story candidly unfolds elements of friendship, youth, rivalry, and backgrounds in an understandable manner, which makes it quite agreeable.)

*U*3: 映画は数年前でしたか。ＤＶＤで観てみたいと思います。 (The movie was released a few years ago, right? I would like to see it on DVD.)

*S*4: ちょっと分からないんですが、無駄なシーンが一切なく、サクサク進んでテンポがいいので、気持ちいいです (I am not so familiar with that, but the movie progresses without any superfluous scenes, briskly and with a good tempo, which is quite satisfying.)

*U*4: 期待できます。 (That sounds promising.)

2.5.2 Effectiveness of Response Change

Integration with User Internal State Estimator

We incorporate the user internal state estimators into the dialogue system. The system judges whether a user possesses knowledge, interest, and willingness at every dialogue turn and then changes its responses as described in Section 2.3.5. In the following experiments, we use the estimators trained on the *Filtered* data, which is expected to provide higher accuracy in overall estimation. We set the positive and negative thresholds to judge whether a user possesses knowledge, interest, and willingness. Each internal state is classified as "high" when its estimation score exceeds the positive threshold, and as "low" if it falls below the negative threshold. In this study, the positive and negative thresholds for knowledge and interest are set at 1.5 and -1.5, and those for willingness are set at 1.0 and -1.0. When multiple response change rules are to be applied, we prioritize the one linked to the internal state with the largest absolute value in the estimated score.

	w/ RC	*w/o RC*
Recommendation Success	3.44	**3.48**
Naturalness	**3.46**	3.20
Satisfaction	**3.34**	3.15

Table 2.9: Results of the questionnaire. The scores represent the average ratings given by the workers.

Dialogue-level Evaluation

We collect 299 dialogues via crowdsourcing using the dialogue system with response changes (*w/ RC*). We also collect 297 dialogues using the system without response changes (*w/o RC*) for comparison.[8]

After the dialogue, we asked the workers if they had watched the recommended movie. As a result, approximately 67% of all workers had not watched the recommended movie. In addition, we asked workers to answer the following 5-point Likert-scale questionnaires (5 is the best):

- **Recommendation Success**: The system has made you want to watch the recommended movie.
- **Naturalness**: The system responses were natural.
- **Satisfaction**: The system responses satisfied you.

Table 2.9 shows the results of the questionnaire. We did not observe a significant difference in Recommendation Success between *w/ RC* and *w/o RC*. On the other hand, *w/ RC* was 0.26 points higher than *w/o RC* on Naturalness and 0.19 points higher on Satisfaction. A Wilcoxon rank-sum test, conducted at a significance level of 5%, revealed p-values of 0.017 and 0.123 for Naturalness and Satisfaction, respectively, indicating a significant improvement in Naturalness. The results show that by estimating the user internal state and modifying

[8] We have updated some modules from the system used in the corpus construction. The update was applied to both the *w/ RC* and *w/o RC* systems, ensuring a fair comparison. Details are given in Appendix A.2.

User Internal State Type	*w/ RC*	*w/o RC*
I. No knowledge of a person	58	48
II. No knowledge of a movie	121	124
III. Knowledge of a movie	201	175
IV. Deep knowledge of a movie	56	42
V. No interest in news	5	5
VI. No interest in a movie title	19	15
VII. No interest in a person	2	0
VIII. No willingness	57	51
Total	519	460

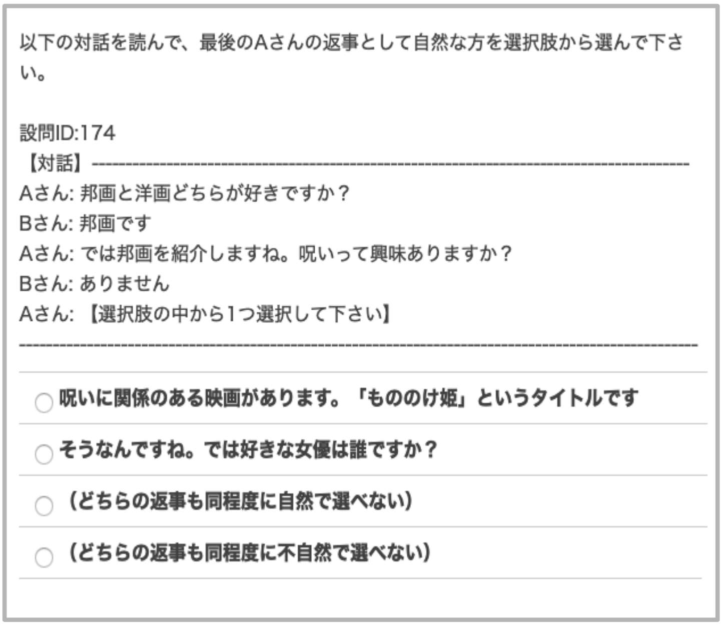
Table 2.10: Statistics of branch points for each user internal state type.

the response accordingly, our system can improve the naturalness of the system's utterances without sacrificing the accomplishment of the movie recommendation.

Utterance-level Evaluation

Our dialogue system is fundamentally scenario-based. Consequently, except for the branch points (i.e., dialogue turns where response change occurs), *w/ RC* and *w/o RC* produce identical responses. Thus, we compare the changed response with the original responses at the utterance level. Moreover, we have designated these branch points based on different types of user internal states, as detailed in Table 2.1. Therefore, this evaluation allows us to examine the effect of the response change on each user internal state type.

We use the collected dialogues in Section 2.5.2 to extract pairs of responses with and without response change. For a fair comparison, we extract the branch points from the dialogues involving both humans and *w/ RC*, as well as humans and *w/o RC*. In dialogues with *w/ RC*, we compare the original response from its scenario with the changed one. In dialogues with *w/o RC*, we identify the branch points using user internal state estimators and then compare the changed response at this point with the original response from its scenario. Table 2.10 shows the results of extracting pairs of responses for each user internal state type. We sample up to 15 pairs from dialogues between humans and *w/ RC*, and humans



以下の対話を読んで、最後のAさんの返事として自然な方を選択肢から選んで下さい。

設問ID:174
【対話】--------------------
Aさん: 邦画と洋画どちらが好きですか？
Bさん: 邦画です
Aさん: では邦画を紹介しますね。呪いって興味ありますか？
Bさん: ありません
Aさん: 【選択肢の中から1つ選択して下さい】

- ○ **呪いに関係のある映画があります。「もののけ姫」というタイトルです**
- ○ **そうなんですね。では好きな女優は誰ですか？**
- ○ **（どちらの返事も同程度に自然で選べない）**
- ○ **（どちらの返事も同程度に不自然で選べない）**

Figure 2.7: The user-interface for utterance-level evaluation.

and *w/o RC*, totaling a maximum of 30 pairs, which we then use for evaluation.

We conduct the utterance-level evaluation via crowdsourcing. Figure 2.7 shows the user-interface for this evaluation. Workers are shown the dialogue context, the changed response, and the original response, and are asked to select which response is more natural. The workers are not informed which response is the changed response. We also add two options: *Natural* ("Both responses are equally natural") and *Unnatural* ("Both responses are equally unnatural"). 10 workers evaluate each response pair.

Table 2.11 shows the results of the utterance-level evaluation. For all user internal state types, the naturalness of the responses generated by *w/ RC* outperformed those by *w/o RC*. These results indicate that our dialogue system can provide more natural responses by estimating the proposed user internal states and accordingly tailoring the responses.

User Internal State Type	*w/ RC*	*w/o RC*	*Natural*	*Unnatural*
I. No knowledge of a person (30)	**240**	33	11	16
II. No knowledge of a movie (30)	**180**	44	32	44
III. Knowledge of a movie (30)	**87**	84	50	79
IV. Deep knowledge of a movie (30)	**135**	111	11	43
V. No interest in news (10)	**54**	11	6	29
VI. No interest in a movie title (30)	**170**	71	6	53
VII. No interest in a person (2)	**11**	8	0	1
VIII. No willingness (30)	**134**	100	30	36
Total (192)	**1,011**	462	146	301

Table 2.11: The utterance-level evaluation results per each user internal state type. The numbers in the table indicate the total number of votes by workers, and the numbers in parentheses indicate the number of samples.

Error Analysis

Although the scores of *w/ RC* for user internal state types III and IV were higher than those of *w/o RC*, the effect was limited compared to the other speaker internal state types. Hence, we analyze the reasons for each of them.

III. Knowledge of a movie In this user internal state type, the ending of the utterance is changed from an informative tone ("よ", *yo*) to a consenting tone ("よね", *yone*). Considering that the number of votes for *Natural* was somewhat high, it can be inferred that changing one character at the end of the utterance did not make a big difference in the impression. The votes for *Unnatural* were also as high as those for *w/ RC* and *w/o RC*. This may be because the sentences extracted as selling points from the reviews did not match the dialogue context. Example (11) shows an example where 6 out of 10 workers judged *Unnatural*. In this example, just before changing the response, the user says, "I like the earlier movie better." in $U3$. Still, the system ignores that utterance, which makes the system's utterance seem unnatural. Future work will improve the method by retrieving selling points using dialogue contexts.

(11) *S*: 好きな映画監督は誰ですか? (Who is your favorite movie director?)

U: スピルバーグ (Spielberg.)

S: ではスティーヴン・スピルバーグさんが監督の映画を紹介しますね (Now I'll introduce a movie directed by Steven Spielberg.)

U: お願いします。 (Please do.)

S: 「ブリッジ・オブ・スパイ」という映画です (The movie is titled "Bridge of Spies.")

U: 知りませんが、どんな内容ですか。(I don't know it. What is it about?)

S: 2015 年に公開された作品です。スティーヴンスピルバーグ監督にコーエン兄弟の脚本、主演がトムハンクスと期待せずにはいられない組み合わせに、見事に期待に応える良作だと評判ですよ (This movie was released in 2015. Directed by Steven Spielberg, written by the Coen brothers and starring Tom Hanks, it's a combination you can't help but expect, and it's a good film that lives up to its expectations as a great movie!)

U: もっと前の映画が好きです。 (I like the earlier movie better.)

S_{before}: ベルリンの壁、ベルリンにかかる橋、チェックポイント、東ベルリン街中などなど当時の感じを丁寧に表現撮影しており見ごたえがありますよ (The Berlin Wall, the bridges over Berlin, the checkpoint, and the streets of East Berlin are all carefully photographed to capture the atmosphere of those days.)

S_{after}: ベルリンの壁、ベルリンにかかる橋、チェックポイント、東ベルリン街中などなど当時の感じを丁寧に表現撮影しており見ごたえがありますよね (The Berlin Wall, the bridges over Berlin, the checkpoint, and the streets of East Berlin are all carefully photographed to capture the atmosphere of those days, aren't they?)

IV. Deep knowledge of a movie In this user internal state type, the system assumes that the user has already watched the recommended movie. Consequently, it uses responses like "Please watch it again." As shown in Table 2.11, there are not many votes for *Unnatural* but many votes for *w/o RC*. This is likely due to the estimation errors of the user's watching experience. In Example (12), the user has watched some of the "OUTRAGE" series, but has not watched the recommended movie, "OUTRAGE CODA." Like this example, suggesting "Please watch it again" to a user who has not watched the movie yet can appear quite unnatural.

(12) *S*: お笑いタレントのビートたけしが 30 日放送のＴＢＳ「新・情報７ＤＡＹＳニュースキャスター」にリモート生出演と話題となっています (Comedian Beat Takeshi has become a topic of discussion for his remote live appearance on the TBS program "Shin Joho 7DAYS Newscaster," which aired on the 30th.)

U: そうなんですか。知らなかったです (Is that so? I was not aware of that.)

S: はい、ビートたけしさん、「アウトレイジ最終章」という映画に出演されています (Yes, Beat Takeshi is featured in the movie "Outrage Coda.")

U: アウトレイジは見たことがあります。でも最終章はみてないかも (I have watched Outrage, but I might not have watched the "Coda.")

S: 韓国フィクサーの立ち位置が今一つで、タケシも今までより控えめですが、西田敏行のズル賢い役者ぶりが目立ちますよね (The positioning of the Korean fixer is somewhat unclear, and Takeshi's role is more subdued than before, but Toshiyuki Nishida's cunning portrayal as an actor is quite prominent.)

U: 韓国人がでてくるんですね (So, it includes a Korean character.)

S: 過去の出演者達がみんな歳をとって迫力不足だが、小心者のピエール瀧が今回は盛り上げてくれますよね (Although the past cast members have all aged and lack some of their former intensity, Pierre Taki, playing a timid character, provides a notable uplift to the film this time.)

U: ピエール瀧がアウトレイジというのはなんだか真実味があります (Pierre Taki's presence in Outrage seems quite authentic.)

Watching Experience	*w/ RC*	*w/o RC*	*Natural*	*Unnatural*
watched (23)	**122**	70	6	32
not watched (7)	13	**41**	5	11
Total (30)	**135**	111	11	43

Table 2.12: The votes for utterance naturalness for each user's watching experience. The numbers in the table indicate the total number of votes by workers, and the numbers in parentheses represent the number of samples.

S_{before}: 是非見てみてください (Please take a look.)
S_{after}: 是非<u>もう一度</u>見てみてください (Please watch it <u>again</u>.)

Next, we align 30 samples for this user internal state type (i.e., IV. Deep knowledge of a movie) with the questionnaire results of each user's watching experience, as described in Section 2.5.2. The results show that of the 30 samples, 23 were correctly estimated as "watched," while 7 were incorrectly estimated as "not watched." Table 2.12 shows the comparison results of naturalness for each user's watching experience. The votes of *w/ RC* were higher for the "watched" samples, but the votes of *w/o RC* were higher for the "not watched" samples. These results indicate that utterance naturalness was negatively affected by errors in estimating the user's watching experience.

2.6 Summary of This Chapter

In this study, we tackled the modeling of the user internal state in dialogue to appropriately interpret the user's intention. Based on the analysis of the collected dialogues, we modeled the user internal state along three axes: knowledge, interest, and willingness. We constructed a dialogue corpus by annotating the collected dialogues with the three modeled internal states via crowdsourcing. The user internal state estimator trained on our dialogue corpus demonstrated high accuracy from the target user utterance and dialogue context.

Furthermore, we constructed a dialogue system that changes its response according to the proposed user internal states. We designed response-changing rules according to the presence or absence of knowledge, interest, and willingness. Our system can produce more natural responses by using the estimation results of the trained user internal state estimator and appropriately modifying responses based on the designed rules.

With the recent development of neural network technology, neural network-based models are required to have explainability to clarify their behaviors and the reasons behind the predictions. We believe that our modeling of the user internal state will serve as a meaningful basis for understanding the behavior of neural network systems, and will provide a foundation for research into explainability.

Chapter 3

Response Generation Based on External Knowledge

3.1 Introduction

In recent years, research on dialogue-based recommendation systems, which suggest items to users through interactive dialogues, has gained significant attention. This study specifically concentrates on the domain of movie recommendations. A movie recommendation dialogue unfolds in two stages: (1) eliciting the user's preferences and selecting a movie from the candidates, and (2) providing detailed information about the chosen movie. This study specifically addresses this second stage.

To provide in-depth information, the use of external knowledge is crucial. There has been much research on incorporating external knowledge in dialogue, and many kinds of knowledge-grounded dialogue datasets have been proposed (Dinan et al., 2019; Liu et al., 2020). These datasets often use plain texts or knowledge graphs as external knowledge. If the hierarchically structured knowledge is available in recommendation dialogues, it allows for more appropriate knowledge selection and informative response generation. However, there is no dialogue dataset with hierarchically structured knowledge to provide rich information for a single target (e.g., a movie).

To address the aforementioned problem, we propose a dialogue dataset, Japanese

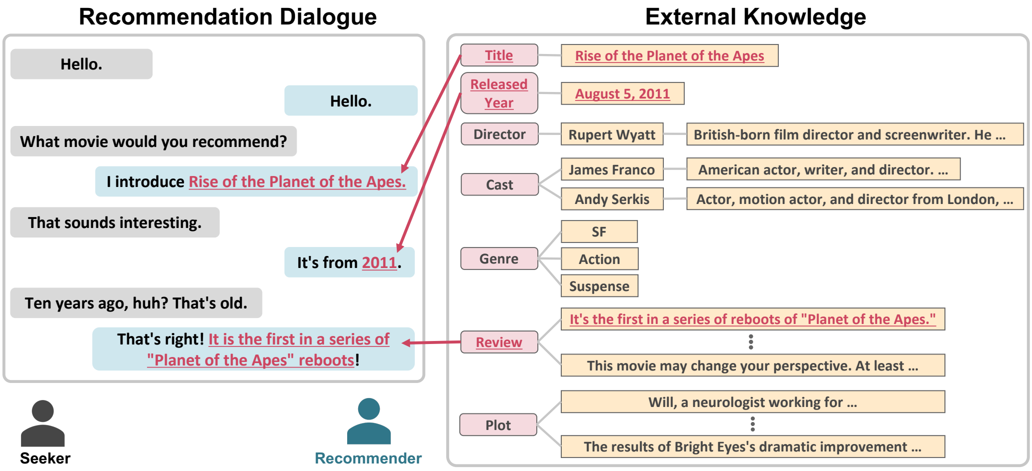


Figure 3.1: An example of JMRD dataset. The underlined parts of the external knowledge indicate the knowledge items used in the dialogue.

Movie Recommendation Dialogue (**JMRD**). This dataset consists of about 5,200 dialogues between crowd workers, each focused on movie recommendations. Each dialogue has 23 turns on average. Our dataset is unique in providing detailed movie recommendations, drawing on a wide range of movie-related information. It has a significant number of dialogue turns, enabling comprehensive discussions. Specifically, as shown in Figure 3.1, one speaker (recommender) recommends a movie to the other speaker (seeker). Only recommenders have access to movie-related knowledge, and they are encouraged to incorporate as much of this external knowledge as possible into their utterances. Furthermore, recommenders must note the specific movie information they refer to during the dialogue. This procedure associates every utterance of the recommenders with the corresponding external knowledge. The external knowledge is hierarchically structured into knowledge types common to all movies (e.g., "Title", "Released Year") and knowledge contents specific to each movie (e.g., "Rise of Planet of the Apes", "August 5, 2011").

We also propose a strong baseline model for our newly constructed dataset. This model considers the history of knowledge types/contents, noting that the order in which each piece of knowledge is used is essential in recommendation dialogues. The experimental results show that our proposed model can select appropriate knowledge with higher accuracy than the baseline method.

Our contributions are three-fold.

- We construct a movie recommendation dialogue dataset associated with hierarchically structured external knowledge.

- We propose a strong baseline model, which selects knowledge based on hierarchically structured knowledge, for our dataset.

- To the best of our knowledge, we are the first to construct a human-to-human dialogue dataset based on external knowledge in Japanese.

3.2 Related Work

The field of recommendation dialogue has been a topic of interest for some time. However, most studies have focused on task-oriented dialogues, where the goal is to understand a user's preferences from a selection of recommended options, and then make a recommendation based on those preferences (Bordes et al., 2017; Li et al., 2018). Li et al. (2018) propose REDIAL, a human-to-human movie recommendation dialogue dataset. In this setup, the recommender presents several movies in a dialogue while inquiring about the seeker's preferences. Similarly, Kang et al. (2019) collect GoRecDial dataset in a gamified setting. Here, experts decide on a movie similar to the seekers' preference from a limited selection of five movies in a minimal number of turns. OpenDialKG (Moon et al., 2019) is a recommendation and chit-chat dataset linking open-ended dialogues to knowledge graphs. In this study, we focus on the recommendation dialogue, which provides in-depth information about a movie rather than just deciding which movie to recommend.

Research on the knowledge-grounded dialogue has also been growing in the last few years. Zhou et al. (2018) collect a human-to-human chit-chat dialogue dataset by utilizing Wikipedia articles of 30 famous movies. This dataset is unique in that it has two dialogue settings: either only one of the participants can see the knowledge, or both of them can see it. Moghe et al. (2018) also collect chit-chat dialogues about movies based on multiple types of knowledge: plots, reviews, Reddit comments, and fact tables. Wizard of Wikipedia (Dinan et al., 2019) is an open-domain chit-chat dialogue dataset based on Wikipedia articles covering 1,365 topics. It has become a standard benchmark in this research field. Su et al. (2020) collect a large Chinese chit-chat dialogue dataset (246,141 dialogues with 3,010,650 turns) about movies. Other dialogue datasets with external knowledge in Chinese are DuConv (Wu et al., 2019), KdConv (Zhou et al., 2020), and DuRecDial (Liu et al., 2020). DuConv (Wu et al., 2019) combines dialogues with knowledge graphs to track the progress of the dialogue topic. KdConv (Zhou et al., 2020) is also a chit-chat dialogue corpus that consists of relatively long dialogues to allow deep discussions in multiple domains (movies, music, and travel). Liu et al. (2020) focus on multiple dialogue types (e.g., QA, chit-chat, recommendation) and collect a multi-domain dialogue dataset associated with a knowledge graph. Compared to these studies, our work stands out by utilizing hierarchically structured knowledge, encompassing both factoid (e.g., titles) and non-factoid (e.g., reviews) information for making recommendations.

3.3 Dialogue Collection Framework

To build a neural network-based dialogue system, thousands to tens of thousands of dialogue context and response pairs are required. Crowdsourcing is often used to collect dialogue data on this scale. ParlAI (Miller et al., 2017) is a dialogue collection framework that can pair workers in real time and collect dialogues between the paired workers. ParlAI can be connected to Amazon Mechanical Turk, making it easy to perform dialogue collection tasks via crowdsourcing. A number of datasets collected using ParlAI have been made publicly available (Dinan et al., 2019; Rashkin et al., 2019; Smith et al., 2020), contributing greatly to the development of dialogue research. However, it is difficult to collect Japanese dialogue on Amazon Mechanical Turk because it has few native Japanese speakers.

To this end, we built a dialogue collection framework, which incorporates crowdsourcing platforms where more native Japanese speakers can be gathered.

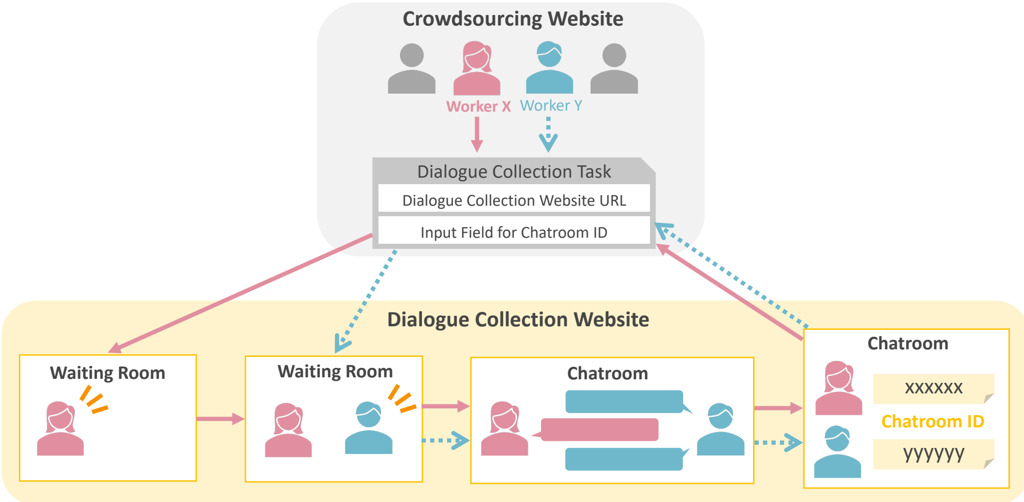


Figure 3.2: Dialogue collection flow using dialogue collection framework

Figure 3.2 shows the dialogue collection flow using our framework.[1]

First, a dialogue collection task is created on a crowdsourcing platform to gather workers. The task includes the dialogue collection website URL and the input field for chatroom ID (details to follow). A worker who participates in the task (e.g., worker X) clicks on the URL of the dialogue collection website and enters the waiting room. If there are no workers in the waiting room, worker X waits until another worker arrives. When another worker (e.g., worker Y) enters the waiting room, worker X and worker Y are automatically paired and moved to the chatroom.

Figure 3.3 shows the screenshot of the chatroom. Workers type their messages into the message box and press the send button. The dialogue ends upon pressing the end button at the bottom of the chatroom, and then each worker's unique chat room ID is displayed. After confirming their chatroom ID, each worker returns to the crowdsourcing platform to complete the task by entering the displayed chatroom ID. The chatroom ID is used as evidence that the worker has completed the dialogue, and is also used to link each worker to the collected dialogue.

[1]Our code is publicly available at `https://github.com/ku-nlp/ChatCollectionFramework`.

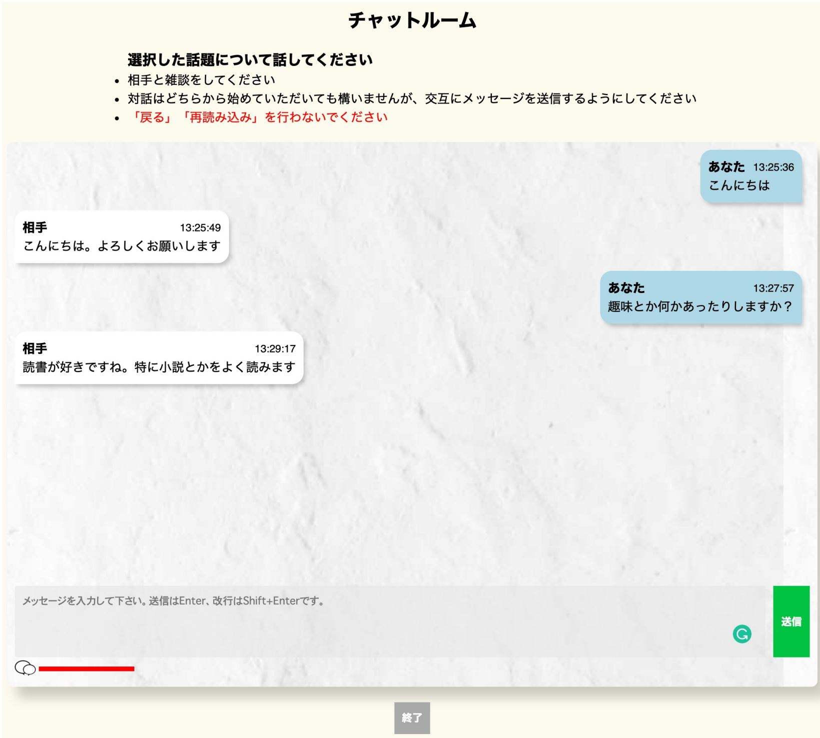


Figure 3.3: The screenshot of the chatroom

3.4 Japanese Movie Recommendation Dialogue

We chose movies as the domain for the recommendation dialogue because movies are generally interesting to everyone and facilitate smooth dialogue. In addition, movie recommendation dialogue is open-domain in nature due to the variety of movie topics, and it is a preferable property for NLP research. In this section, we explain the construction method of the JMRD.

3.4.1 External Knowledge Collection

The external knowledge is mainly collected from web texts such as Wikipedia. First, we select 261 movies based on the box-office revenue ranking.[2] For each of these movies, we collect movie information as external knowledge.

[2] http://www.eiren.org/toukei/index.html

The external knowledge consists of seven knowledge types: title, released year, director, cast, genre, review, and plot, as shown in Figure 3.1. The title, released year, director, cast, and plot are extracted from the Wikipedia article of each movie (we allow at most one director and two casts). For the director and the casts, a brief description is also extracted from the first paragraph of each person's Wikipedia article. For the genre, we use the genre classification of Yahoo! Movies.[3] Reviews are collected by crowdsourcing using Yahoo! Crowdsourcing.[4] Each worker selects a movie that he or she has seen from a list of 261 movies and writes down three recommendations for the selected movie. As a result, we collected an average of 16.5 reviews per movie.

We split the plot into sentences and present only the first ten sentences (or all sentences if fewer than ten) to reduce the burden of the recommender. Furthermore, we use the reviews written by the workers as they are, without splitting the sentences. We randomly selected five reviews between 15 and 80 characters long for each movie from the collected reviews. Those five reviews are used as the reviews for that movie.

3.4.2 Dialogue Collection

Settings

In dialogue collection, we use the chat collection framework described in Section 3.3. The two workers engaging in the movie recommendation dialogue have different roles: one is the **recommender**, and the other is the **seeker**.

Recommender The recommender first decides which movie to recommend. Figure 3.4 shows the recommender's chatroom before deciding which movie to recommend. A movie database is provided on the left side of the recommender's screen, allowing the recommender to search for movies by genre or text-based queries. The movie chosen for recommendation may be one that the recommender wishes to suggest, or it may be one that aligns with the seeker's preferences, as determined through a few message exchanges.

[3] https://movies.yahoo.co.jp/

[4] https://crowdsourcing.yahoo.co.jp/

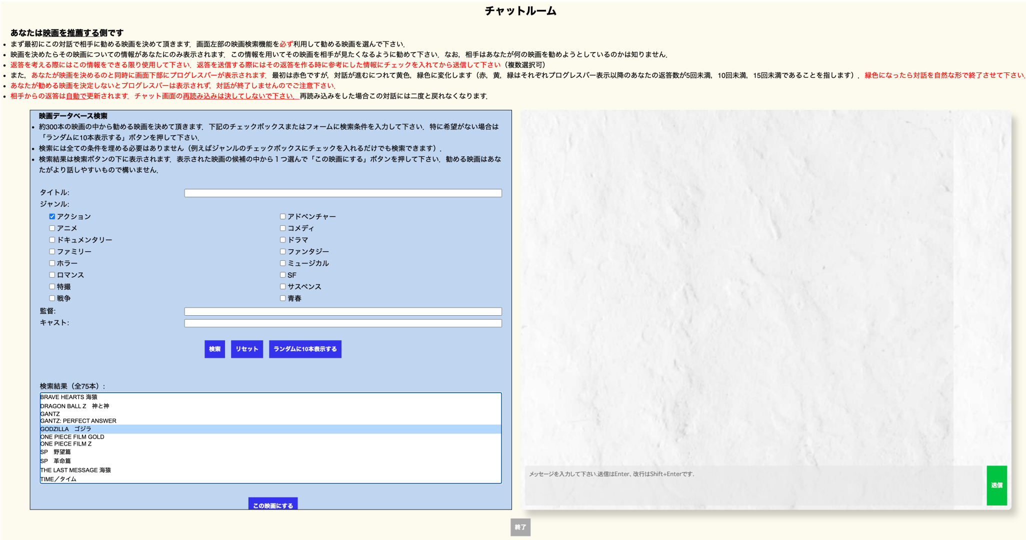


Figure 3.4: The screenshot of the chatroom for recommenders before they decide on the movie to recommend.

Figure 3.5: The screenshot of the chatroom for recommenders after they have decided on the movie to recommend.

The recommender can access the detailed movie knowledge after deciding which movie to recommend. The left side of Figure 3.5 shows an example of

Figure 3.6: The screenshot of the chatroom for seekers.

movie knowledge. The recommender is instructed to use the presented knowledge as much as possible to recommend the movie. When the recommender sends their utterance, they must select the knowledge referred to by the utterance (multiple selection is allowed). For the utterance that does not use any knowledge, such as greetings, the recommender can select the “no knowledge” option.

Seeker The seeker is only instructed to enjoy and learn more about the recommended movie, and they can talk freely. This instruction refers to that of Wizard of Wikipedia (Dinan et al., 2019). As shown in Figure 3.6, the seeker is only shown the chat screen and cannot access knowledge about the movie.

The dialogue can be initialized by either the recommender or the seeker. The dialogue extends for at least 20 turns following the selection of the movie.

Statistics

Table 3.1 shows the statistics of the collected dialogues. We collected a total of 5,166 dialogues. The recommender’s average number of words per utterance is more than three times larger than that of the seeker. This is likely because the recommender needs to talk more than the seeker to provide information to

# dialogues	5,166
# utterances (R)	57,714
# utterances (S)	59,160
# movies	261
# workers	322
Avg. # turns per dialogue	22.6
Avg. # words per utterance (R)	23.8
Avg. # words per utterance (S)	6.9
Avg. # knowledge used per utterance	1.3
Avg. # knowledge used per dialogue	10.8

Table 3.1: Statistics of JMRD. R and S denote recommender and seeker respectively. We use Juman++ (Morita et al., 2015; Tolmachev et al., 2018) for word segmentation.

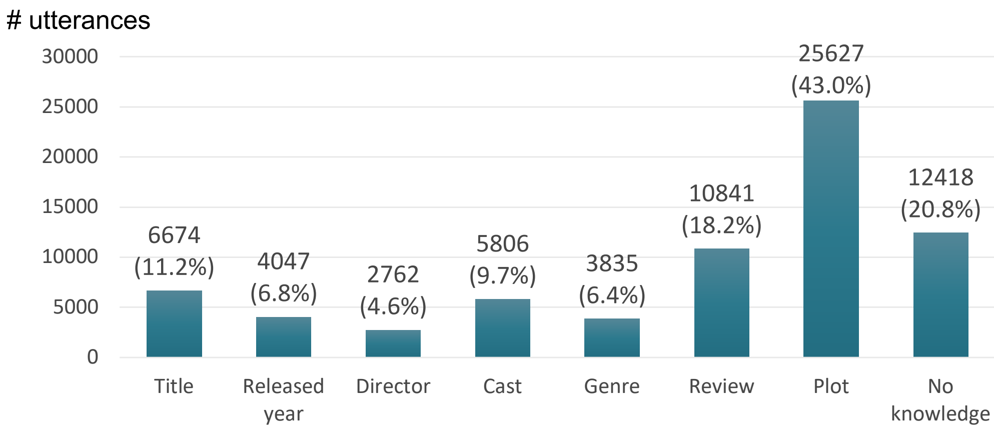


Figure 3.7: Distribution of external knowledge used.

recommend a movie. The average number of knowledge items per utterance is 1.3, and the recommender tends to mention each knowledge item separately. An average of 10.8 different types of knowledge was used per dialogue, indicating that we could collect dialogues with various types of external knowledge.

Figure 3.7 shows the distribution of the knowledge types used. The number of utterances that did not use any knowledge was only about 20% of the total, indicating that the majority of utterances incorporate some form of external

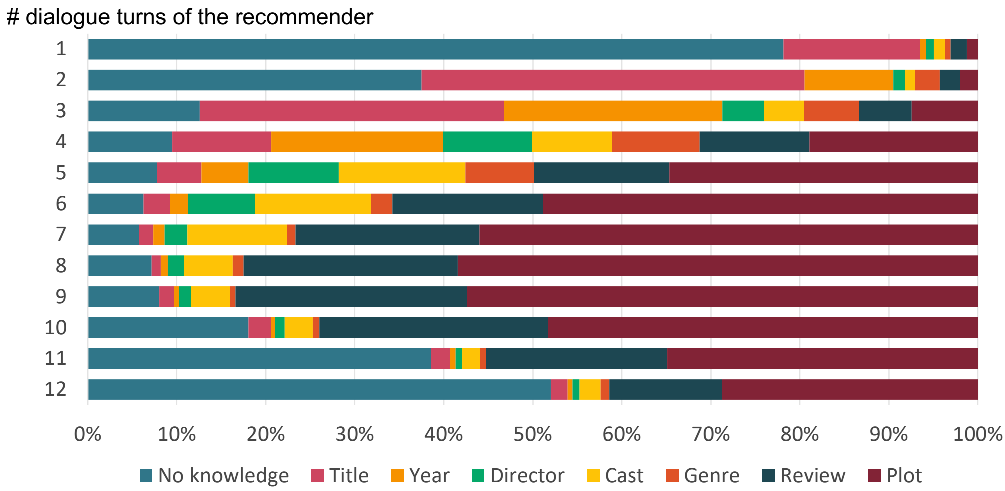


Figure 3.8: Distribution of external knowledge used in each dialogue turn of the recommender. The information up to turn 12 is shown here.

knowledge. Additionally, non-factoid texts such as reviews and plots tend to be used more frequently.

Furthermore, Figure 3.8 shows the distribution of the knowledge used in each dialogue turn of the recommender. In the early part of the dialogue, numerous utterances lack knowledge, such as greetings. Then, the recommenders often use factoid information such as title, released year, director, and cast. In the latter part of the dialogue, non-factoid information, such as reviews and plots, is often used to convey specific content. In addition, beyond ten turns, the percentage of "No knowledge" increases again, as more generic recommendations such as "please check it out" are used. This analysis demonstrates that our dataset is capable of analyzing human recommendation strategies.

Post-task Questionnaire

We ask the dialogue participants to answer the following post-task questionnaire in some of the collected dialogues (= 4,410 dialogues).

Q1: Do you like movies?

Q2: Did you enjoy the dialogue?

	Q1	Q2	Q3	Q4	Q5
Recommender	4.36	4.00	3.94	4.01	-
Seeker	4.26	3.83	2.72	-	3.82

Table 3.2: Results of the questionnaire.

Q3: Do you know the movie you recommended (or that was recommended to you)?

Q4: Do you think you have recommended the movie well?

Q5: Do you want to watch the recommended movie?

All questions are answered on a 5-point Likert scale, with five being the best and one being the worst. The choices for Q1, Q2, Q4, and Q5 are [agree/somewhat agree/neutral/somewhat disagree/disagree]. The choices for Q3 are [have seen the movie and remember the contents well/have seen the movie and remember some contents/have never seen the movie but know the plot/have never seen the movie but know only the title/do not know at all]. Q4 is for recommenders only, and Q5 is for seekers only.

Table 3.2 shows the results of the questionnaire. We found that most of the workers were highly interested in the topic of movies (Q1), and both recommenders and seekers enjoyed the dialogue, although it was relatively long, more than 20 turns (Q2). In addition, from Q3, we can see that the recommenders recommended movies they knew, whereas the seekers were often recommended movies they did not know. Finally, from Q4 and Q5, it was confirmed that the collected dialogues sufficiently achieved the purpose of movie recommendation.

3.5 Proposed Model

Our JMRD is characterized by its structured external knowledge and the utilization of multiple pieces of external knowledge in relatively long dialogues. Thus, leveraging these features, we propose a model that selects external knowledge and generates responses accordingly.

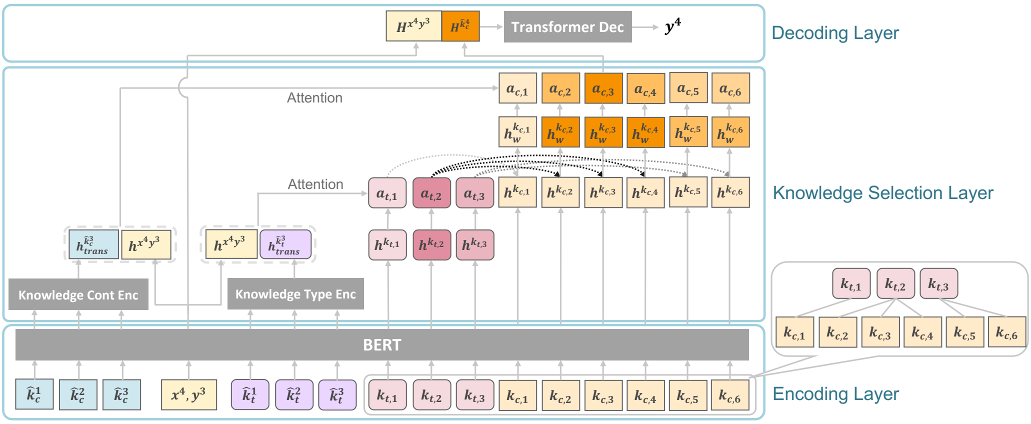


Figure 3.9: Overview of the proposed model. In this figure, the model generates the response y^4 at time $l = 4$. Knowledge Cont Enc, Knowledge Type Enc, and Transformer Dec denote the knowledge content encoder, the knowledge type encoder, and the transformer decoder, respectively.

3.5.1 Outline

Each dialogue $\mathcal{D} = \{(x^l, y^l)\}_{l=1}^{L}$ in the dataset is paired with a knowledge pool $\mathcal{K} = (\mathbf{k_t}, \mathbf{k_c})$ about the movie recommended in that dialogue, where x^l and y^l are the utterances of the seeker and recommender at turn l, and L is the number of turns in $\mathcal{D}$. In addition, $\mathbf{k_t}$ $(= \{k_{t,1}, \ldots, k_{t,m}, \ldots, k_{t,M}\})$ are the knowledge types, $\mathbf{k_c}$ $(= \{k_{c,1}, \ldots, k_{c,n}, \ldots, k_{c,N}\})$ are knowledge contents, and M, N are the number of knowledge types and knowledge contents contained in $\mathcal{K}$, respectively. At turn l, given the dialogue context (= the current seeker's utterance x^l and the last recommender's utterance y^{l-1}), the previously selected knowledge types $\{\hat{k}_t^1, \ldots, \hat{k}_t^{l-1}\}$, and previously selected knowledge contents $\{\hat{k}_c^1, \ldots, \hat{k}_c^{l-1}\}$, our target is to select a piece of knowledge $\hat{k}_c^l$ from $\mathbf{k_c}$ and generate response y^l utilizing $\hat{k}_c^l$. We call the previously selected knowledge types the "knowledge type history" and the previously selected knowledge contents the "knowledge content history" in this study.

Figure 3.9 shows the overview of the proposed model. The proposed model mainly consists of the Encoding Layer, the Knowledge Selection Layer, and the Decoding Layer. We describe each of the components in the following sections.

3.5.2 Encoding Layer

The encoding layer is used to obtain the following representations: dialogue context, knowledge types, knowledge contents, knowledge type history, and knowledge content history. We use BERT (Devlin et al., 2019) as the encoder. For encoding the dialogue context, we obtain the hidden state $H^{x^l y^{l-1}}$ via BERT, and then perform average pooling to obtain $h^{x^l y^{l-1}}$ (Cer et al., 2018):

$$H^{x^l y^{l-1}} = BERT(x^l, y^{l-1}) \tag{3.1}$$

$$h^{x^l y^{l-1}} = avgpool(H^{x^l y^{l-1}}) \in \mathbb{R}^d, \tag{3.2}$$

where d is the hidden size. We insert [SEP] between x^l and y^{l-1}, and insert [CLS] and [SEP] at the beginning and the end of the entire input string, respectively.

In the case of knowledge types, we insert [CLS] and [SEP] at the beginning and the end of the input string, respectively. After that, we get $\{h^{k_{t,m}}\}_{m=1}^{M}$ by feeding it to BERT in the same way. For the knowledge contents, we input the knowledge type in addition to the knowledge contents, following the method of Dinan et al. (2019). We insert a new special token [KNOW SEP] between the knowledge type and the knowledge content and further insert [CLS] and [SEP] at the beginning and the end of the input string, respectively. The resulting string is input to BERT to obtain $\{h^{k_{c,n}}\}_{n=1}^{N}$ likewise. We also compute the representations of knowledge type history $\{h^{\hat{k}_t^i}\}_{i=1}^{l-1}$ and that of knowledge content history $\{h^{\hat{k}_c^i}\}_{i=1}^{l-1}$.

3.5.3 Knowledge Selection Layer

We encode the knowledge type history via the transformer encoder (Vaswani et al., 2017). This transformer encoder (we call this "knowledge type encoder") adds a positional embedding for each turn (= turn embedding) to the input so that the model reflects in which turn each knowledge type was used (Meng et al., 2021). We concatenate the last output of this encoder $h_{trans}^{\hat{k}_t^{l-1}}$ with the hidden state of the dialogue context $h^{x^l y^{l-1}}$ as the query, and regard $\{h^{k_{t,m}}\}_{m=1}^{M}$ as the key. The attention over knowledge types $a_t \in \mathbb{R}^M$ is calculated as follows:

$$
\begin{aligned}
a_t &= [a_{t,1}, \ldots, a_{t,m}, \ldots, a_{t,M}] \\
&= \mathit{softmax}(Q_t K_t^{\top}) \\
Q_t &= \mathit{MLP}([h_{trans}^{\hat{k}_t^{l-1}}; h^{x^l y^{l-1}}]) \\
K_t &= \mathit{MLP}([h^{k_{t,1}}, \ldots, h^{k_{t,M}}]) \\
[h_{trans}^{\hat{k}_t^1}, \ldots, h_{trans}^{\hat{k}_t^{l-1}}] &= \mathit{KTE}([h^{\hat{k}_t^1}, \ldots, h^{\hat{k}_t^{l-1}}]),
\end{aligned}
\tag{3.3}
$$

where $MLP(\cdot)$ is a multilayer perceptron, KTE is the knowledge type encoder, and $[\cdot;\cdot]$ is the vector concatenation operation.

We compute the weighted hidden state of the knowledge contents $\{h_w^{k_{c,n}}\}_{n=1}^N$ based on the calculated attention a_t. This weighted hidden state is used to calculate the attention over the knowledge contents. Suppose the number of knowledge contents belonging to the m-th knowledge type is N_m, and the same weight $a_{t,m} \in a_t$ is given to all of them. In that case, the M-dimensional a_t can be extended to the N-dimensional $a'_t \in \mathbb{R}^N$ as follows, because N_m satisfies $\sum_{m=1}^M N_m = N$:

$$
a'_t = [a_{t,1}, \ldots, \underbrace{a_{t,m}, \ldots, a_{t,m}}_{N_m}, \ldots, a_{t,M}] \tag{3.4}
$$

Using a'_t, the weighted hidden states of the knowledge contents $\{h_w^{k_{c,n}}\}_{n=1}^N$ can be obtained as follows:

$$
[h_w^{k_{c,1}}, \ldots, h_w^{k_{c,N}}] = a'_t [h^{k_{c,1}}, \ldots, h^{k_{c,N}}] \tag{3.5}
$$

The knowledge content history is encoded by the transformer encoder as well. This transformer encoder, which we call "knowledge content encoder", has the same setting as the knowledge type encoder, but they do not share any parameters. We concatenate the last output of the encoder $h_{trans}^{\hat{k}_c^{l-1}}$ with $h^{x^l y^{l-1}}$ as the query, and regard the weighted hidden states of knowledge contents $\{h_w^{k_{c,n}}\}_{n=1}^N$ as the key. We can calculate the attention over the knowledge contents $a_c \in \mathbb{R}^N$ as follows:

$$\begin{aligned}a_c &= softmax(Q_c {K_c}^\top) \\ Q_c &= MLP([h^{\hat{k}_c^{l-1}}_{trans}; h^{x^l y^{l-1}}]) \\ K_c &= MLP([h_w^{k_{c,1}}, \ldots, h_w^{k_{c,N}}]) \\ [h^{\hat{k}_c^1}_{trans}, \ldots, h^{\hat{k}_c^{l-1}}_{trans}] &= KCE([h^{\hat{k}_c^1}, \ldots, h^{\hat{k}_c^{l-1}}]),\end{aligned} \tag{3.6}$$

where KCE is the knowledge content encoder. Finally, we select a knowledge content $\hat{k}_c^l$ at time l from the probability distribution of a_c.

3.5.4 Decoding Layer

At time l, the dialogue context x^l, y^{l-1} and the knowledge content $\hat{k}_c^l$ selected by the knowledge selection layer, are input to the transformer decoder to generate the response y^l. Specifically, we feed the concatenated embedding $H^{x^l y^{l-1} \hat{k}_c^l} = [H^{x^l y^{l-1}}; H^{\hat{k}_c^l}]$ to the decoder. The word generation probability $p(y_j^l)$ over the vocabulary V when the decoder generates the j-th word can be written as follows:

$$\begin{aligned}p(y_j^l) &= softmax(MLP(h_{dec}^{l,j})) \in \mathbb{R}^{1\times|V|} \\ h_{dec}^{l,j} &= TD(H^{x^l y^{l-1} \hat{k}_c^l}, emb(y_{<j}^l))) \in \mathbb{R}^{1\times d},\end{aligned} \tag{3.7}$$

where TD is the transformer decoder, $y_{<j}^l$ are the words generated up to the j-th word, $emb(y_{<j}^l)$ are the word embeddings of $y_{<j}^l$, which is initialized with the word embedding of BERT.

We use copy mechanism (Gu et al., 2016; See et al., 2017) to make it easier to generate knowledge words and follow the method used in Meng et al. (2021).

3.5.5 Learning Objective

Similar to Dinan et al. (2019), we combine the negative log-likelihood loss for the generated response $\mathcal{L}_{nll}$ with the cross-entropy loss for knowledge selection $\mathcal{L}_{knowledge}$ modulated by a weight λ, which is the hyperparameter. The final loss function $\mathcal{L}$ is as:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{nll} + \lambda\mathcal{L}_{knowledge} \tag{3.8}$$

3.6 Experiments

3.6.1 Settings

We randomly split the dialogues into the train (90%), validation (5%), and test sets (5%). Input texts are truncated to the maximum input length of 64 tokens for dialogue contexts and knowledge contents, and 5 tokens for knowledge types. In addition, a maximum of 20 turns of knowledge history can be entered for both knowledge types and knowledge contents. Our proposed dataset may have multiple pieces of knowledge associated with a recommender's utterance, but we use only one of them in this study for simplicity. For utterances with multiple knowledge items, we select the one with the highest Jaccard coefficient between the word sets of the recommender's utterance and the knowledge as the correct one. To input "No knowledge," we use the special token [NO KNOW] in place of knowledge type and content.

3.6.2 Baseline

We use an end-to-end Transformer Memory Network (TMN) (Dinan et al., 2019) as a baseline. This model encodes the dialogue context and each knowledge respectively and selects knowledge by calculating the dot-product attention between them. It also performs end-to-end response generation using the selected knowledge. To make a fair comparison with our proposed model, we have replaced the original transformer encoder with a BERT encoder. We call this model TMN BERT.

As a baseline to consider knowledge history, we add the knowledge content encoder to TMN BERT and concatenate its output with the hidden states of the dialogue context. We call this model TMN BERT+KH. Knowledge selection is made by calculating the attention between the knowledge candidates and the concatenated hidden states. Other conditions are the same as in TMN BERT.

In addition, we use Random baseline that selects knowledge randomly.

3.6.3 Implementation Details

We use the NICT BERT Japanese pre-trained model (with BPE)[5] as the encoder. This BERT model is also used to initialize the word embedding in the transformer decoder. The transformer encoders for knowledge type and knowledge content, and the transformer decoder have the same architecture, consisting of 2 attention heads, 5 layers, and the size of the hidden layer is 768, and the filter size is 3072. We train the models for 100 epochs, with a batch size of 512, and 0.1 gradient clipping. We do early stopping if no improvement of the validation loss is observed for five consecutive epochs. All models are learned with Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an initial learning rate = 0.00005. We use an inverse square root learning rate scheduler with the first 1,000 steps allocated for warmup. In addition, we set the hyperparameter λ to 0.95. For decoding, we use beam search with a beam of size 3. We add a restriction to prevent the same bigram from being generated multiple times.

3.6.4 Evaluation Metrics

We evaluate the models with automatic evaluation metrics. For knowledge selection, we use accuracy (**Acc**). For response reproducibility, we measure **BLEU**$_{tgt}$**-4** (Papineni et al., 2002), which is the 4-gram overlap between a generated response and a target response. We also use unigram F1 (**F1**) following the evaluation setting in Dinan et al. (2019). Additionally, we use **Jaccard** and **BLEU**$_{know}$**-4** to evaluate whether the knowledge is reflected in the generated response. **Jaccard** is the Jaccard coefficient of the set of words in the generated response and the set of words in the selected knowledge content. **BLEU**$_{know}$**-4** is the BLEU-4 computed between the generated response and the selected knowledge content.

3.6.5 Results

Table 3.3 shows the results of knowledge selection. TMN BERT+KH, which incorporates a mechanism for considering knowledge history into the baseline TMN BERT, demonstrated improved Acc, suggesting the importance of knowledge his-

[5] `https://alaginrc.nict.go.jp/nict-bert/index.html`

Model	Acc
Random	3.90 ($\pm$0.67)
TMN BERT	47.53 ($\pm$1.26)
TMN BERT+KH	48.19 ($\pm$0.92)
Ours	**48.87** ($\pm$1.68)

Table 3.3: The evaluation results for knowledge selection. Scores are the average of eight runs of the experiment with different random seeds, and 95% confidence intervals are shown in parentheses. The bold scores indicate the best results across all models.

	response reproducibility		knowledge reflection	
	F1	BLEU$_{tgt}$-4	Jaccard	BLEU$_{know}$-4
Random	23.40 ($\pm$0.76)	3.95 ($\pm$0.53)	5.90 ($\pm$0.23)	0.49 ($\pm$0.13)
TMN BERT	42.98 ($\pm$0.17)	20.46 ($\pm$0.66)	39.38 ($\pm$1.39)	25.50 ($\pm$0.95)
TMN BERT+KH	42.92 ($\pm$0.33)	20.44 ($\pm$0.62)	39.49 ($\pm$1.20)	25.55 ($\pm$1.00)
Ours	**43.44** ($\pm$0.64)	**20.72** ($\pm$0.56)	**39.90** ($\pm$0.96)	**26.26** ($\pm$0.86)

Table 3.4: The evaluation results for response generation. For the notation, please refer to Table 3.3.

tory. Our proposed method, which incorporates a mechanism for considering knowledge structure, further improved Acc.

Table 3.4 shows the results of response generation. The proposed method outperformed other methods in both response reproducibility and knowledge reflection. We believe this improvement is due to more accurate knowledge selection based on knowledge history and structure.

3.6.6 Case Study

Table 3.5 shows an example of knowledge selection and response generation. TMN BERT, which does not consider the history of knowledge, selects the plot despite being at the beginning of the dialogue. Furthermore, the utterance generated by TMN BERT does not reflect the selected knowledge. In contrast, our pro-

	Dialogue	Knowledge
R$_1$:	Nice to meet you.	no knowledge
S$_1$:	Hello.	-
R$_2$:	I am pleased to meet you.	no knowledge
S$_2$:	What movies do you recommend?	-
TMN BERT	I will introduce a movie called Do You Like Disney Movies?	Danny Ocean immediately breaks his parole rules (no interstate movement) and reunites with his partner Rusty Ryan in Los Angeles. He confides in Ryan about a new theft scheme he had hatched while in prison. (Plot)
Ours:	Today I will introduce Ocean's Eleven.	Ocean's Eleven (Title)
Gold:	How about Ocean's Eleven?	Ocean's Eleven (Title)

Table 3.5: Examples of generated responses by our model and the baseline model. R and S denote recommender and seeker, respectively. Subscript numbers indicate the number of turns in the dialogue. The knowledge type is indicated in parentheses in the Knowledge column.

posed model introduces the movie title that has not been mentioned earlier in this dialogue by considering the history of knowledge.

As illustrated by the generated response of TMN BERT, the generated utterances may not reflect the selected knowledge or may contain words inconsistent with the selected knowledge. This issue is referred to as the hallucination problem (Roller et al., 2021; Shuster et al., 2021), and we leave the solution to this problem as future work.

3.7 Summary of This Chapter

We proposed JMRD, a hierarchically structured knowledge-grounded movie recommendation dialogue dataset. We also proposed an end-to-end dialogue system that utilizes hierarchically structured knowledge to select knowledge and generate responses as a strong baseline for our dataset. The experimental results show that our model can select more appropriate knowledge than baselines.

As far as we know, this is the first Japanese dialogue dataset associated with external knowledge. We hope our dataset facilitates further research on movie recommendation dialogue based on structured external knowledge (especially in Japanese dialogue research).

In response generation, we can observe that the utterances do not reflect the knowledge in some cases, even when the knowledge is selected correctly. There is still much room for improvement in knowledge reflection, and we leave this as future work.

Chapter 4

Engagingness Analysis of External Knowledge-Based Responses

4.1 Introduction

More and more dialogue research has utilized external knowledge to enable dialogue systems to generate rich and informative responses (Ghazvininejad et al., 2018; Zhou et al., 2018; Moghe et al., 2018; Dinan et al., 2019; Zhao et al., 2020). The major focus of such research is on how to select appropriate external knowledge and accurately reflect it in the response (Kim et al., 2020; Zhan et al., 2021; Rashkin et al., 2021; Li et al., 2022).

However, as shown in Figure 4.1, a good speaker not only informs the dialogue partner of external knowledge but also incorporates his or her own knowledge, experiences, and opinions effectively, which makes the dialogue more engaging. The extent to which models specializing in reflecting given external knowledge can achieve such an engaging behavior has not yet been explored quantitatively.

In this study, we first analyze how humans incorporate speaker-derived information by annotating the utterances in an existing knowledge-grounded dialogue corpus. Each entity in the utterances is annotated with its information source,

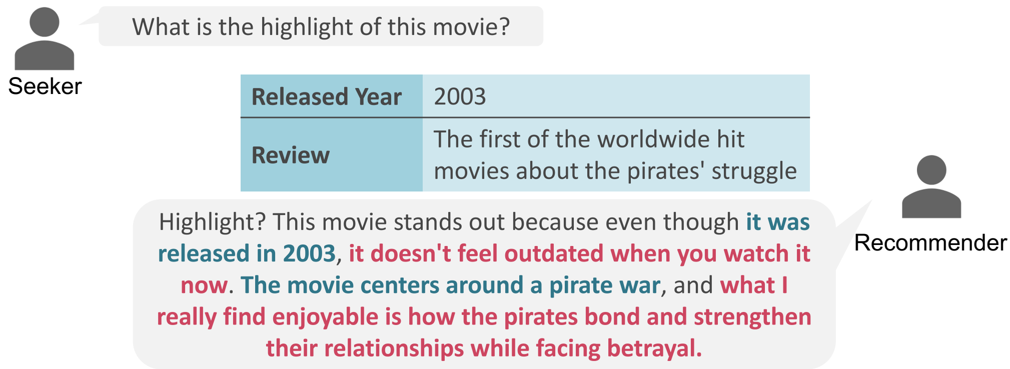


Figure 4.1: An example of Japanese Movie Recommendation Dialogue (Kodama et al., 2022). The table above the recommender's utterance indicates the external knowledge used in that utterance. The recommender incorporates not only database-derived information but also speaker-derived information.

either derived from external knowledge (database-derived) or the speaker's own knowledge, experiences, and opinions (speaker-derived). The analysis of the annotated dataset showed that engaging utterances contained more speaker-derived information.

In addition, we train a BART-based response generation model in a standard way, i.e., by minimizing perplexity, and investigate the extent to which it incorporates speaker-derived information. The experimental result showed that the response generation model did not incorporate speaker-derived information into their utterances as often as humans do. This result implies that minimizing perplexity is insufficient to increase engagingness in knowledge-grounded response generation and suggests room for improvement in the training framework.

4.2 Information Source Annotation

This section describes the annotation scheme for information sources and the annotation results.

4.2.1 Scheme

We annotate Japanese Movie Recommendation Dialogue (JMRD) (Kodama et al., 2022) with information sources. JMRD is a human-to-human knowledge-grounded dialogue corpus in Japanese. A recommender recommends a movie to a seeker. Each utterance of the recommender is associated with movie information as external knowledge. Each piece of knowledge consists of a knowledge type (e.g., movie title) and the corresponding knowledge contents (e.g., "Marvel's The Avengers").

In this study, we extract entities from the recommender's utterances and annotate them with their information source. Entities are nouns, verbs, and adjectives and are extracted together with their modifiers to make it easier to grasp their meanings. Entities are extracted using Juman++ (Morita et al., 2015; Tolmachev et al., 2018), a widely-used Japanese morphological analyzer. Annotators classify the extracted entities into the following information source types:

Database-derived: The entity is based on the external knowledge used in that utterance.

Speaker-derived: The entity is based on the knowledge, experiences, and opinions that the recommender originally has about the recommended movie.

Other: The entity does not fall under the above two types (e.g., greetings).

Example (1) shows an annotation example. The recommender's utterance contains two entities, action scenes and spectacular. The former is classified as database-derived because the recommender uses the external knowledge {*Genre, Action*} in that utterance. Conversely, the latter entity is categorized as speaker-derived because the information *spectacular* is not based on the external knowledge used.

(1) Utterance: The action scenes$_{\text{(database)}}$ are spectacular$_{\text{(speaker)}}$!
External Knowledge: {Genre, Action}

We recruited professional annotators, who are native Japanese speakers, to annotate these information source types. One annotator was assigned to each dialogue.

	Train	Dev	Test	Total
# dialogues	4,575	200	300	5,075
# recommender's utterances	51,080	2,244	3,347	56,671
# entities	235,771	10,320	15,734	261,825
# database-derived	166,958	7,223	10,476	184,657
# speaker-derived	51,170	2,303	4,095	57,568
# other	17,643	794	1,163	19,600

Table 4.1: Statistics of the information source annotation.

After the annotation, another annotator double-checked the annotated results. The annotation of all dialogues cost approximately 1.4 million JPY.

4.2.2 Result

Table 4.1 shows the annotation statistics. While JMRD is a knowledge-grounded dialogue corpus and thus inherently contains many database-derived entities, it also contains about 60,000 speaker-derived entities. This result verifies that humans incorporate their own knowledge, experiences, and opinions into their utterances, even in dialogues to convey external knowledge.

4.3 Analysis of Human Utterances

Based on the information source annotations, we first analyze human utterances at the dialogue level and utterance level.

4.3.1 Dialogue-level Analysis

4,328 dialogues in JMRD have post-task questionnaires on 5-point Likert scale (5 is the best.) We consider the seekers' answers to the question (i.e., Did you enjoy the dialogue?) as a measure of dialogue engagingness and analyze the relationship between this and the ratio of each information source label.

Figure 4.2 shows that dialogues with high engagingness scores tend to have more speaker-derived entities (or less database-derived) than those with low en-

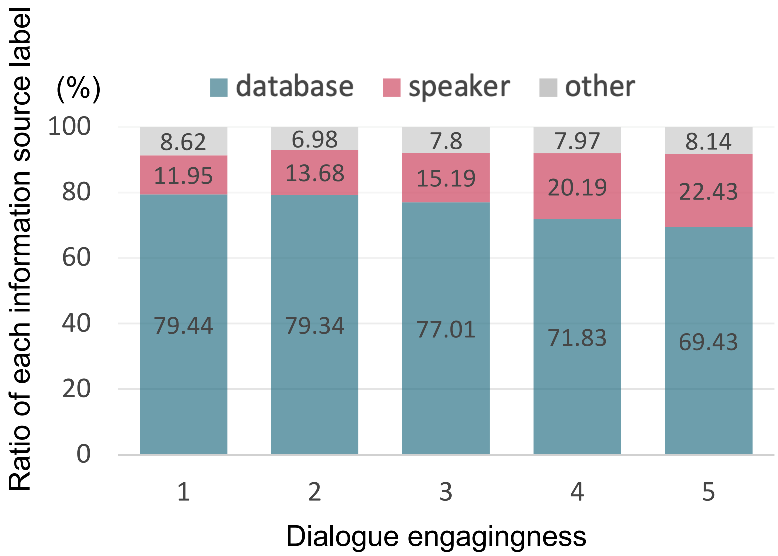


Figure 4.2: Relationship between dialogue engagingness and ratio of each information source label.

gagingness scores. When constructing JMRD, recommenders were given a certain amount of external knowledge and asked to use that knowledge to respond. However, recommenders highly rated by their dialogue partners incorporated not only the given external knowledge but also speaker-derived information to some extent in their dialogues.

4.3.2 Utterance-level Analysis

We conduct the utterance-level evaluation via crowdsourcing. We randomly extract 500 responses along with their dialogue contexts (= 4 previous utterances) from the test set. For each utterance, workers rate utterance engagingness (i.e., Would you like to talk to the person who made this response?) on a 5-point Likert scale, with 5 being the best. Three workers evaluate each utterance, and the scores are averaged.

The average score for utterances with speaker-derived entities was 3.31, while those without speaker-derived entities was 3.07. Student's t-test with p = 0.05 revealed a statistically significant difference between these scores.

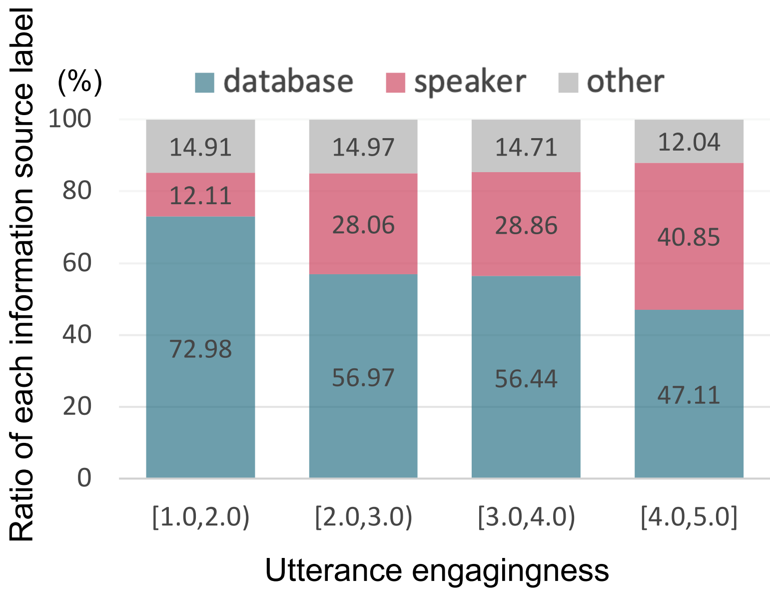


Figure 4.3: Relationship between utterance engagingness and ratio of each information source label.

Furthermore, Figure 4.3 shows the relationship between utterance engagingness and the ratio of each information source label. This figure shows that utterances with high scores tend to have more speaker-derived entities. This trend is consistent with that of the dialogue engagingness. The above analysis demonstrates that information obtained from the speaker's own experience is an important factor in utterance engagingness.

Does subjective knowledge contribute to engagingness? The knowledge type used in JMRD can be divided into subjective knowledge (review) and objective knowledge (title, etc.). Reviews are the opinions of individuals who have watched movies and have similar characteristics to speaker-derived information. We then examine whether there is a difference in engagingness between utterances using subjective and objective knowledge. The average engagingness scores were 3.32 and 3.16[1], respectively, and Student's t-test with p = 0.05 revealed no statistically significant difference.

4.4 Analysis of System Utterances

We investigate the distribution of information source labels in the responses of the model trained on the knowledge-grounded dialogue dataset. First, we train a Response Generator (Section 4.4.1) with the dialogue contexts and external knowledge as input and responses as output. Next, an Information Source Classifier (Section 4.4.2) is trained with responses and external knowledge as input and information source labels as output. Then, the Information Source Classifier infers the information source labels for the system responses generated by the Response Generator. Finally, we analyze the distribution of inferred information source labels.

4.4.1 Response Generator

We use a BART$_{large}$ (Lewis et al., 2020) model as a backbone.[2] The input to the model is formed as follows:

$$\begin{aligned}&[CLS]u_{t-4}[SEP]u_{t-3}[SEP]u_{t-2}[SEP]\\&\qquad u_{t-1}[SEP][CLS_K]kt^1[SEP]kc^1[SEP]...\\&\qquad\qquad [CLS_K]kt^M[SEP]kc^M[SEP],\end{aligned} \tag{4.1}$$

where t is the dialogue turn, u_t is the t-th response, and kt^i and kc^i ($1 <= i <= M$) are the knowledge type and knowledge content associated with the target response, respectively (M is the maximum number of knowledge associated with u_t.) $[CLS_K]$ is a special token. We feed the gold knowledge into the model to focus on how knowledge is reflected in the responses. The model learns to minimize perplexity in generating u_t.

[1] We exclude utterances referring to both of subjective and objective knowledge from this result.

[2] https://huggingface.co/ku-nlp/bart-large-japanese

Dialogue contexts, knowledge (knowledge types and contents), and target responses are truncated to the maximum input length of 256, 256, and 128, respectively. The model is trained for up to 50 epochs with a batch size of 512 and 0.5 gradient clipping. We apply early stopping if no improvement of the loss for the development set is observed for three consecutive epochs. We use AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ and an initial learning rate $= 1e-5$. We use an inverse square root learning rate scheduler with the first 1,000 steps allocated for warmup. During decoding, we use the beam search with a beam size of 3.

We evaluate the quality of response generation with the SacreBLEU (Post, 2018).[3] BLEU-1/2/3/4 scored high, 81.1/73.5/71.0/69.9. This result is reasonable because the gold knowledge was given.

4.4.2 Information Source Classifier

We fine-tune a RoBERTa$_{large}$ (Liu et al., 2019) model.[4] The Information Source Classifier performs a sequence labeling task to estimate BIO[5] labels of the information source. The input to the model is formed as follows:

$$[CLS]u_t[SEP][CLS_K]kt^1[SEP]kc^1[SEP]... \\ [CLS_K]kt^M[SEP]kc^M[SEP] \quad (4.2)$$

The model learns to minimize the cross-entropy loss between the model outputs and the correct BIO labels of the information source.

Target responses and knowledge (knowledge types and contents) are truncated to the maximum input length of 128 and 384, respectively. The model is trained for up to 20 epochs with a batch size of 64 and 0.5 gradient clipping. We apply early stopping if no improvement of the f1 score for the development set is observed for three consecutive epochs. We use AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ and an initial learning rate $= 1e-5$.

[3] `https://github.com/mjpost/sacrebleu`

[4] `https://huggingface.co/nlp-waseda/roberta-large-japanese-seq512`

[5] B, I, and O stand for Begin, Inside and Outside, respectively.

	Prec.	Rec.	F1
database-derived	94.92	95.61	95.27
speaker-derived	80.88	84.39	82.60
other	82.93	64.15	72.34
micro avg.	90.52	90.48	90.50

Table 4.2: Results of the sequence labeling by Information Source Classifier.

Dist. (%)	Human (gold)	Human (pred)	System (pred)
database-derived	66.22	66.75	85.48
speaker-derived	26.33	27.49	10.66
other	7.45	5.77	3.86

Table 4.3: Distributions of information source labels for human and system responses. "gold" indicates the label annotated by humans, and "pred" indicates the label inferred by the Information Source Classifier.

We use an inverse square root learning rate scheduler with the first 1,000 steps allocated for warmup.

Table 4.2 shows precision, recall, and F1 scores for each label and micro average scores across all labels. The scores were calculated using seqeval.[6] The micro average F1 score was 90.50, which is accurate enough for the further analysis.

4.4.3 Analysis for Inferred Labels

The information source labels for system responses are inferred using the classifier trained in Section 4.4.2. Table 4.3 shows distributions of information source labels for human and system responses. For a fair comparison, the human responses are also given labels inferred by the classifier (denoted as **Human (pred)**), although they have gold labels (denoted as **Human (gold)**). **Human (gold)** and **Human (pred)** have similar distributions, indicating that the accuracy of the classifier is sufficiently high. For **System (pred)**, the percentage of database-derived labels

[6] `https://github.com/chakki-works/seqeval`

		Engagingness
Context	... Recommender: This movie is an animation movie released in 2015. Seeker: I see.	
Knowledge	{director, Takahiko Kyogoku}, {cast, Emi Nitta}, {cast, Yoshino Nanjo}	
Response	Human: The director is Takahiko Kyogoku, and the voice actors are Emi Nitta and Yoshino Nanjo. These two are also singers.	4.00
	System: The director is Takahiko Kyogoku. The voice actors are Emi Nitta and Yoshino Nanjo.	2.33

Table 4.4: An example of the human and system response. In the Knowledge row, the left side in the curly brackets indicates the knowledge type and the right side indicates the knowledge content. The blue and red parts refer to database-derived and speaker-derived information, respectively.

increased significantly (66.75%→85.48%) and that of speaker-derived information decreased significantly (27.49%→10.66%). This result shows that the response generation model, trained in a standard way, was not able to use speaker-derived information as often as humans do.

Table 4.4 shows an example of human and system responses along with the engagingness scores. The system was able to reflect given knowledge in the response appropriately but did not incorporate additional speaker-derived information, such as the information two voice actors also work as singers.

For further analysis, we investigate the average ratios of speaker-derived information by knowledge type used. Table 4.5 shows the result. Significant drops were observed for reviews (31.42%→6.32%) and plots (13.68%→2.32%). This is probably because reviews and plots are relatively long and informative external knowledge, so the system judged there was no need to incorporate additional speaker-derived information.

Combined with our observation that speaker-derived information improves

Ratio (%)	Human (gold)	Human (pred)	System (pred)
Title	30.21	34.12	27.09
Released Year	16.41	22.31	6.56
Director	13.94	11.96	4.50
Cast	36.11	45.34	23.45
Genre	10.47	15.14	5.49
Review	27.72	31.42	6.32
Plot	13.98	13.68	2.32
No knowledge	57.49	63.08	55.99

Table 4.5: Average ratios of speaker-derived labels per knowledge type used.

engagingness, the current model is likely to have lower engagingness due to its inability to effectively incorporate speaker-derived information. Such an ability is hardly learned by simply optimizing a model to reduce the perplexity of response generation, suggesting the need for a novel learning framework.

4.5 Summary of This Chapter

We analyzed the distribution of speaker-derived information in human and system responses in the knowledge-grounded dialogue. The analysis showed that the use of speaker-derived information, as well as external knowledge, made responses more engaging. We also confirmed that the response generation model trained in a standard way generated less speaker-derived information than humans.

It is difficult to make good use of speaker-derived information by simply minimizing the perplexity of the model because a wide variety of speaker-derived information appears in each dialogue. We hope our published annotated corpus becomes a good launch pad for tackling this issue.

Chapter 5

Response Generation Based on User Internal States and External Knowledge

5.1 Introduction

In human dialogues, individuals pay careful attention to their interlocutor's internal state (Chiba et al., 2014), including their level of understanding and emotional state. Particularly in recommendation dialogues, where a recommender recommends something to a seeker, it is crucial to estimate what the seeker knows and what he/she is interested in. This understanding enables us to offer recommendations that better align with the seeker's preferences.

In the past few years, many large language models (LLMs) have been actively developed and have achieved remarkable performance in various natural language processing tasks (Brown et al., 2020; Zhang et al., 2022; Chowdhery et al., 2022; OpenAI, 2023). Current LLMs are able to generate human-like responses without specialized modules to consider the interlocutors. However, it remains an open question whether LLMs need to explicitly consider the seeker's internal state and how to effectively implement it. To answer this question, we need high-quality dialogue data with careful and fine-grained annotations of the seeker's internal state. Unfortunately, there are no existing recommendation dialogue datasets with internal state annotation.

One possible solution is to annotate existing recommendation dialogue datasets (Li et al., 2018; Kang et al., 2019; Jia et al., 2022) with the seeker's internal state. However, the internal state labels annotated by a third party may not accurately reflect the actual state. Kajiwara et al. (2021) point out a difference between the subjective emotions the document writer intends to convey and the objective emotions the document reader receives. Thus, it is necessary to collect new dialogues and have the seekers annotate them with the subjective internal state.

To account for the aforementioned requirements, we constructed **RecMind**, a Japanese movie recommendation dialogue dataset. As illustrated in Table 5.1[1], the recommender recommends movies based on the seeker's preferences in a multi-turn dialogue. We treat noun phrases as entities and annotate each of them with the seeker's level of knowledge and interest at three levels (*High*, *Neutral*, and *Low*). In this process, the seeker assigns subjective labels to each entity, reflecting their own perception. Conversely, the recommender estimates the seeker's internal state based on the interactions with the seeker and assigns objective labels.

Our dataset also features high-quality dialogues with high dialogue enjoyment and recommendation success. Compared to JMRD (Kodama et al., 2022), an existing Japanese movie recommendation dialogue dataset, our dataset stands out with longer seeker's utterances. This characteristic allows us to observe the internal states of a wide variety of entities.

Using the constructed dataset, we analyze the relationship between the seeker's internal state and the recommendation success. Our analysis reveals that entities without knowledge but with interest contribute to the successful recommendation. This finding suggests that the recommender should focus on topics or subjects that the seeker lacks knowledge of yet is interested in.

Based on the analysis results, we also propose a LLM-based response generation framework that explicitly considers the seeker's internal state. Specifically, we apply Chain-of-Thought prompting (Wei et al., 2022) and estimate the seeker's

[1]Examples of dialogues presented in this study are originally in Japanese and were translated by the authors.

Dialogue	Entity	Knowledge	Interest
R1: What kind of movies do you usually watch?	-	-	-
S1: I often watch Japanese movies regardless of their genres.	Japanese movies	*H* / *H*	*H* / *H*
R2: OK, Japanese movies. Do you know an eccentric movie called "DESTINY: The Tale of Kamakura"?	Japanese movies	*H* / *H*	*H* / *H*
	eccentric movie	*H* / *L*	*H* / *N*
	"DESTINY: The Tale of Kamakura"	*N* / *L*	*H* / *N*
S2: I have never seen it. What kind of movie is it?	-	-	-
R3: Through strange experiences around an unsuccessful mystery writer living in Kamakura and his new wife, they gradually discover their destiny. That's the story. It's interesting to see specters and ghosts. Because this is Kamakura, he said.	Strange experiences around an unsuccessful mystery writer living in Kamakura and his new wife	*L* / *L*	*H* / *N*
	their destiny	*L* / *L*	*H* / *L*
	specters and ghosts	*L* / *L*	*H* / *L*
S3: Sounds interesting. Could I see any street scenery of Kamakura?	street scenery of Kamakura	*H* / *N*	*H* / *H*
R4: The scenery of Kamakura does not appear so much. The main characters go to the land of Hades by train, so you can see a fantasy-like scene.	The scenery of Kamakura	*H* / *N*	*H* / *H*
	the land of Hades	*N* / *L*	*H* / *L*
	train	*N* / *L*	*H* / *L*
	a fantasy-like scene	*H* / *L*	*H* / *L*
S4: I remembered that I had heard of this movie. Is Masato Sakai in it, I think?	-	-	-
R5: That's right. His new wife is played by Mitsuki Takahata.	His new wife	*H* / *L*	*H* / *N*
	Mitsuki Takahata	*H* / *H*	*H* / *N*
S5: I see. Both are my favorite actors, so I would love to see them!	-	-	-

Table 5.1: An example of RecMind. R and S denote the recommender and seeker, respectively. The Entity column lists the entities extracted from the dialogue. Each entity has subjective/objective labels for knowledge and interest. *H*, *N*, and *L* denote *High*, *Neutral*, and *Low*, respectively.

internal state before generating a response. The human evaluation results demonstrate that our proposed method outperforms the baseline method, which does not explicitly consider the seeker's internal state, in both consistency and the success of recommendations.

In summary, our contributions are as follows.

- We proposed RecMind, a Japanese movie recommendation dialogue dataset with subjective and objective annotations of the seeker's internal state at the entity level.
- We found entities the seeker has no knowledge about but has an interest in contributed to the successful recommendation.
- We proposed the response generation framework that explicitly considers the seeker's internal state, applying Chain-of-Thought prompting (Wei et al., 2022).

5.2 Related Work

Our work centers on the interlocutor's internal states in dialogues. Besides, we describe the prior work on recommendation dialogue datasets.

5.2.1 Internal State

We focus on knowledge and interest as internal states in recommendation dialogues. Here, we introduce previous studies that deal with knowledge and interest in dialogues.

Miyazaki et al. (2013) proposed a method to estimate callers' levels of knowledge about particular themes (e.g., troubleshooting of products and services) in call center dialogues. Their annotations are performed at the dialogue level, whereas our dataset is annotated at the entity level. This allows for more fine-grained knowledge-state tracking and analysis. Inspired by the theory of mind (Premack and Woodruff, 1978) and the common ground (Clark, 1996), Bara et al. (2021) created MINDCRAFT dataset which considers the user's knowledge for situated dialogue in collaborative tasks. Given the necessary knowledge and skills, two workers are asked to create a specific object together in the 3D virtual blocks world of Minecraft. Every predetermined time interval, players must answer a question about the common ground (e.g., "Do you think the other player knows how to make YELLOW_WOOL?"). In this study, we consider the user's knowledge in a more realistic dialogue that contains both chit-chat and recommendations.

Modeling interlocutor's interest has been actively studied in the field of recommendation dialogue (Kang et al., 2019; Liu et al., 2020; Zhou et al., 2020; Jia et al., 2022). In GoRecDialog (Kang et al., 2019), each worker is given a set of five movies. The seeker's set represents their watching history, and the recommender's set represents the candidate movies to choose from. The recommender should recommend the appropriate movie among the candidates to the seeker. DuRecDial (Liu et al., 2020) is a recommendation dialogue dataset containing multiple dialogue types, such as question-answering and chit-chat. The recommender attempts to elicit the seeker's preferences, and the seeker responds based on a predefined user profile. These studies focus on the preferences for predefined objects (e.g., movies, user profiles). Our dataset differs in that we annotate all entities appearing in dialogues with the seeker's interest.

5.2.2 Recommendation Dialogue Dataset

Many previous studies have released recommendation dialogue datasets (Li et al., 2018; Kang et al., 2019; Moon et al., 2019; Liu et al., 2020; Zhou et al., 2020; Kodama et al., 2022; Jia et al., 2022). INSPIRED (Hayati et al., 2020) is a movie recommendation dialogue dataset associated with the recommendation strategy label (e.g., preference confirmation, personal experience). Using this annotation, they analyzed the recommender's strategies and pointed out that using sociable strategies (e.g., sharing personal opinions) more frequently leads to successful recommendations. We provide the labels of the seeker's internal state (i.e., knowledge and interest) from both the recommender's and the seeker's points of view.

5.3 Data Collection

We collect data via crowdsourcing through a data supplier in Japan. In this section, we describe how we collect the RecMind dataset.

5.3.1 Dialogue Collection Settings

Workers

The two workers engaging in a dialogue have distinct roles: **recommender** and **seeker**. The recommenders recommend movies that align with the seeker's preferences, taking into account the seeker's current internal state. The seekers actively participate in the dialogue and ask questions if there is anything they do not know about what the recommender says.

It is assumed that recommendations from recommenders unfamiliar with movies might be short-sighted or less engaging because their knowledge about movies is sparse. Thus, we have two requirements for recommenders: (1) to be a movie enthusiast and (2) to watch an average of at least ten movies per year. We do not have any specific requirements for seekers.

Tasks for Workers

Workers are required to complete the four specific tasks: dialogue, annotation of the seeker's internal state, annotation of external knowledge[2], and questionnaire.

Dialogue During a dialogue, the recommender suggests one or more movies to the seeker. Recommenders must actively gather enough information from the seeker through dialogue. They should also be attentive to the seeker's preferences rather than suggesting movies based on their own tastes. Meanwhile, seekers are encouraged to openly share their preferences and ask questions about any unknowns. Each participant is required to respond at least eight times.

[2] In this study, knowledge means the seeker's internal state of knowledge, and external knowledge means the information the recommenders refer to in dialogues.

Annotation of Seeker's Internal State The seekers annotate each entity in the dialogues with the subjective labels of the level of knowledge and interest, while the recommenders annotate the entity with the objective labels.

The options for knowledge are as follows:

High The seeker has knowledge regarding the entity.

Neutral The entity cannot be said to be either *High* or *Low*. Or the level of knowledge for the entity cannot be judged from the given context.

Low The seeker does not have knowledge regarding the entity.

The options for interest are as follows:

High The seeker is interested in the entity.

Neutral The entity cannot be said to be either *High* or *Low*. Or the level of interest for the entity cannot be judged from the given context.

Low The seeker is not interested in the entity.

In addition to the above three options, we introduce an additional option, denoted as *Error*. This option is used when the extracted span is not a valid entity. We discard entities that the recommender or the seeker labeled as *Error*. The annotation can be performed either during or after the dialogue.

Annotation of External Knowledge Following the previous research on knowledge-grounded dialogues (Dinan et al., 2019; Wu et al., 2019), recommenders annotate their own utterances with the piece of external knowledge when they refer to it. The annotation is not required for utterances that do not refer to external knowledge, such as greetings and utterances containing the personal knowledge of the recommenders. However, the recommenders are required to always annotate their utterances with the title of the recommended movies when mentioning them.[3] This is to track recommended movies in the dialogues.

[3]For dialogues missing the annotation of the recommended movies, authors read the dialogues and annotated them with the movie titles.

	Question	Choice
Q1	How many movies do you watch per year?	5: 20 or more, 4: 10 to 19, 3: 5 to 9, 2: 3 to 4, 1: 2 or less
Q2	Do you know the movie you recommended? (for R) Do you know the movie that was recommended? (for S)	5: have watched the movie and remembered the contents well 4: have watched the movie and remembered some of the contents 3: have never watched the movie but know the plots 2: have never watched the movie and know only the title 1: do not know at all
Q3	Did you enjoy the dialogue?	5: agree, 4: somewhat agree, 3: neutral, 2: somewhat disagree, 1: disagree
Q4	Do you think you have recommended the movie well? (for R) Do you want to watch the recommended movie? (for S)	5: agree, 4: somewhat agree, 3: neutral, 2: somewhat disagree, 1: disagree

Table 5.2: Questions and choices of the questionnaire. R and S denote recommender and seeker, respectively. The number at the beginning of each choice indicates the score for that choice.

Questionnaire After the dialogue, workers answer the questionnaire shown in Table 5.2. We assign a score of 5 to 1 to each choice for each question.

5.3.2 Dialogue Collection System

We develop a web-based dialogue collection system for data collection. This system is an extension of ChatCollectionFramework[4], by adding a movie search tool and an internal state annotation tool. Figures 5.1 and 5.2 show the screenshots of the recommender's and the seeker's chatrooms, respectively.

[4] `https://github.com/ku-nlp/ChatCollectionFramework`

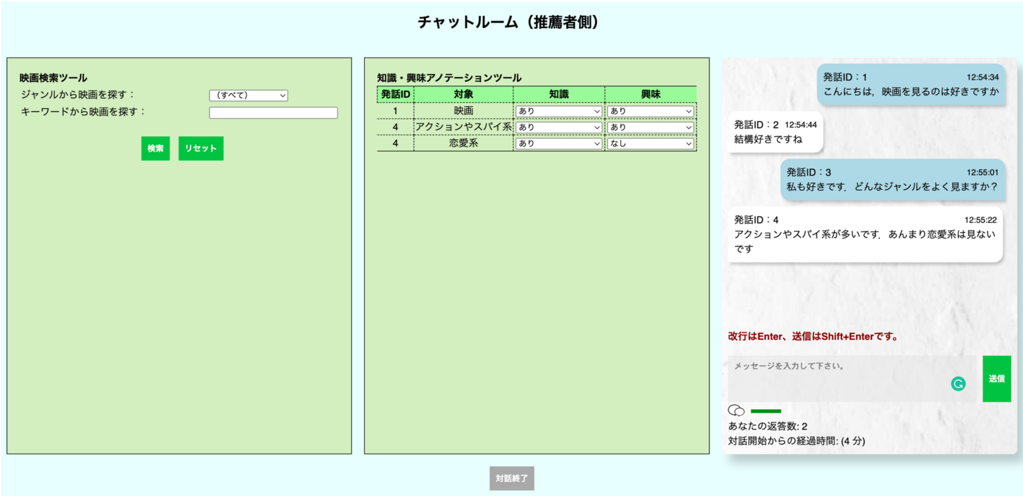


Figure 5.1: Screenshot of the recommender's chatroom

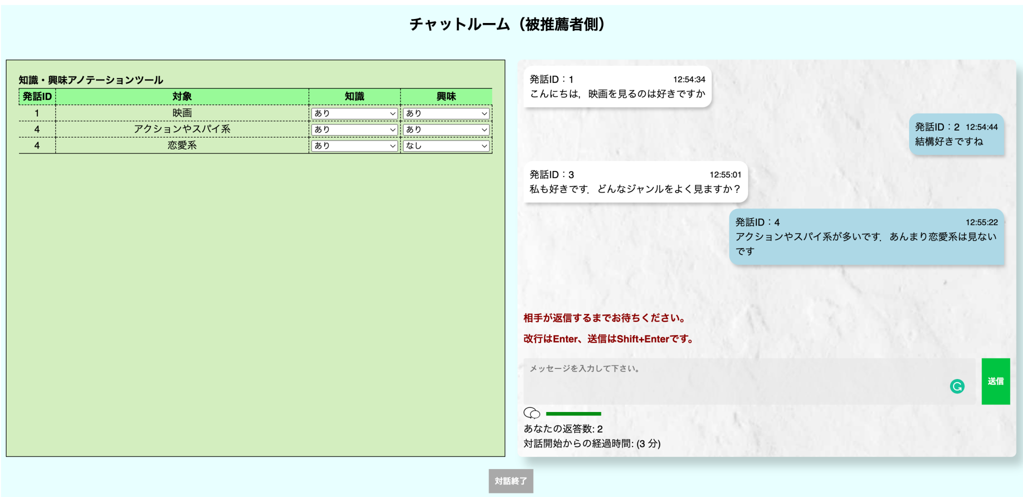


Figure 5.2: Screenshot of the seeker's chatroom

Movie Search Tool

We create a movie search tool to assist recommenders in dialogues. We first curate 2,317 popular movie titles and their genres from a Japanese movie information website, Yahoo! Movies.[5] We then collect metadata for each movie from

[5] https://movies.yahoo.co.jp/

Wikipedia. Metadata consists of the title, release date, running time, directors, casts, original work, theme song, production country, box office, and plots.[6] Additionally, we include reviews for 261 movies from JMRD (Kodama et al., 2022) as part of the metadata.

During dialogue collection, recommenders use this tool to search and check movie information, as shown in the left part of Figure 5.1. Searching can be done by genres or text-based queries. We save the search log with the corresponding recommender's utterance as one of the records of the recommender's behaviors. When sending an utterance, recommenders can annotate it with the referred external knowledge by clicking the checkbox on the side of each piece of external knowledge. This tool is only displayed on the recommender's screen, that is, the seekers cannot see movie information.

Internal State Annotation Tool

The internal state annotation tool displays the entities to be annotated on the screen of both the recommenders and the seekers, as shown in the center part of Figure 5.1 and the left part of Figure 5.2. Entities are automatically extracted from utterances to reduce the load of workers. We consider noun phrases as entities. Modifiers are extracted together to make it easier to grasp their meanings. We use linguistic features from the Japanese morphological analyzer Juman++ (Morita et al., 2015; Tolmachev et al., 2018) and the Japanese syntactic analyzer KNP (Kurohashi and Nagao, 1994) for entity extraction.

5.3.3 Statistics

Dialogue and Questionnaire

Table 5.3 shows the statistics of RecMind. We collected 1,201 dialogues consisting of an average of 17.5 utterances. 739 movies were used in the dialogues, indicating our dataset contains diverse recommendation dialogues.

The bottom row in Table 5.4 shows the questionnaire results. The results of Q2 show that the recommenders often recommend movies that the seeker does

[6] Some metadata may be missing.

# dialogues	1,201
# utterances (R / S)	10,697 / 10,317
Avg. # utterances per dialogue	17.5
# movies	739
# workers (R / S)	27 / 46
# searches	5,596
# external knowledge	5,250
# entities (knowledge / interest)	52,586 / 52,246

Table 5.3: Statistics of RecMind. R and S denote recommender and seeker, respectively.

not know.

Comparison with JMRD Table 5.4 also shows the comparison results with JMRD (Kodama et al., 2022), a knowledge-grounded recommendation dialogue in the same language and domain. The result of Q3 shows that the recommendation process is more enjoyable for both recommenders and seekers in our dataset. Regarding Q4, the result shows that our recommendations are more successful. Notably, the average score of Q4 by seekers improved from 3.82 to 4.51, highlighting that our dialogues are high-quality recommendation dialogue.

We next compare the average number of words per utterance. The results demonstrate that our dataset has longer utterances than JMRD. Especially, the seeker's utterances of RecMind are more than four times longer than those of JMRD, which could facilitate the analysis of the seeker's internal state. We additionally compare the average count of external knowledge use per recommender's utterance and observe a decrease from 1.24 to 0.75 in our dataset. This decrease is because we did not mandate recommenders to use external knowledge, except when mentioning movie titles. We believe that it is unnecessary to link external knowledge to every utterance because humans only refer to external knowledge when necessary.

Influence of Recommender's Movie Knowledge As noted in Section 5.3.1, we recruited movie enthusiasts who watched ten movies or more per year as recom-

	Q1		Q2		Q3 (↑)		Q4 (↑)		Words (↑)		Ext. K. (↓)
	R	S	R	S	R	S	R	S	R	S	-
JMRD	-	-	3.94	2.72	4.00	3.83	4.01	3.82	23.80	6.87	1.24
RecMind^{-}	2.57	3.66	3.80	1.58	3.99	4.27	3.61	4.47	**41.90**	**31.48**	0.75
RecMind	4.73	3.54	3.17	1.79	**4.29**	**4.42**	**4.27**	**4.51**	41.07	31.08	**0.49**

Table 5.4: Results of the questionnaire and the comparison with JMRD. "Words" indicates the average number of words per utterance and "Ext. K." indicates the average use count of external knowledge per recommender's utterance. R and S denote recommender and seeker, respectively. "$^{-}$" means the results of the dialogue collection by the recommenders who are not movie enthusiasts. Best results are in bold. The scores for Q1 and Q2 are not bolded because a higher (or lower) score does not imply superiority of any kind.

menders. To verify the effectiveness of this recruitment, we collected 74 dialogues from recommenders who watched fewer than ten movies per year. This data collection followed the same methodology as described in Section 5.3.1, except for the number of movies the recommenders watched.

Table 5.4 shows the results. The average score of Q3 by seekers decreased from 4.42 to 4.27, and that of Q4 from 4.51 to 4.47. Furthermore, the scores for Q3 and Q4 by recommenders, which means self-evaluation, also decreased from 4.29 to 3.99 and from 4.27 to 3.61, respectively. These results indicate that movie enthusiasts are likely to deliver more enjoyable dialogues and recommend successfully.

While the length of utterances is comparable, the number of external knowledge used increases from 0.49 to 0.75. This is because the recommenders who are not movie enthusiasts tend to rely on external knowledge more frequently to compensate for their lack of knowledge about movies.

Internal State

RecMind has 52,586 and 52,246 entities annotated with the seeker's knowledge and interest, respectively. Tables 5.5 and 5.6 show the statistics of the seeker's

Obj. \ Sub.	*High*	*Neutral*	*Low*	Total
High	20,664	3,084	4,794	28,542
Neutral	6,737	1,791	3,583	12,111
Low	5,154	1,502	5,277	11,933
Total	32,555	6,377	13,654	-

Table 5.5: Statistics of knowledge annotation.

Obj. \ Sub.	*High*	*Neutral*	*Low*	Total
High	28,244	4,338	746	33,328
Neutral	11,838	3,716	1,018	16,572
Low	1,346	549	451	2,346
Total	41,428	8,603	2,215	-

Table 5.6: Statistics of interest annotation.

internal state annotations. For subjective knowledge labels, *High* is the most common, followed by *Low*. The distribution for subjective interest labels is more imbalanced than knowledge labels with *High* being particularly dominant. This is probably because recommenders usually advance a dialogue toward topics of interest to the seekers. For objective labels, the number of *Neutral* labels increases in both knowledge and interest. This is because it is difficult for recommenders to judge the seeker's internal state of some entities.

We calculate the agreement and Pearson correlation between the subjective and objective labels. The agreement is 0.53 for knowledge and 0.62 for interest labels, and the Pearson correlation is 0.27 for knowledge and 0.21 for interest. This result suggests that it is difficult to substitute subjective labels with objective ones.

Relationship between Knowledge and Interest We explore the correlation between subjective knowledge and interest labels for the same entities. The Pearson correlation coefficient is 0.12, indicating no correlation. This result means

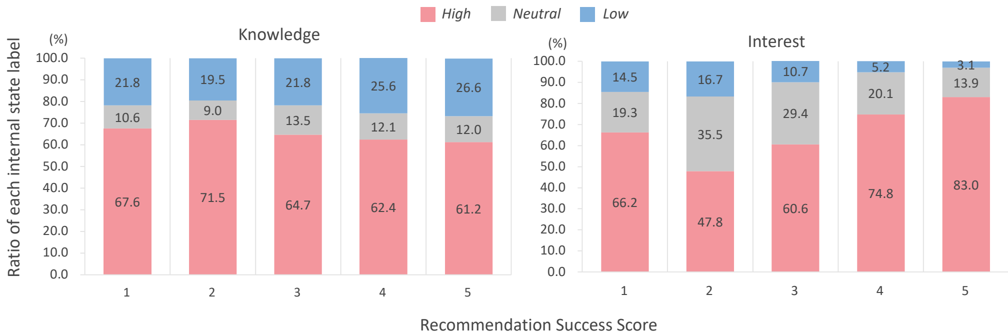


Figure 5.3: Relationship between recommendation success score and the ratio of each internal state label.

that knowledge and interest represent different facets of the internal state.

Contribution of Seeker's Internal State to Recommendation Success We investigate the relationship between the subjective seeker's internal state and recommendation success at the dialogue level. We use the seeker's answer to Q4 (i.e., "Do you want to watch the recommended movie?") as an indication of recommendation success. Figure 5.3 shows that dialogues with high recommendation success scores tend to have more *Low* knowledge entities. For interest, on the other hand, dialogues with high recommendation success scores tend to have more *High* interest entities.

We next analyze the dialogues with entities of *Low* knowledge and *High* interest in comparison with those dialogues without these kinds of entities. The average recommendation success score for the former dialogues is 4.59, while that for the latter dialogues is 4.18. Student's t-test result reveals that the difference is statistically significant at the $p = 0.05$ level. The above analysis results indicate it is important in recommendation dialogues to identify and mention the topics where the seeker has no knowledge but has an interest.

Next, we explore the relationship between the subjective seeker's internal state and recommendation success at the utterance level for detailed analysis. To this end, we randomly selected 1,000 pairs of recommender's utterances and preceding dialogue context from our constructed dataset. We then ask crowdworkers to

Knowledge	Interest	✓	✗
High	*High*	**3.61**	**3.61**
High	*Low*	3.59	**3.61**
Low	*High*	**3.72***	3.53
Low	*Low*	3.56	**3.61**

Table 5.7: Difference in recommendation success by each entity. ✓and ✗ denote the presence and absence of the entity in the utterance, respectively. The asterisk (*) indicates that the difference is statistically significant at the $p = 0.05$ level. Wilcoxon rank-sum test is used as a statistical test.

evaluate whether the utterance makes the interlocutor interested in watching a movie, using a 5-point Likert scale (5 is the best). Three workers evaluate each utterance, and the scores are averaged. Table 5.7 shows the results. The score is high when the recommender's utterance includes entities with *Low* knowledge and *High* interest. The above results confirm that the recommender can effectively recommend by mentioning entities the seeker does not know but is interested in, even at the utterance level.

5.4 Experiment

The analysis in Section 5.3.3 suggests the importance of understanding the seeker's internal state at the entity level. Thus, we propose a response generation framework that explicitly considers the seeker's internal state at the entity level. In this section, we describe our proposed method and verify its effectiveness.

5.4.1 Proposed Method

We propose a LLM-based response generation framework that explicitly considers the seeker's internal state labels. Specifically, we apply Chain-of-Thought prompting (Wei et al., 2022) to our task and estimate the seeker's internal state prior to generating a response.

Figure 5.4 shows an overview of our proposed method. Given the movie information and the dialogue history as the input for the model, the model first

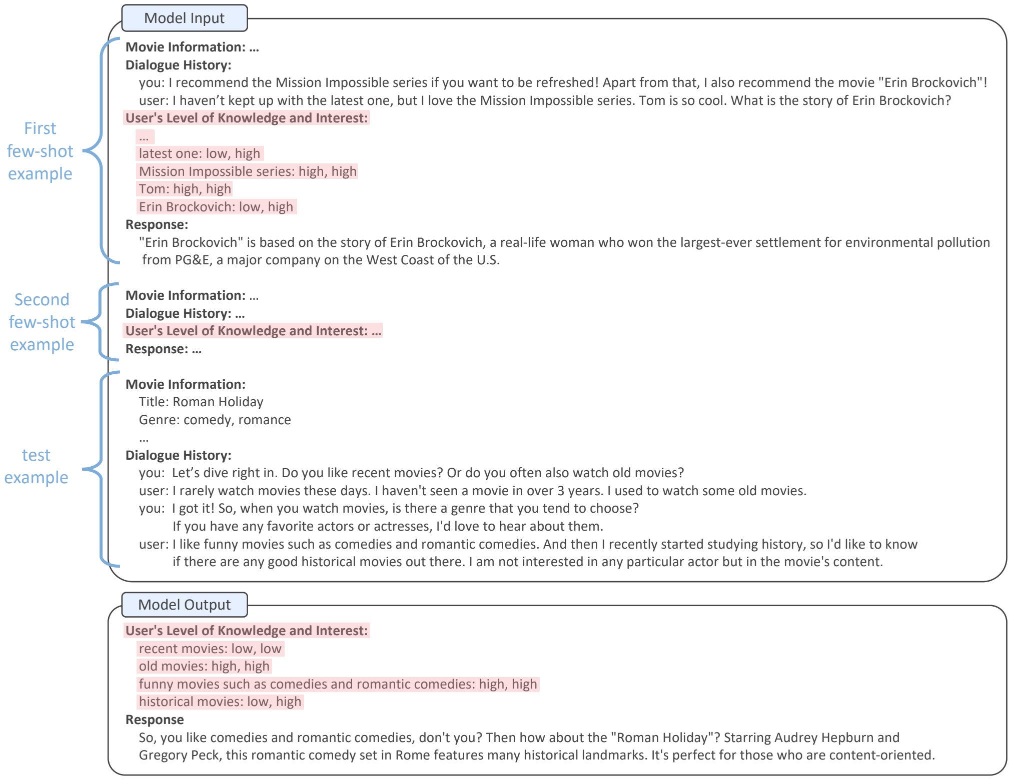


Figure 5.4: Overview of our proposed method. The internal state estimation, which is highlighted, is performed only for the proposed method and not for the baseline method.

extracts target entities. It then estimates the seeker's level of knowledge and interest in each entity at three levels: *High*, *Neutral*, and *Low*, respectively. Finally, the model generates a response referring to the estimated internal state.

5.4.2 Experimental Settings

Model

We use GPT-4 (`gpt-4-0613`) (OpenAI, 2023), which achieves outstanding performance on various language-related tasks, as the base model. We selected GPT-4 because of its remarkable performance in JGLUE (Kurihara et al., 2022), the general natural language understanding benchmark for Japanese.[7]

We compare the following four methods with a baseline that generates responses without estimating the seeker's internal state:

CoT (sub): uses the subjective seeker's internal state labels in the few-shot examples.

CoT (obj): uses the objective seeker's internal state labels in the few-shot examples.

CoT (sub, gold): uses CoT (sub) method but is given the correct labels of the seeker's internal state in the test example, and only the response generation is performed. This method is for the purposes of an ablation study.

CoT (obj, gold): uses CoT (obj) method but is given the correct labels of the seeker's internal state in the test example, and only the response generation is performed. This method is for the purposes of an ablation study.

Dataset

We randomly split the collected dialogues into 85%:15% for training and test data, respectively. We selected the candidates for few-shot examples from the training data based on the following two criteria: (1) including all types of entity labels for knowledge and interest within the dialogue context, and (2) ensuring that the response incorporates an entity with *Low* knowledge and *High* interest. The second constraint is based on the findings in Section 5.3.3, and was established to use higher-quality responses as few-shot examples. Consequently, we obtained 217 few-shot examples for CoT (sub) and 150 few-shot examples for CoT (obj). As for the test example, we randomly selected 500 examples from the test split only using the first criterion. For each test example, we then randomly chose two few-shot examples from the candidate pool.

[7] http://nejumi.ai/

Model	Consistency	Seeker's Knowledge	Seeker's Interest	Tailored Information	Recommendation Success
CoT (sub)	52.2*	51.5	52.5*	51.4	52.1*
CoT (obj)	51.4	52.1*	52.2*	52.3*	51.3
CoT (sub, gold)	54.5*	54.2*	54.8*	55.0*	56.0*
CoT (obj, gold)	53.0*	51.6	53.0*	52.7*	53.5*

Table 5.8: Results of the response generation. The asterisk (*) indicates that the difference is statistically significant at the $p = 0.05$ level using a binomial test.

5.4.3 Result

In this study, we conduct a human evaluation to assess the quality of the responses generated by the proposed methods. Specifically, we present the responses of each method in Section 5.4.2 and the baseline method to crowdworkers along with the corresponding dialogue context. Subsequently, these crowdworkers are requested to select which response is superior concerning the following five evaluation metrics.

Consistency The response is consistent with dialogue context.

Seeker's Knowledge The response considers the seeker's level of knowledge.

Seeker's Interest The response considers the seeker's level of interest.

Tailored Information The response provides more information that the seeker does not know but is interested in.

Recommendation Success The response is more likely to entice the seeker to watch the recommended movie.

Table 5.8 presents the win rates compared to the baseline method. Our proposed methods, CoT (sub) and CoT (obj), surpassed the baseline across all metrics. Notably, the improvements were statistically significant in the categories of Consistency, Seeker's Interest, and Recommendation Success for CoT (sub), and in Seeker's Knowledge, Seeker's Interest, and Tailored Information for CoT (obj).

In addition, when correct labels were provided for the seeker's internal state estimation, there was a further improvement in the win rate. Notably, CoT (sub, gold) exhibited a higher win rate than CoT (obj, gold), indicating that considering the subjective (i.e., actual) seeker's internal state is effective in generating responses.

5.4.4 Discussion

In this section, we analyze the results of the seeker's internal state estimation, which is an intermediate task in our proposed framework. We consider the results divided into entity extraction and internal state classification.

Entity Extraction

We use precision and recall scores for exact matching as strict evaluation metrics and use the character-level F1 score as a lenient evaluation metric. To calculate the character-level F1 score, we first calculate the maximum character-level F1 score between each gold entity and the predicted entities. Then, we compute the average of these maximum values across all gold entities.

The precision and recall scores for the CoT (sub) model were observed to be 44.1 and 47.8, respectively, while the CoT (obj) model yielded scores of 42.7 and 46.3. These figures are relatively low, indicating a challenge in the model's ability to accurately estimate the precise spans of entities, particularly in terms of determining which modifiers should be included within the entity span. In contrast, the character-level F1 scores for the respective models exhibited higher values, achieving 76.2 and 76.1. This disparity in performance suggests that while the model encounters difficulties with precise entity span estimation, it is relatively adept at estimating approximate spans.

Seeker's Internal State Classification

We assess the classification performance of the seeker's internal state labels for successfully extracted entities using F1 score metric.

Table 5.9 shows the results. In the context of knowledge and interest estimation, CoT (sub) and CoT (obj) demonstrated superior accuracy in predicting *High* and *Low* levels compared to human interlocutors (recommenders). However,

	Knowledge			Interest		
	High	*Neutral*	*Low*	*High*	*Neutral*	*Low*
CoT (sub)	**74.2**	9.9	**49.5**	**84.7**	23.1	**26.9**
Recommender	70.4	**14.4**	46.4	76.3	**27.6**	25.5
CoT (obj)	**73.1**	14.2	**47.8**	**83.0**	20.4	**22.8**
Recommender	72.2	**16.5**	39.8	76.6	**28.1**	19.2

Table 5.9: Results of seeker's internal state classification.

for *Neutral*, humans outperformed these models, indicating potential areas for further improvement. Additionally, when comparing CoT (sub) and CoT (obj), CoT (sub) generally achieved higher accuracy, suggesting the effectiveness of utilizing subjective labels.

Furthermore, knowledge and interest were estimated with relatively high accuracy for the *High* category. Conversely, the *Low* category exhibited lower accuracy, particularly regarding interest estimation. This lower performance is likely due to the imbalanced distribution of labels within the dataset. However, the primary focus of the study remains on the accurate identification of topics with *High* interest in the context of recommendation dialogues rather than the identification of *Low* interest topics. Consequently, this finding does not significantly detract from the overall utility of our proposed framework in recommendation scenarios.

5.5 Summary of This Chapter

We constructed RecMind, a recommendation dialogue dataset that features both subjective and objective annotations of the seeker's internal state at the entity level. Our dataset also has engaging dialogues with longer seeker's utterances, characterized by high scores in dialogue enjoyment and recommendation success. We also proposed a response generation framework that explicitly considers the seeker's internal state, applying Chain-of-Thought prompting to our task. The experimental results showed that our proposed method could generate responses that are more consistent and tailored to the seeker than the baseline method.

Our dataset has diverse and fine-grained annotations, which are useful for various tasks such as internal state estimation, external knowledge selection, and dialogue response generation. We hope our dataset will be useful for future research on recommendation dialogues.

Chapter 6

Conclusion

6.1 Overview

In this thesis, we addressed the two dialogues system challenges: the understanding of the user internal states and the utilization of external knowledge. To tackle these challenges, we adopted corpus-based approaches, attempting to improve dialogue systems.

In Chapter 2, we studied the dialogue management based on user internal states. We curated 1,060 dialogues between a rule-based dialogue system and humans. We then annotate all the user utterances with knowledge, interest, and willingness, respectively. Using the constructed corpus, we train classifiers to estimate each user internal state. These classifiers achieved a high estimation accuracy of approximately 80% for each internal state when allowing an error of ± 1 on a 7-point scale. Finally, we added rules to change responses according to the estimated user internal states. The dialogue-level and utterance-level evaluations showed that the response change improved the naturaleness of system responses.

In Chapter 3, we studied dialogue response generation based on external knowledge. We first constructed Chat Collection Framework, which enables us to construct large-scale, human-to-human dialogue datasets in Japanese. Utilizing this framework, we newly constructed the JMRD dataset, which is a Japanese external knowledge-grounded dialogue dataset. This dataset consists of about 5,200 dialogues, where the recommender utterances are grounded on movie-related external knowledge. Employing the constructed dataset, we presented a strong baseline model that simultaneously selects external knowledge and generates responses based on it.

In Chapter 4, we analyzed the engagingness for external knowledge-grounded responses. We extracted the entities from the recommender utterances in JMRD dataset and annotated them with its information source: database-derived, speaker-derived, or other. We then analyzed the annotated corpus and showed the use of speaker-derived information, as well as external knowledge, made responses more engaging.

In Chapter 5, we studied response generation based on both user internal states and external knowledge. We proposed RecMind, a Japanese movie recommendation dialogue dataset with subjective and objective annotations of the seeker's internal state at the entity level. We curated this data by recruiting movie enthusiasts who have a deep knowledge of movies, leading to more engaging dialogues. Through the analysis of this dataset, we found entities the seeker has no knowledge about but has an interest in contributed to the successful recommendation. Furthermore, we proposed the LLM-based response generation framework that explicitly considers the seeker's internal state, applying Chain-of-Thought prompting. The experimental results showed that our proposed method can generate responses that are more consistent and tailored to the seeker than baselines.

6.2 Future Prospects

6.2.1 Refining User Internal State Estimation

In this thesis, we tackled the estimation of three user internal states: knowledge, interest, and willingness from dialogue contexts. We employed two methods: a supervised fine-tuning approach using BERT and an in-context learning approach using LLM. However, there is still room for improvement in the accuracy of these methods.

One promising approach is to utilize common sense relationships to structure and refine the estimation of user internal states. For instance, it is reasonable to assume that a person interested in "Tom Cruise" might also have an interest in movies featuring him. Explicitly incorporating such common sense could lead to more accurate estimation.

6.2.2 Modeling User Internal States for Long-term Dialogues

In this thesis, we modeled user internal states in recommendation dialogues using knowledge, interest, and willingness. This modeling has effectively helped dialogue systems understand user intentions and generate responses. However, this modeling does not fully capture the user internal states.

Dialogue systems will be expected to assist users over the long term (Xu et al., 2022), engaging in multiple dialogues over an extended period. Rapport is an interesting internal state in such long-term dialogues. Sensing the extent to which the rapport has been established with the user and altering the communication style accordingly could lead to a more natural and comfortable user experience.

6.2.3 Improving External Knowledge Retrieval based on User Internal States

Most existing research on external knowledge-grounded dialogue systems has relied on only the dialogue context for retrieving relevant external knowledge. However, the user internal states could be used to enhance the external knowledge retrieval. For example, a system could prioritize information that aligns with the user's interests or adapt the complexity of the information based on the user's knowledge level. This explicit use of user internal states in retrieving external knowledge could lead to more personalized and engaging responses.

Appendix A

Dialogue Management Based on User Internal States

A.1 Annotation Reliability for Each Annotator

In this section, we describe the annotation reliability for each annotator. The annotation reliability for each annotator is measured by the match rate m with other annotators. The match rate m for each annotator is calculated as follows:

$$m = \frac{\text{\# of annotations by target annotator agreed upon with other annotators}}{\text{\# of annotations by target annotator}} \tag{A.1}$$

The average match rate $\overline{m}$ for *All* data was 0.51. If we take annotators with match rates below 0.33 as low reliable annotators, then out of the 458 annotators in the study, 41 were identified as low reliability. This figure represents about 9% of the total, showing that most annotators were reliable.

Interestingly, when we excluded the data from these 41 annotators, Krippendorff's alpha values for knowledge, interest, and willingness were 0.45, 0.45, and 0.39, respectively. These values are quite similar to the original alpha values (knowledge: 0.41, interest: 0.40, willingness: 0.35), indicating that the exclusion of the unreliable annotators did not significantly impact the results.

A.2 Improvement of Dialogue System

This section describes some improvements for our movie recommendation dialogue system. These improvements are applied only at the evaluation stage (Section 2.5.2), and not at the corpus construction stage (Section 2.4.1).

Movie Database Expansion The number of movies stored in the movie database was increased from 213 to 331 so that the latest movies can be used for recommendation.

Filtering Selling Points We refined our method for extracting selling points in reviews. First, we analyzed the selling point sentences extracted using the previous word2vec-based method. As a result, we found that some sentences are appropriate as human utterances, such as "I found that movie interesting and will go to see it again next week," but are not suitable as dialogue systems. Thus, we excluded review sentences that met all of the following three conditions from the selling point candidates.

Condition 1 : Subject of any predicate in review is the review author (i.e., the user who wrote the review)

Condition 2 : Predicates that meet Condition 1 are intentional.

Condition 3 : Predicates that meet Condition 1 are inappropriate to use as a position in a dialogue system

Condition 1 is determined by identifying the nominative case of the predicate using the existing analyzer, *Base+coref+noun+bridge* model (Ueda et al., 2020) trained on the Kyoto University Web Document Read Corpus (KWDLC) (Hangyo et al., 2012) and the Kyoto University text corpus (Kawahara et al., 2002). In the experiments conducted with three different random seeds on the test data of KWDLC, the average F1 score for the zero exophora resolution was 0.745.

Regarding Condition 2, we used the existing rule-based classifiers.[1]

[1] https://github.com/ku-nlp/ishi

Condition 3 is judged by whether the predicate is included in a specially curated list of 53 inappropriate predicates. This list was derived by initially extracting 103 predicates, each occurring more than 50 times, from 2,901 predicates that met Conditions 1 and 2. Subsequently, we meticulously selected 53 of these.

Refining Sentiment Analysis for Review Texts We also changed the method of determining the sentiment of the review sentence. In this study, we used an existing sentiment classifier (Saito et al., 2019). Using this classifier, we divided the review sentence into event units and regarded the sentiment of the last event as the sentiment of the entire review sentence. The sentiment score ranges from 1 to -1, with 1 being the most positive and -1 being the most negative. This sentiment score and the number of characters in each sentence are multiplied to obtain a score, from which the top 50 sentences are extracted as candidates for recommended highlight sentences. This process is conducted for each movie, and within each scenario, two sentences are randomly selected from the top 50 for use.

Bibliography

[1] Jaime Acosta. Using emotion to gain rapport in a spoken dialog system. In Ulrich Germann, Chirag Shah, Svetlana Stoyanchev, Carolyn Penstein Rosé, and Anoop Sarkar, editors, Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium, pages 49–54, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL `https://aclanthology.org/N09-3009`.

[2] Daniel De Freitas Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. ArXiv, abs/2001.09977, 2020.

[3] Tsunehiro Arimoto, Hiroaki Sugiyama, Masahiro Mizukami, Hiromi Narimatsu, and Ryuichiro Higashinaka. Analysis of satisfaction and topics in repeated conversation through days. In Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts, London, United Kingdom, September 2019. SEMDIAL. URL `http://semdial.org/anthology/Z19-Arimoto_semdial_0020.pdf`.

[4] Tsunehiro Arimoto, Hiroaki Sugiyama, Hiromi Narimatsu, and Masahiro Mizukami. Daily text chat collection and analysis for developing long-term relational dialogue systems. Proceedings of the Annual Conference of JSAI, JSAI2023:3O1OS2c02–3O1OS2c02, 2023. doi: 10.11517/pjsai.JSAI2023.0_3O1OS2c02.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, 2015.

[6] Rafael E. Banchs and Haizhou Li. IRIS: a chat-oriented dialogue system based on the vector space model. In Min Zhang, editor, Proceedings of the ACL 2012 System Demonstrations, pages 37–42, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `https://aclanthology.org/P12-3007`.

[7] Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1112–1125, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.85. URL `https://aclanthology.org/2021.emnlp-main.85`.

[8] Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. In Trevor Cohn, Yulan He, and Yang Liu, editors, Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3897–3909, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.347. URL `https://aclanthology.org/2020.findings-emnlp.347`.

[9] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. In ICLR, 2017. URL `https://openreview.net/forum?id=S1Bb3D5gg`.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger,

Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[11] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In Eduardo Blanco and Wei Lu, editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2029. URL https://aclanthology.org/D18-2029.

[12] Yuya Chiba, Masashi Ito, Takashi Nose, and Akinori Ito. User modeling by using bag-of-behaviors for building a dialog system sensitive to the interlocutor's internal state. In Kallirroi Georgila, Matthew Stone, Helen Hastie, and Ani Nenkova, editors, Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 74–78, Philadelphia, PA, U.S.A., June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4310. URL https://aclanthology.org/W14-4310.

[13] Mathieu Chollet, Helmut Prendinger, and Stefan Scherer. Native vs. non-native language fluency implications on multimodal interaction for interpersonal skills training. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 386—-393, 2016.

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma,

Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[15] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf`.

[16] H.H. Clark. Using Language. Cambridge University Press, 1996.

[17] Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In Proceedings of the ACM Web Conference 2022, WWW '22, page 2998–3008, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512020. URL `https://doi.org/10.1145/3485447.3512020`.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.

BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[19] John M. Digman. Personality structure: Emergence of the five-factor model. Annual Review of Psychology, 41(1):417–440, 1990. URL `https://doi.org/10.1146/annurev.ps.41.020190.002221`.

[20] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In International Conference on Learning Representations, 2019. URL `https://openreview.net/forum?id=r1l73iRqKm`.

[21] Yue Feng, Yunlong Jiao, Animesh Prasad, Nikolaos Aletras, Emine Yilmaz, and Gabriella Kazai. Schema-guided user satisfaction modeling for task-oriented dialogues. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2079–2091, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.116. URL `https://aclanthology.org/2023.acl-long.116`.

[22] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, William B. Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. 2018.

[23] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In Proc. Interspeech 2019, pages 1891–1895, 2019. doi:

10.21437/Interspeech.2019-3079. URL http://dx.doi.org/10.21437/Interspeech.2019-3079.

[24] David Griol and Zoraida Callejas. A neural network approach to intention modeling for user-adapted conversational agents. Intell. Neuroscience, 2016, jan 2016. ISSN 1687-5265. doi: 10.1155/2016/8402127. URL https://doi.org/10.1155/2016/8402127.

[25] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In Katrin Erk and Noah A. Smith, editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1631–1640, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1154. URL https://aclanthology.org/P16-1154.

[26] Ao Guo, Atsumoto Ohashi, Ryu Hirai, Yuya Chiba, Yuiko Tsunomori, and Ryuichiro Higashinaka. Influence of user personality on dialogue task performance: A case study using a rule-based dialogue system. In Alexandros Papangelis, Paweł Budzianowski, Bing Liu, Elnaz Nouri, Abhinav Rastogi, and Yun-Nung Chen, editors, Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pages 263–270, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4convai-1.25. URL https://aclanthology.org/2021.nlp4convai-1.25.

[27] Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. Fine-grained post-training for improving retrieval-based dialogue systems. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1549–1558, Online, June 2021. Association for Computational Linguistics. doi: 10.

18653/v1/2021.naacl-main.122. URL https://aclanthology.org/2021.naacl-main.122.

[28] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In Ruli Manurung and Francis Bond, editors, Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, pages 535–544, Bali, Indonesia, November 2012. Faculty of Computer Science, Universitas Indonesia. URL https://aclanthology.org/Y12-1058.

[29] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. INSPIRED: Toward sociable recommendation dialog systems. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8142–8152, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.654. URL https://aclanthology.org/2020.emnlp-main.654.

[30] Michimasa Inaba and Kenichi Takahashi. Estimating user interest from open-domain dialogue. In Kazunori Komatani, Diane Litman, Kai Yu, Alex Papangelis, Lawrence Cavedon, and Mikio Nakano, editors, Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, pages 32–40, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5004. URL https://aclanthology.org/W18-5004.

[31] Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. Talking with ERICA, an autonomous android. In Raquel Fernandez, Wolfgang Minker, Giuseppe Carenini, Ryuichiro Higashinaka, Ron Artstein, and Alesia Gainer, editors, Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 212–215, Los Angeles, September 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3625. URL https://aclanthology.org/W16-3625.

[32] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. En-

gagement recognition in spoken dialogue via neural network by aggregating different annotators' models. In INTERSPEECH 2018, pages 616–620, 2018.

[33] Takuya Ishihara, Katsumi Nitta, Fuminori Nagasawa, and Shogo Okada. Estimating interviewee's willingness in multimodal human robot interview interaction. In Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct, ICMI '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360029. doi: 10.1145/3281151.3281153. URL `https://doi.org/10.1145/3281151.3281153`.

[34] Meihuizi Jia, Ruixue Liu, Peiying Wang, Yang Song, Zexi Xi, Haobin Li, Xin Shen, Meng Chen, Jinhui Pang, and Xiaodong He. E-ConvRec: A large-scale conversational recommendation dataset for E-commerce customer service. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5787–5796, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.622`.

[35] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2095–2104, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.169. URL `https://aclanthology.org/2021.naacl-main.169`.

[36] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. Recommendation as a communication game:

Self-supervised bot-play for goal-oriented dialogue. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1951–1961, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1203. URL `https://aclanthology.org/D19-1203`.

[37] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese relevance-tagged corpus. In Manuel González Rodríguez and Carmen Paz Suarez Araujo, editors, Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2002/pdf/302.pdf`.

[38] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3405–3424, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.274. URL `https://aclanthology.org/2021.emnlp-main.274`.

[39] Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. Sequential latent

knowledge selection for knowledge-grounded dialogue. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=Hke0K1HKwr.

[40] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

[41] Takashi Kodama, Ryuichiro Higashinaka, Koh Mitsuda, Ryo Masumura, Yushi Aono, Ryuta Nakamura, Noritake Adachi, and Hidetoshi Kawabata. Collection of meta information with user-generated question answer pairs and its reflection for improving expressibility in response generation. Journal of Natural Language Processing, 28(1):136–159, 2021. doi: 10.5715/jnlp.28.136.

[42] Takashi Kodama, Ribeka Tanaka, and Sadao Kurohashi. Construction of hierarchical structured knowledge-based recommendation dialogue dataset and dialogue system. In Song Feng, Hui Wan, Caixia Yuan, and Han Yu, editors, Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 83–92, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.dialdoc-1.9. URL https://aclanthology.org/2022.dialdoc-1.9.

[43] Kazunori Komatani and Shogo Okada. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8, 2021. doi: 10.1109/ACII52823.2021.9597447.

[44] Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. Flexible guidance generation using user model in spoken dialogue systems. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 256–263, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075129. URL https://aclanthology.org/P03-1033.

[45] Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. Collecting sharable multimodal dialogue data and analysis of annotation disagreement tendency. SIG-SLUD, 5(02):45–50, 2018. (in Japanese).

[46] Klaus Krippendorff. Content Analysis: An Introduction to Its Methodology (second edition). Sage Publications, 2004.

[47] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 2957–2966, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.317.

[48] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. Computational Linguistics, 20(4):507–534, 1994. URL https://aclanthology.org/J94-4001.

[49] Bishal Lamichhane. Evaluation of chatgpt for nlp-based mental health applications, 2023.

[50] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

[51] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In Katrin Erk and Noah A. Smith, editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1094. URL https://aclanthology.org/P16-1094.

[52] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, pages 9748–9758, Red Hook, NY, USA, 2018. Curran Associates Inc.

[53] Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2810–2823, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.202. URL https://aclanthology.org/2022.naacl-main.202.

[54] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1147.

[55] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. OpenSubtitles2018:

Statistical rescoring of sentence alignments in large, noisy parallel corpora. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1275.

[56] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[57] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. Towards conversational recommendation over multi-type dialogs. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1036–1049, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.98. URL https://aclanthology.org/2020.acl-main.98.

[58] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

[59] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In Alexander Koller, Gabriel Skantze, Filip Jurcicek, Masahiro Araki, and Carolyn Penstein Rose, editors, Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 285–294, Prague, Czech Republic, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4640. URL https://aclanthology.org/W15-4640.

[60] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-

Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic Turing test: Learning to evaluate dialogue responses. In Regina Barzilay and Min-Yen Kan, editors, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1116–1126, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1103. URL https://aclanthology.org/P17-1103.

[61] Junyu Lu, Xiancong Ren, Yazhou Ren, Ao Liu, and Zenglin Xu. Improving contextual language models for response retrieval in multi-turn conversation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, page 1805–1808, New York, NY, USA, 2020. Association for Computing Machinery. URL https://doi.org/10.1145/3397271.3401255.

[62] Toyomi Meguro, Ryuichiro Higashinaka, Kohji Dohsaka, Yasuhiro Minami, , and Hideki Isozaki. Effects of personality traits on listening-oriented dialogue. In Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS), pages 104–107, 2009.

[63] Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. Initiative-aware self-supervised learning for knowledge-grounded conversations. In SIGIR, pages 522–532, 2021.

[64] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013. URL http://arxiv.org/abs/1301.3781.

[65] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. ParlAI: A dialog research software platform. In Lucia Specia, Matt Post, and Michael Paul, editors, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 79–84, Copenhagen,

Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2014. URL https://aclanthology.org/D17-2014.

[66] Chiaki Miyazaki, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. Estimating callers' levels of knowledge in call center dialogues. In INTERSPEECH 2013, pages 2866–2870, Lyon, France, 2013.

[67] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. Towards exploiting background knowledge for building conversation systems. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2322–2332, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1255. URL https://aclanthology.org/D18-1255.

[68] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 845–854, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1081. URL https://aclanthology.org/P19-1081.

[69] Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2292–2297, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1276. URL https://aclanthology.org/D15-1276.

[70] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In 23rd International Conference on Intelligent User Interfaces,

IUI '18, page 153–164, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349451. doi: 10.1145/3172944.3172969. URL https://doi.org/10.1145/3172944.3172969.

[71] Mikio Nakano, Kazunori Komatani, Kotaro Funakoshi, Yukiko Nakano, and Manabu Okumura. Dialogue Systems. Corona Publishing Co., Ltd., 2015.

[72] Olabiyi Oluwatobi and Erik Mueller. DLGNet: A transformer-based model for dialogue response generation. In Tsung-Hsien Wen, Asli Celikyilmaz, Zhou Yu, Alexandros Papangelis, Mihail Eric, Anuj Kumar, Iñigo Casanueva, and Rushin Shah, editors, Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, pages 54–62, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.7. URL https://aclanthology.org/2020.nlp4convai-1.7.

[73] OpenAI. GPT-4 Technical Report, 2023.

[74] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[75] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[76] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party

dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050.

[77] Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

[78] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1(4):515–526, 1978.

[79] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Assigning personality/profile to a chatting machine for coherent conversation generation. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/595. URL https://doi.org/10.24963/ijcai.2018/595.

[80] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

[81] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das.

Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 704–718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.58. URL `https://aclanthology.org/2021.acl-long.58`.

[82] Alan Ritter, Colin Cherry, and William B. Dolan. Data-driven response generation in social media. In Regina Barzilay and Mark Johnson, editors, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 583–593, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `https://aclanthology.org/D11-1054`.

[83] Pedro Rodriguez, Paul Crook, Seungwhan Moon, and Zhiguang Wang. Information seeking in the spirit of learning: A dataset for conversational curiosity. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8153–8172, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.655. URL `https://aclanthology.org/2020.emnlp-main.655`.

[84] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL `https://aclanthology.org/2021.eacl-main.24`.

[85] Jun Saito, Yugo Murawaki, and Sadao Kurohashi. Minimally supervised learning of affective events using discourse relations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5758–5765, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1581. URL `https://aclanthology.org/D19-1581`.

[86] Björn W. Schuller, Niels Köhler, Ronald Müller, and Gerhard Rigoll. Recognition of interest in human conversational speech. In INTERSPEECH 2006, pages 793–796, 2006.

[87] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In Regina Barzilay and Min-Yen Kan, editors, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL `https://aclanthology.org/P17-1099`.

[88] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith, editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://aclanthology.org/P16-1162`.

[89] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, page 3776–3783, 2016.

[90] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In Chengqing Zong and Michael Strube, ed-

itors, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1577–1586, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1152. URL https://aclanthology.org/P15-1152.

[91] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3784–3803, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.320. URL https://aclanthology.org/2021.findings-emnlp.320.

[92] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage, 2022.

[93] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2021–2030, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.183. URL https://aclanthology.org/2020.acl-main.183.

[94] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. Generating responses with a specific emotion in dialog. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3685–3695,

Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1359. URL https://aclanthology.org/P19-1359.

[95] S. S. Stevens. On the theory of scales of measurement. Science, 103(2684): 677–680, 1946.

[96] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[97] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3):339–374, 2000. URL https://aclanthology.org/J00-3003.

[98] Hui Su, Xiaoyu Shen, Zhou Xiao, Zheng Zhang, Ernie Chang, Cheng Zhang, Cheng Niu, and Jie Zhou. MovieChats: Chat like humans in a closed domain. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6605–6619, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.535. URL https://aclanthology.org/2020.emnlp-main.535.

[99] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[100] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du,

YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022.

[101] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A morphological analysis toolkit for scriptio continua. In Eduardo Blanco and Wei Lu, editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 54–59, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2010. URL `https://aclanthology.org/D18-2010`.

[102] Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. Integrated anaphora and coreference resolution with bert and refinement network. In Proceedings of 26th Annual Meeting of Association for Natural Language Processing, pages 1101–1104, 2020. (in Japanese).

[103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 5998–6008, 2017.

[104] Oriol Vinyals and Quoc V. Le. A neural conversational model. In Proceedings of ICML Deep Learning Workshop, 2015.

[105] Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. Proceedings of the AAAI Conference on Artificial Intelligence, 35:14006–14014, 05 2021.

[106] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, 2022. URL `https://openreview.net/forum?id=_VjQlMeSB_J`.

[107] 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.

[108] Terry Winograd. Understanding natural language. Academic Press, 1972.

[109] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. CoRR, abs/1901.08149, 2019. URL `http://arxiv.org/abs/1901.08149`.

[110] Wanqi Wu and Tetsuya Sakai. Response generation based on the big five personality traits. 2020. URL `https://api.semanticscholar.org/CorpusID:235666838`.

[111] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. Proactive human-machine conversation with explicit conversation goal. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3794–3804, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1369. URL `https://aclanthology.org/P19-1369`.

[112] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.356. URL `https://aclanthology.org/2022.acl-long.356`.

[113] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 6056–6077, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.370. URL `https://aclanthology.org/2023.emnlp-main.370`.

[114] Fanghua Ye, Zhiyuan Hu, and Emine Yilmaz. Modeling user satisfaction dynamics in dialogue via Hawkes process. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8875–8889, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.494. URL `https://aclanthology.org/2023.acl-long.494`.

[115] Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. Augmenting knowledge-grounded conversations with sequential knowledge transition. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5621–5630, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.446. URL `https://aclanthology.org/2021.naacl-main.446`.

[116] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

[117] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[118] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 270–278, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.30. URL https://aclanthology.org/2020.acl-demos.30.

[119] Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. Knowledge-grounded dialogue generation with pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3377–3390, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.272. URL https://aclanthology.org/2020.emnlp-main.272.

[120] Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn

knowledge-driven conversation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7098–7108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.635. URL https://aclanthology.org/2020.acl-main.635.

[121] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 708–713, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1076. URL https://aclanthology.org/D18-1076.

[122] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. Towards topic-guided conversational recommender system. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, Proceedings of the 28th International Conference on Computational Linguistics, pages 4128–4139, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.365. URL https://aclanthology.org/2020.coling-main.365.

[123] Xianda Zhou and William Yang Wang. MojiTalk: Generating emotional responses at scale. In Iryna Gurevych and Yusuke Miyao, editors, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1128–1137, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1104. URL https://aclanthology.org/P18-1104.

[124] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In 2015 IEEE International Conference on Computer Vision (ICCV),

pages 19–27, Los Alamitos, CA, USA, dec 2015. IEEE Computer Society. doi: 10.1109/ICCV.2015.11. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.11.

List of Major Publications

[1] Takashi Kodama, Hirokazu Kiyomaru, Yin Jou Huang, and Sadao Kurohashi. RecMind: Japanese Movie Recommendation Dialogue with Seeker's Internal State. In *Arxiv*, 2024.

[2] Takashi Kodama, Hirokazu Kiyomaru, Yin Jou Huang, Taro Okahisa, and Sadao Kurohashi. Is a Knowledge-based Response Engaging?: An Analysis on Knowledge-Grounded Dialogue with Information Source Annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 237–240, 2023. (**Best Paper Award**).

[3] Takashi Kodama, Ribeka Tanaka, and Sadao Kurohashi. Construction of Hierarchical Structured Knowledge-based Recommendation Dialogue Dataset and Dialogue System. In *Proceedings of The 2nd DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 83–92, 2022.

[4] Takashi Kodama, Frederic Bergeron, Hayato Atarashi, Ribeka Tanaka, Wataru Sakata, and Sadao Kurohashi. A Framework for Collecting Japanese Dialogues Available in Crowdsourcing. In *Proceedings of 27th Annual Meeting of Association for Natural Language Processing*, pages 859–863, 2021. (in Japanese).

[5] Takashi Kodama, Ribeka Tanaka, and Sadao Kurohashi. Dialogue Management by Estimating User's Internal State Using the Movie Recommendation Dialogue. *Journal of Natural Language Processing*, 28(1):104–135, 2021. (in Japanese).

List of Other Publications

[6] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. KWJA: A Unified Japanese Analyzer Based on Foundation Models. In *Proceedings of the 61st*

Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 538–548, 2023.

[7] Taro Okahisa, Ribeka Tanaka, Takashi Kodama, Yin Jou Huang, Yugo Murawaki, and Sadao Kurohashi. Domain Knowledge Elicitation in a Corpus of Online Interview Dialogues on Culinary Arts. *Journal of Natural Language Processing*, 30(2):773–799, 2023. (Technical paper, in Japanese).

[8] Takashi Kodama, Nobuhiro Ueda, Kazumasa Omura, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Japanese Morphological Analysis Using Text Generation Model. In *Proceedings of 29th Annual Meeting of Association for Natural Language Processing*, pages 339–344, 2023. (in Japanese).

[9] Takumi Yoshikoshi, Takashi Kodama, and Sadao Kurohashi. Explicit Use of Topicality in Dialogue Response Generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 222–228, 2022.

[10] Taro Okahisa, Ribeka Tanaka, Takashi Kodama, Yin Jou Huang, and Sadao Kurohashi. Constructing a Culinary Interview Dialogue Corpus with Video Conferencing Tool. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*, pages 3131–3139, 2022.

[11] Takashi Kodama, Ryuichiro Higashinaka, Koh Mitsuda, Ryo Masumura, Yushi Aono, Ryuta Nakamura, Noritake Adachi, and Hidetoshi Kawabata. Collection of Meta Information with User-Generated Question Answer Pairs and its Reflection for Improving Expressibility in Response Generation. *Journal of Natural Language Processing*, 28(1):136–159, 2021.

[12] Akiko Aizawa, Frederic Bergeron, Junjie Chen, Fei Cheng, Katsuhiko Hayashi, Kentaro Inui, Hiroyoshi Ito, Daisuke Kawahara, Masaru Kitsuregawa, Hirokazu Kiyomaru, Masaki Kobayashi, Takashi Kodama, Sadao Kurohashi, Qianying Liu, Masaki Matsubara, Yusuke Miyao, Atsuyuki Morishima, Yugo Murawaki, Kazumasa Omura, Haiyue Song, Eiichiro Sumita,

Shinji Suzuki, Ribeka Tanaka, Yu Tanaka, Masashi Toyoda, Nobuhiro Ueda, Honai Ueoka, Masao Utiyama, and Ying Zhong (in alphabetical order). A System for Worldwide COVID-19 Information Aggregation. In *Proceedings of Workshop on NLP for COVID-19 (Part 2) at EMNLP2020*, 2020.

[13] Takashi Kodama, Ryuichiro Higashinaka, Koh Mitsuda, Ryo Masumura, Yushi Aono, Ryuta Nakamura, Noritake Adachi, and Hidetoshi Kawabata. Generating Responses that Reflect Meta Information in User-Generated Question Answer Pairs. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 5433–5441, 2020.

[14] Hirokazu Kiyomaru, Nobuhiro Ueda, Takashi Kodama, Yu Tanaka, Ribeka Tanaka, Daisuke Kawahara, and Sadao Kurohashi. CausalityGraph: A System to Organize Causes, Results, and Solutions of Events based on Structural Language Analysis. In *Proceedings of 26th Annual Meeting of Association for Natural Language Processing*, pages 1125–1128, 2020. (in Japanese).

[15] Takashi Kodama, Ribeka Tanaka, and Sadao Kurohashi. Collection and Analysis of Meaningful Dialogue by Constructing a Movie Recommendation Dialogue System. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue: Poster Abstracts*, 2019.