Spoken Dialogue System for Information Navigation based on Statistical Learning of Semantic and Dialogue Structure

Koichiro Yoshino

Abstract

The thesis addresses a framework of a spoken dialogue system that navigates information of text documents, such as news articles, based on statistical learning of semantic and dialogue structure. Conventional spoken dialogue systems require a clear goal of the user, but this assumption does not always hold. Moreover, domain knowledge and task flows of the systems are conventionally hand-crafted. This process is costly and hampers domain portability.

In this thesis, a task of information navigation that handles ambiguous user queries is proposed. The goal and procedure of the user are not well-defined, but the proposed system probes potential information demands of the user and presents relevant information proactively according to semantic similarity and dialogue context. The back-end domain knowledge of semantic structure is statistically learned from the domain corpus in an unsupervised manner. The proposed system has seven modules for information navigation. These modules are controlled using a dialogue manager based on statistical learning of dialogue structure.

Chapter 2 introduces a scheme of information extraction, which is defined by the Predicate-Argument (P-A) structure and realized in unsupervised statistical learning. A useful information structure for information navigation depends on the domain, and the proposed significance score based on the Naive Bayes model selects a set of important P-A structures from the parsed results of domain texts. A preliminary experiment suggests that the significance score effectively extracts important patterns.

Chapter 3 presents a novel text selection method for language modeling with Web texts for automatic speech recognition (ASR). Compared to the conventional approach based on the perplexity criterion, this method introduces a semantic-level relevance measure (significance score defined in Chapter 2) with the back-end knowledge base used in the dialogue system. It is used to filter semantically relevant sentences in the domain. Experimental evaluations in two different domains showed the effectiveness and generality of this method. The method combined with the perplexity measure results in significant improvement not only in ASR accuracy, but also in semantic and dialogue-level accuracy.

Chapter 4 presents the spoken dialogue module that navigates not only exact information for the user query, but also partially-matched information related to user interests. Based on the information structure represented by P-A structure, the dialogue module conducts question answering and proactive information presentation. In conjunction with the P-A significance score defined in Chapter 2, similarity measures of the P-A components are introduced to select relevant information. An experimental evaluation showed that the proposed system makes more relevant responses compared with the conventional system based on the “bag-of-words” scheme. The proactive presentation module was also implemented based on the relevance measure to the dialogue history.

Chapter 5 addresses an empirical approach to managing the proposed spoken dialogue system based on partially observable Markov decision process (POMDP). The POMDP has been widely used for dialogue management and is formulated for information navigation as a selection of modules. The POMDP-based dialogue manager receives a user intention and user focus, which are classified by spoken language understanding (SLU) based on discriminative models. These dialogue states are used for selecting appropriate modules by policy function, which is optimized by reinforcement learning. The reward function is defined by the quality of interaction to encourage long interaction of information navigation. Experimental evaluations with real dialogue sessions demonstrated that the proposed system outperforms the conventional rule-based system and the POMDP-based system that does not track the user focus in terms of dialogue state tracking and action selection.

Chapter 6 concludes this thesis. The proposed system is designed and trained in a domain-independent manner, so it can be ported to a variety of domains of the information navigation task.

Acknowledgments

First and foremost, I would like to express my gratitudes to Professor Tatsuya Kawahara for his hospitable supervision lasted for five and a half years from the start of my master course. I am very impressed by his considerate supervision, and I have been taught by him not only how to be a good researcher but also how to be a good teacher. His tireless supervision dictated my career decision and I would like to look up to him, a good researcher and a good teacher, in the future.

I wish to thank Professor Sadao Kurohashi and Professor Hisashi Kashima for agreeing to take part as members of my doctoral committee, and for their careful reviews and insightful advices on my thesis.

I would also like to thank Professor Shinsuke Mori for his kind advices and fruitful discussions on both my research and my career.

Thanks to Dr. Yuya Akita, who is an assistant professor of our laboratory, senior member of our student room. He walked me through the technical issues and his assistance was essential to my research progress. Moreover, his supports on research environment for students had salutary effects on my research.

Dr. Katsuya Takanashi gave me so many important advices on a view point of linguistics. I thank him for his valuable suggestions.

I would also like to thank Professor Shun Ishizaki and Professor Kiyoko Uchiyama, who were supervisors of my undergraduate at Keio university. They gave me a catalyst to join the area of speech and language processing, especially the area of dialogue systems.

I would like to thank Dr. Shinji Watanabe and Dr. John R. Hershey, supervisors of my internship at Mitsubishi Electric Research Laboratories, and speech team members; Dr. Jonathan Le Roux, Mr. Bret A. Harsham and Mr. Yuki Tachioka. The experiments and discussions at the internship program improved my understanding of statistical dialogue

management, and I finished this work gracefully with them.

I would like to thank Dr. Jason D. Williams for his valuable and detailed advice on SIGDIAL 2014 mentoring program. This work was improved by his kind advices.

I learned the basis of spoken dialogue systems from Dr. Teruhisa Misu and Dr. Ryuichiro Higashinaka, during the internships at National Institute of Information and Communication Technology and NTT Communication Science Laboratories. I appreciate their considerate mentoring.

Special thanks to members of Kurohashi laboratory; Professor Daisuke Kawahara, Dr. Tomohide Shibata and Dr. Toshiaki Nakazawa, and alumni; Dr. Ryohei Sasano, Dr. Yugo Murawaki and Dr. Masatsugu Hangyo for their endless support on natural language processing.

Thanks to Professor Mamoru Komachi and Dr. Graham Neubig who provided solid answers to my questions and helpful advices.

Mr. Shunsuke Aihara, who was a senior member of the laboratory on my undergraduate, advised me on the implementation of my research. I can never thank him enough.

I would like to thank Takuma Ohtsuka, Ryo Yonetani and Shoko Tarumoto, fellows of my doctoral course, who encouraged my course work. I can not list all of friends here, but my course work was also encouraged by the help of my friends.

Of course, my time at Kyoto University was supported by members of Kawahara lab, students and staffs. Especially I would like to thank Ms. Mayumi Abe, secretary of our laboratory, who helped me so much. I also would like to thank Mr. Daniel Flannery and Ms. Mirzaei Maryam Sadat who checked my English, and Mr. Hirokuni Maeta with whom I often discuss technical or philosophical problems.

Parents, Yuka and all of the rest of the family have been endlessly encouraging and supportive. Finally, thank you Kasumi for always supporting me in my doctoral course.

Contents

Chapter 1

Introduction

1.1 Background

Studies on spoken dialogue systems now enter a new stage. A large number of spoken dialogue systems have been investigated and many systems are now deployed in the real world, most typically as smart phone applications, which interact with a diversity of users. However, a large majority of current applications is based on a specific task description which includes a definite task goal and necessary slots, such as place and date, for the task completion. Users are required to follow these concepts and they need to be aware of the clear task goal according to the system's capability. On the other hand, keyword search systems and question answering systems with speech interface are also developed for smart-phone applications. Such systems can provide answers to a variety of queries from users, but these systems do not conduct dialogue which involves an interaction with users, as they do not incorporate the domain knowledge and dialogue histories. Moreover, these systems work well only for simple keyword queries and factoid questions, but it is hard to deal with ambiguous user queries or non-factoid questions.

Some spoken dialogue agent systems in smart-phone applications are based on the combination of the above two schemes (Kawahara, 2009): a task-oriented system based on well-defined task domain knowledge using relational databases (RDB) and an open-domain system based on question answering using information retrieval techniques. The first scheme can achieve a well-defined task by using a structured database, but this scheme cannot be applied to Web information in which the structure and task are not well defined. The second scheme has been studied to handle large-scale texts such as the Web, but most of the conventional systems adopt a “bag-of-words” model, and naive statistical matching often generates irrelevant responses which have nothing to do with the user’s requests. The majority of current spoken dialogue systems are based on these two frameworks, thus, the capability of the systems is usually limited to factoid questions such as “when” or “how tall”, or pre-defined tasks such as “what is today’s weather?”. **Figure 1.1** depicts the position that the current systems are working on.

When users ask something beyond the system’s capability of the goal-oriented systems (upper left of the figure), current systems usually reply “I can’t answer the question”, or turns to the Web search and returns the retrieval list in the display (lower right of the figure). This kind of dialogue is not a natural interaction since people want to converse with them besides simple commands. A user-friendly conversational system should not reply with “I can’t answer the question” even if the system cannot find the result exactly matching the user query. Instead, it should present relevant information according to the user’s intention and preference by using domain knowledge and dialogue management that considers the dialogue history. The goal of this study is a conversational system with speech interface which can engage in information navigation. It is also plotted in Figure 1.1 (upper right of the figure). This thesis addresses a scheme that solves this problem by using statistical learning of semantic and dialogue structure from natural language texts, without constructing any RDB.

1.1.1 Conventional Goal-Oriented Systems

A large number of task-oriented spoken dialogue systems have been developed since 1980s. The norm of current spoken dialogue systems starts with single navigation systems such as Voyage (Hong et al., 1997) and Airline Travel Information System (ATIS (Dahl et al., 1994)). The task-oriented dialogue systems provide exact information or achieve the requested task, such as finding a restaurant in the area or searching flight to the destination. These systems achieve the well-defined task according to the well-organized domain-dependent task flow and database structure. This norm of spoken dialogue systems has been adopted in a variety of tasks and domains, weather information system (Zue et al., 2000), train information systems (Aust et al., 1995; Lamel et al., 2002), and bus navigation systems (Komatani et al., 2005; Raux et al., 2005). Such systems are also extended to

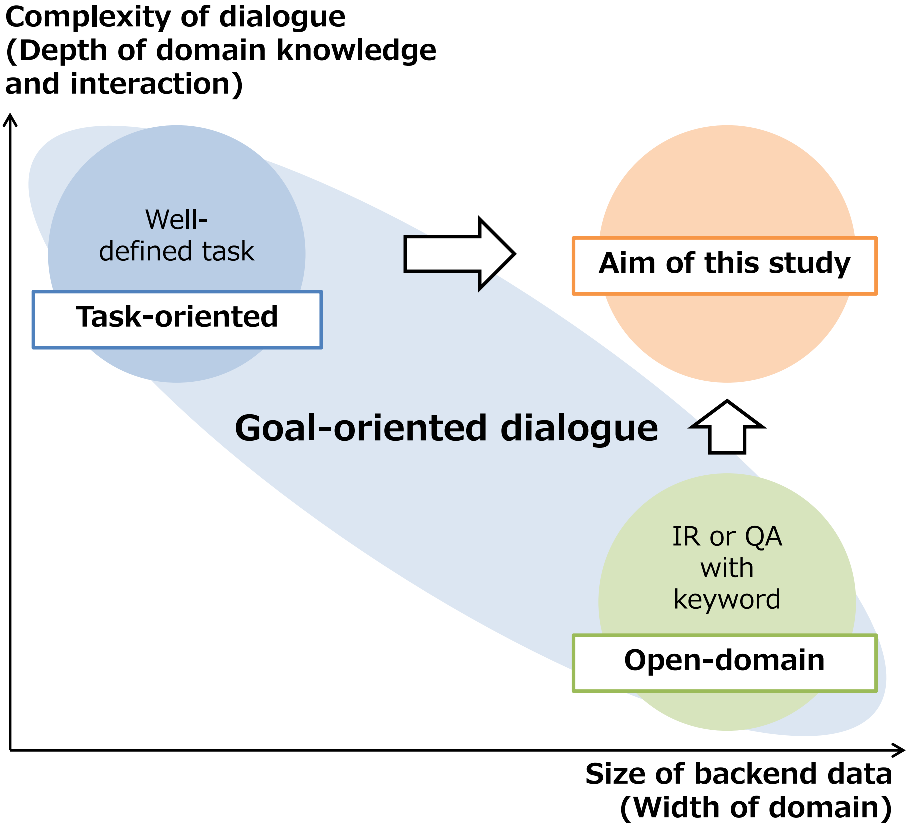


Figure 1.1: *Positions of current systems and this study.*

multi-domain systems such as Galaxy (Seneff et al., 1998) and Communicator (Walker et al., 2001). These systems have some components described below to operate a task based on the task flow and data structure.

On the other hand, question answering (QA) systems based on the information retrieval (IR) technique are developed since 1990s. The origin of existing open-domain QA system is MURAX (Kupiec, 1993), which finds an answer of factoid questions from encyclopedia. FAQ Finder (Burke et al., 1997) answers user's request by searching a database that describes frequent questions based on similarity between the user query and frequent questions. These systems also assume a clear goal of the user, i.e. what the user wants to know, and solve the problem of QA by constructing question-answer pairs from its back-end knowledge base.

The goal of these systems is to present information that is clearly requested by the user. Therefore, the systems are often called goal-oriented.

1.1.2 Goal, Task and Domain Knowledge

The typical type of conventional spoken dialogue systems assumes a clear goal. The goal is a unique destination of a dialogue that is shared by the user and the system. For example, the goal is searching a flight from Tokyo to New York in the airline travel task, or knowing the waiting time for the next bus in the bus navigation task. QA systems based on Web search also assume a clear user query, for which the unique answer exists on the Web. The majority of the QA systems deal with "factoid question", for example, "How tall is the Mt. Fuji". The task of spoken dialogue systems is defined to achieve the goal of the system. It is reduced to determine the necessary slots such as the destination place and date in the database, or question type and named entities (NEs). Domain knowledge is defined as a scope of the defined task. For example, the knowledge that "Charles de Gaulle" is an airport name in Paris is essential in the airline travel task. Domain knowledge can have some other types according to the task: tags of NEs or an element assigned to a slot. The aim of conventional dialogue systems is successfully reaching to the user goal as soon as possible.

The task flow and domain knowledge are conventionally hand-crafted. An example of the task structure in the multi-task navigation system, one of the typical smart-phone applications is shown in **Figure 1.2**. In this task structure, the system starts the task from greeting, and requests the user to select a food type or location or other keywords. If the system has some task candidates such as "phone call" or "show map", the system asks the user to select one. Finally, the system selects one task and makes the action requested by the user. Domain knowledge sources are essential for finishing the task. For example, "Phone call $RESTAURANT" is a typical experssion of the restaurant reservation task (the third task from the left). This strategy assumes that users already decided the task goal before the dialogue and they follow the task flow of the system.

1.1.3 Architecture of Conventional Systems

General spoken dialogue systems typically have components of automatic speech recognition (ASR), spoken language understanding (SLU), dialogue management (DM), natural language generation (NLG) and text-to-speech (TTS). The system architecture is shown in **Figure 1.3**. First, the ASR module recognizes the speech input of the user utterance by

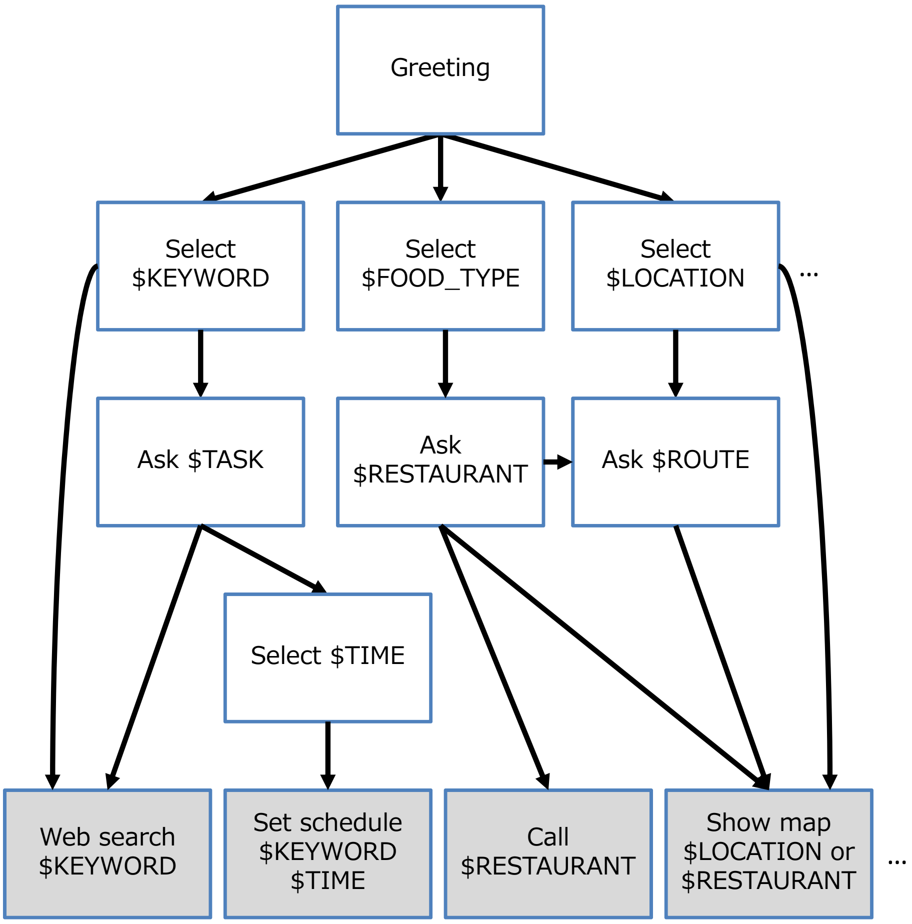


Figure 1.2: *An example of task structure in multi-task navigation system.*

using a language model that matches to the task and domain of the system. The recognition result is interpreted by the SLU module by referring to the SLU result and the knowledge base of the system such as RDB and NE classes, and the DM module controls what to say by referring to the knowledge base. The NLG module generates a response by using the knowledge base and the result of DM, and the TTS module synthesizes the speech output to the user. These modules share the task flow and domain knowledge such as vocabulary in the domain and semantic slots or named entity classes.

Figure 1.4 depicts one example dialogue session of the system that is defined in Figure 1.2. In this example, the user wanted to eat something, but the system could not answer an appropriate response such as suggesting some restaurants because the user query was very ambiguous. In the second utterance, the user gave a clarified demand to the system,

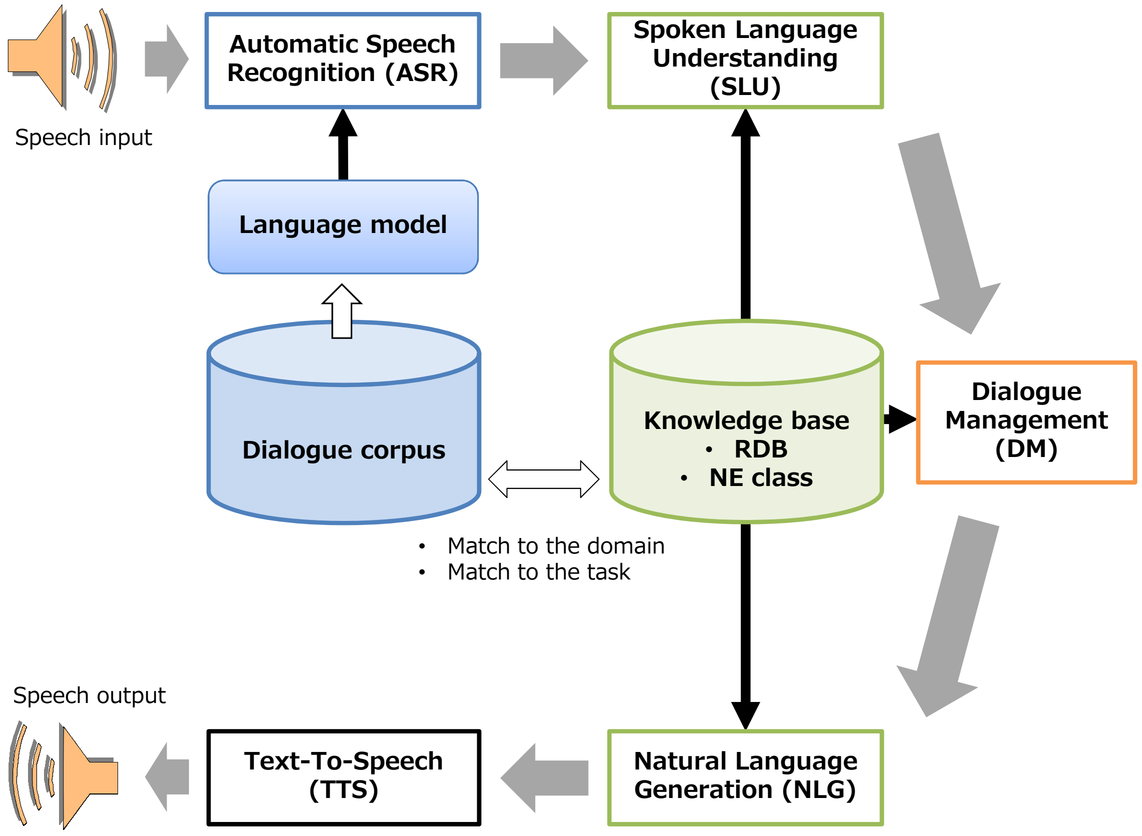
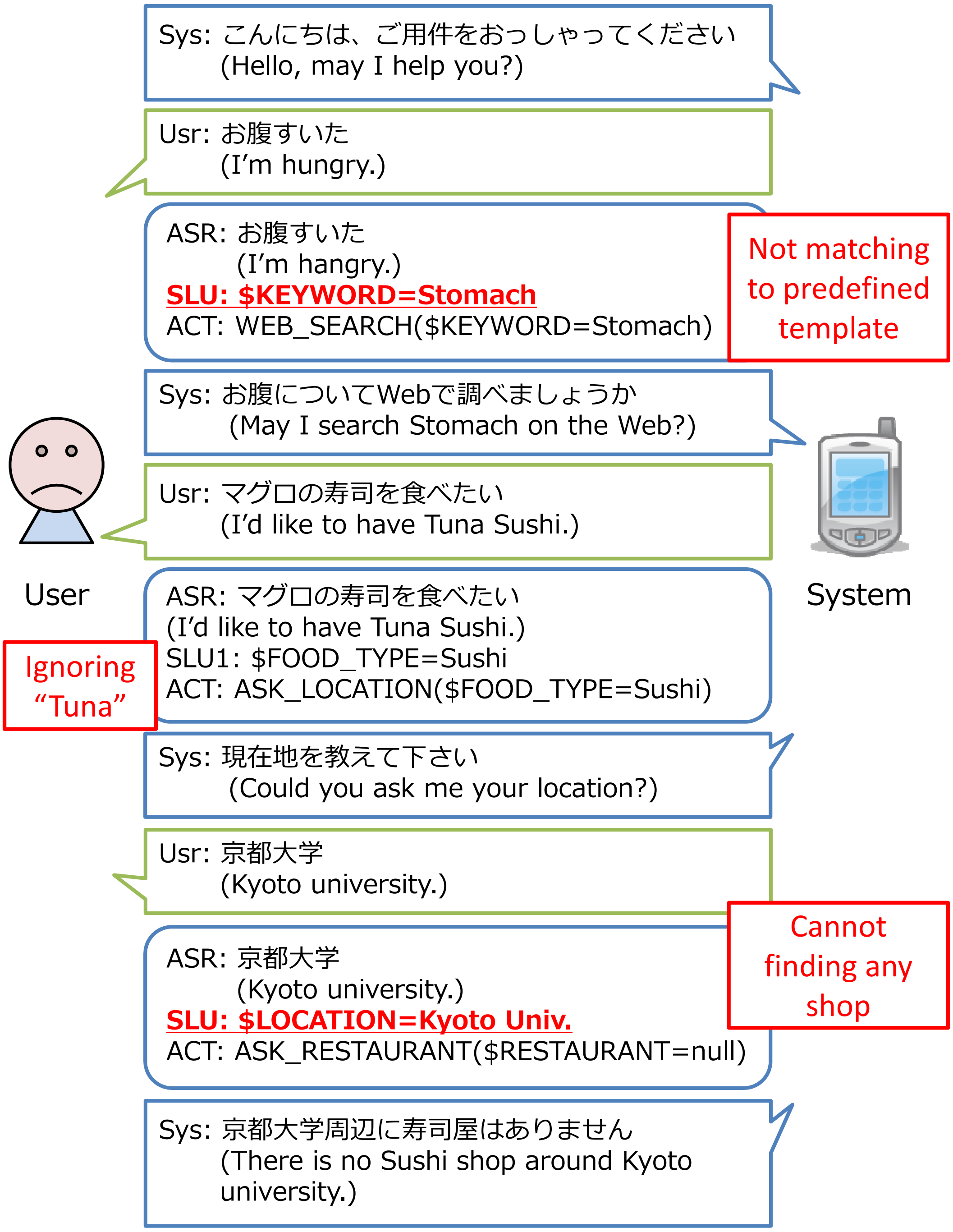


Figure 1.3: *Architecture of conventional spoken dialogue systems.*

but the system ignored an element of information (=Tuna) because the utterance is beyond the capability of the understanding module of the system. The system requested the location of the user for completing the restaurant navigation task, and the user answered the request. However, the system could not find any restaurant that matched to the request and responded with an inattentive answer.

1.2 Problems

The thesis addresses two problems inherited in the conventional spoken dialogue systems. First, dialogue systems traditionally assume that the user has a clear goal of dialogue, and the system is designed to reach the expected goal as soon as possible. However, this assumption does not always hold. Second, well-organized domain knowledge and task flow are required to control the system. They are usually hand-crafted, and it is costly to prepare



Sys: こんにちは、ご用件をおっしゃってください
(Hello, may I help you?)

Usr: お腹すいた
(I'm hungry.)

ASR: お腹すいた
(I'm hangry.)
SLU: $KEYWORD=Stomach
ACT: WEB_SEARCH($KEYWORD=Stomach)

Not matching to predefined template

Sys: お腹についてWebで調べましょうか
(May I search Stomach on the Web?)

Usr: マグロの寿司を食べたい
(I'd like to have Tuna Sushi.)

ASR: マグロの寿司を食べたい
(I'd like to have Tuna Sushi.)
SLU1: $FOOD_TYPE=Sushi
ACT: ASK_LOCATION($FOOD_TYPE=Sushi)

Ignoring "Tuna"

Sys: 現在地を教えて下さい
(Could you ask me your location?)

Usr: 京都大学
(Kyoto university.)

ASR: 京都大学
(Kyoto university.)
SLU: $LOCATION=Kyoto Univ.
ACT: ASK_RESTAURANT($RESTAURANT=null)

Cannot finding any shop

Sys: 京都大学周辺に寿司屋はありません
(There is no Sushi shop around Kyoto university.)

Figure 1.4: *An example of spoken dialogue system in navigation system.*

them for a variety of tasks and domains.

1.2.1 Ambiguous Task Goal and User Query

The conventional dialogue systems require a clear task goal of the user. It can accept the user queries that accord to the designed task and goal. However, this assumption does not always hold as the user may not have a clear goal or the user cannot clarify his or her potential demand to the system. Users are embarrassed by the inflexible management and confused if they do not know how to use the system or how to express their demand in the manner that the system expected. For example, users often do not have any specified preference of a restaurant in the navigation task. In the example of Figure 1.4, the user said to the system "I'm hungry", but the system could not find any information for the ambiguous query. In this case, the system should clarify the user demand by presenting some candidates of the restaurants. The user may have some potential demand even if he does not express clearly their demands. The system should clarify the potential information demand through an interaction and fulfill the potential user demand even if the user query is ambiguous. Flexible matching of information and dialogue management of non-goal-oriented dialogue are needed to realize such kind of information navigation.

There is not a clear principle nor established methodology to design and implement such casual conversation systems. Therefore, an empirical data-driven approach is desirable. WikiTalk (Wilcock, 2012; Wilcock and Jokinen, 2013) is a dialogue system that talks about topics in Wikipedia. This system works on the pre-defined scenario that is represented with an automaton, but it forces users to follow the system scenario. Moreover, developers need to implement a new scenario for a new domain or task. A data-driven approach based on phrase-based statistical machine translation (SMT) (Ritter et al., 2011) tries to train response generation from micro-blog data. This approach enables the system to output a variety of responses, which is expected to include information the user want to know, but it does not track any user intention or dialogue state to fulfill what the user really wants to know.

1.2.2 Hand-crafted Domain Knowledge and Task Flow

Deep semantic knowledge sources of the domain are essential for designing the dialogue system. They are conventionally hand-crafted, but it is costly to prepare and tune them. Flexible SLU of the user query is also achieved by using the domain knowledge. Recently, the template-based approach is widely adopted (Grishman, 2003; Ramshaw and Weischedel, 2005) to extract such domain knowledge, but it costs developers to construct the seed templates dedicated to the domain.

The conventional DM also assumes a well-defined domain knowledge and task flow to control the dialogue in an efficient way, but it costs the developer to construct them. Moreover, the traditional spoken dialogue systems do not distinguish the task and the domain, and the developers need to construct a new task-domain structure if either of the task or domain has been changed. Recently, machine learning of dialogue flow has been intensively studied, but its application is limited and it requires a large amount of annotated training data.

The ASR module also need the domain information. Traditionally, the LM of the ASR module for a spoken dialogue system requires a corpus that matches to the task domain of the system. The LM can be trained by using the Web texts, but we still need to select texts relevant to the domain.

1.3 Approaches in this Thesis

A new spoken dialogue system framework that tackles the problems described above is proposed in this thesis. A new task of spoken dialogue system is defined, information navigation. The information navigation does not assume a clear goal and tries to answer ambiguous user queries, while it assumes a domain such as sports and travel. The domain knowledge is automatically extracted from a domain corpus.

1.3.1 Information Navigation that Handles Ambiguous User Queries

In human-human dialogue, people usually have topics they plan to talk about, and they progress the dialogue in accordance with the topics (Schegloff and Sacks, 1973). Dialogue participants have a role of speaker and listener, and they converse with each other by changing their role of speaker and listener. The proposed system realizes the situated information navigation by taking a role of the speaker who provides information to the listener.

An example is shown in **Figure 1.5**. First, the speaker offers a new topic and probes the interest of the listener. If the listener shows interest, the speaker describes details of the

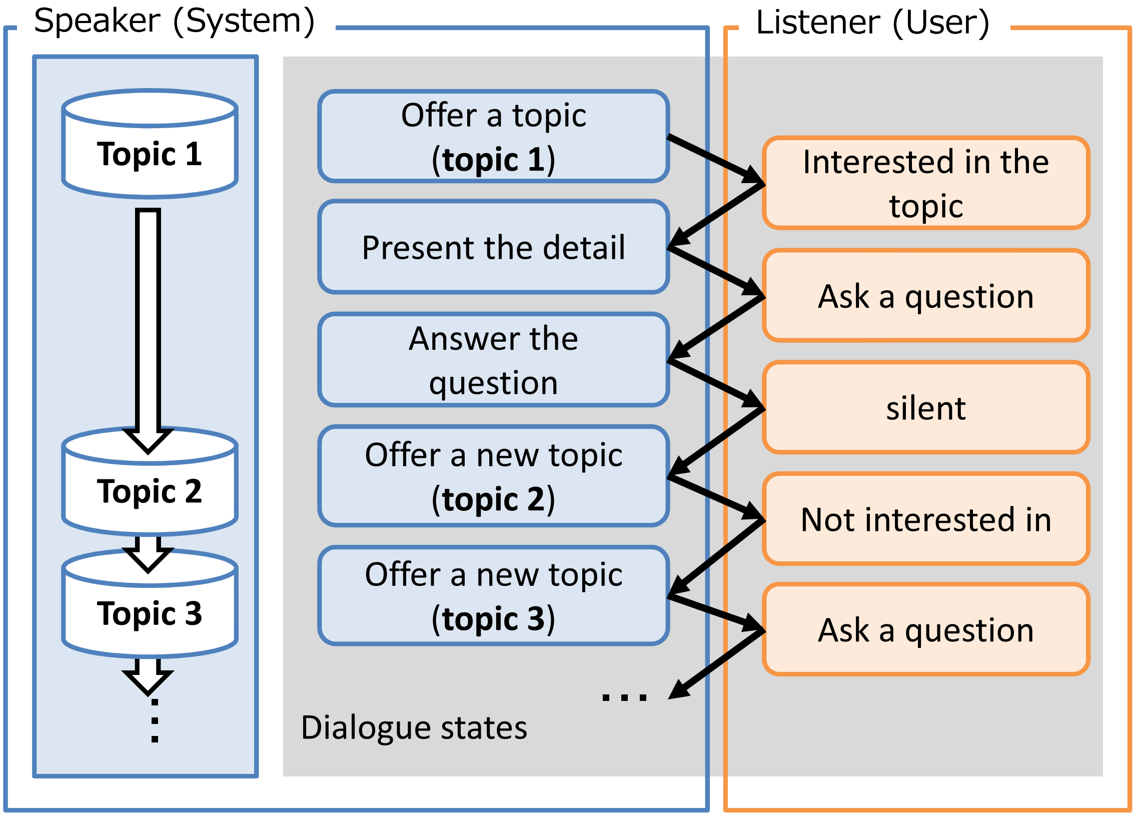


Figure 1.5: *An example of information navigation.*

topic. If the listener asks a specific question, the speaker answers it. On the other hand, if the listener is not interested in the topic, the speaker avoids the details of that topic and changes the topic.

The task of information navigation is designed as a non-goal-oriented dialogue task according to the above-described dialogue manner. When the user demands are not always clear, the information navigation system clarifies the user demands through interactions. The system presents relevant information even if the user request is not necessarily clear and there is no exactly matching result to the user query. Moreover, the system can occasionally present potentially useful information without any explicit request by following the dialogue context. The aim of dialogue is to fulfill information demand of the user through an interaction.

The task design of information navigation is defined as a selection of information navigation modules. The initiative of dialogue comes and goes between the system and the

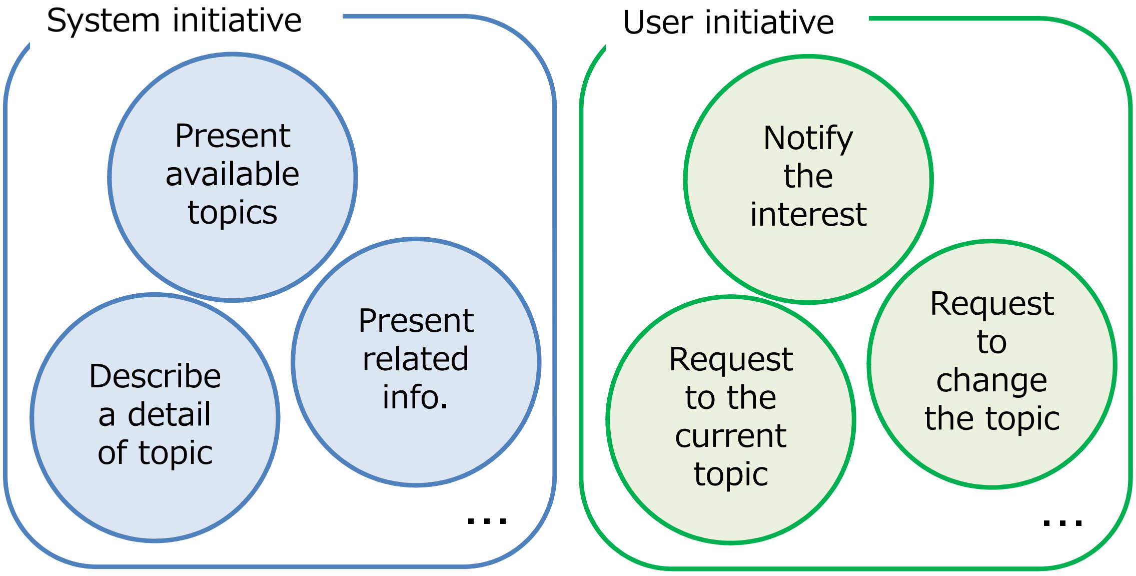


Figure 1.6: *An example of information navigation modules.*

user because it depends on the specification of the user demand. If the user has a specific demand, the user can ask an exact question that matches to his demand. When the user demand is not clear, the system should take an initiative to clarify the user demand by showing candidates that is related to the ambiguous query of the user. Such capability is achieved by modules that refer to the domain knowledge, the user intention and the user focus.

In an information navigation, the system presents topics that it can talk about, describes the detail of the current topic, or presents related topic to the dialogue history when the system has an initiative. In contrast, the system answers the question of the user, replies to the information demand of the user, or receives a request of changing the topic. The function of the system modules depends on the kind of information navigation. An example of information navigation modules is shown in **Figure 1.6**. The details and defined modules in this study are presented in Section 1.4.

1.3.2 Statistical Learning of Domain Knowledge

In the proposed framework, the task design and the domain knowledge are separated. The task of information navigation is designed independently from the domain, and the domain knowledge is automatically extracted from a domain corpus via statistical learning of semantic and dialogue structure. For the information navigation task, important information structure is dependent on the domain. Such information structure is important for improving LM for the ASR module, the SLU and NLG modules including flexible information retrieval. Conventionally, the templates of domain-dependent information structures were hand-crafted, but this heuristic process is so costly, thus it cannot be applied to a variety of domains.

This thesis focuses on the predicate-argument (P-A) structure, a semantic structure in a sentence, generated by a parser. A significance score of domain-dependent P-A structure is statistically defined to extract a set of useful P-A structures for information navigation.

In the proposed framework, dialogue structure is also statistically learned. The learning is conducted on the pre-defined structure of information navigation by referring to the user intention and the user focus. The user intention is a pre-defined dialogue act, which is detected with a classifier. The user focus is an attentional state of the users, which is used for improving the information navigation. It is used to select information the user potentially wants to know. This study designs and develops a news navigation system that uses Web news articles as a knowledge source and presents information based on the users' preference and queries.

The new architecture of a spoken dialogue system is depicted in **Figure 1.7**. In the framework, domain texts are collected from news articles and wisdom of crowds on the Web. The ASR module recognizes the speech input by using the LM that is automatically constructed from the domain texts. The domain knowledge is used to select the appropriate texts for training of LM. The statistical SLU module detects the user intention and user focus with a discriminative model, which is based on the extracted domain knowledge and defined task design of information navigation. The DM module controls dialogue modules by using its belief update of user states (the user intention and the user focus). These dialogue modules are designed by following the design of the task of information navigation. The NLG module generates the most appropriate system response that is determined by DM module, and the TTS module synthesizes the speech output.

Training data collection approaches from the Web are widely applied to the construction of LM of ASR. A new text selection method is proposed to construct the LM that matches not only in the surface word level, but also in the deep semantic structure of expected user

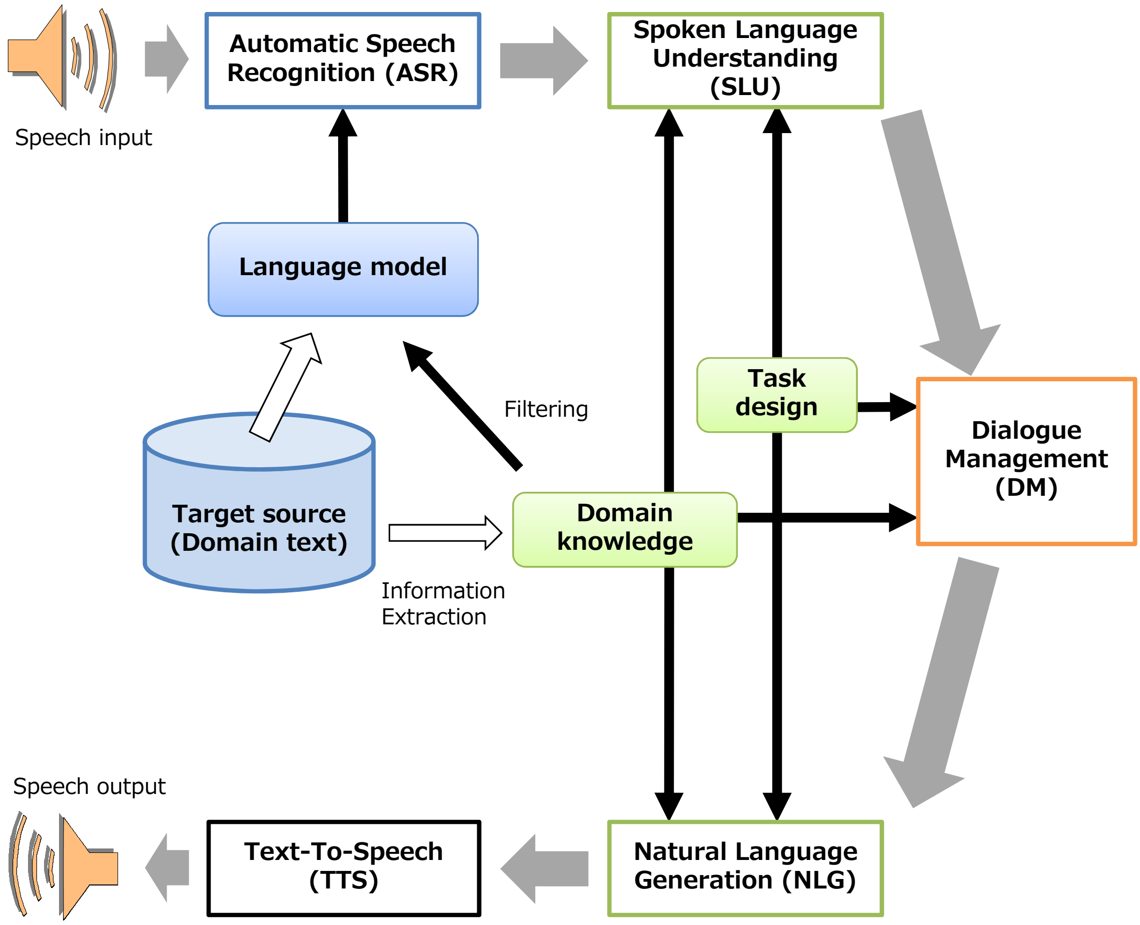


Figure 1.7: *Architecture of the proposed spoken dialogue system.*

query to the system. The system filters the training text for the LM of ASR by considering the domain knowledge. As a result, the dialogue system can work with the ASR system well-matched to the domain. The LM of the ASR module is automatically adapted to the domain.

A flexible question answering module for the information navigation task is realized to handle ambiguous user queries. This module always responds with related information to the user query, even if the user query includes some ambiguity. A proactive presentation module is also implemented to present information that the user is potentially interested in, by selecting information that is related to the dialogue history. The SLU and DM modules incorporate the automatically constructed domain knowledge and output the information (=sentence in the news article) to the NLG module.

1.4 Overview of the System

In this thesis, information navigation is realized by a task of news navigation. In news navigation, the system has a back-end knowledge source described in text and fulfills user information demands through an interaction by referring to the knowledge source. This section presents the overview of the proposed information navigation system in the news navigation task.

1.4.1 News Navigation Task

The news navigation task assumes that the system has a large number of news articles as a back-end knowledge source. The system describes what happened on the day that is written in the articles, and the user can know about the articles through an interaction. The task of news navigation breaks down the task of information navigation into a simpler way. The knowledge source of the system is limited to the news articles, but the articles are updated day by day. The system navigates such dynamic content by parsing the articles and extracting information from huge back-end knowledge source. Moreover, it incorporates a tag of the domain in news articles to extract the domain knowledge from the text source.

The news navigation system is designed based on the dialogue structure of information navigation depicted on Figure 1.5. The system provides topics collected from Web news articles, and the user receives information according to his interests and queries.

1.4.2 System Modules

An overview of the proposed system is illustrated in **Figure 1.8**. The system has seven modules, each of which implements a different dialogue acts. Each module takes as input a recognized user utterance, an analyzed predicate-argument (P-A) structure, and the detected user focus.

The system begins a dialogues with the “topic presentation (TP)” module, which presents a new topic selected from news articles. It chooses the next module based on the user’s response. In this work, it is assumed that each news article corresponds to a single topic, and the system presents a headline of the news in the TP module. If the user shows interest (positive response) in the topic without any specific questions, the system selects the “story telling (ST)” module to give details of the news. In the ST module, the

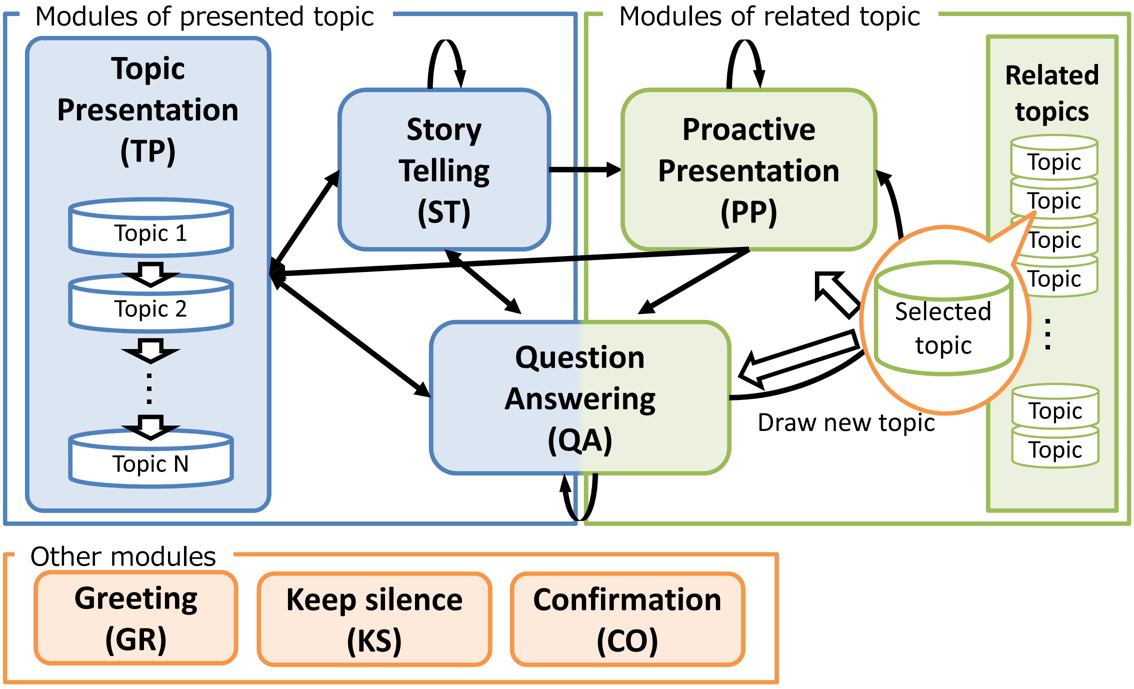


Figure 1.8: *Overview of the information navigation system.*

system provides a summary of the news article by using lead sentences. The system can also provide related topics with the "proactive presentation (PP)" module. This module is invoked by the system's initiative; this module is not invoked by any user request. If the user asks a specific question regarding the topic, the system switches to the "question answering (QA)" module to answer the question. This module deals with questions on the presented topic and related topics.

The modules of PP and QA are based on a dialogue framework which uses the similarity of the P-A structure between user queries and news articles, and retrieves or recommends the appropriate sentence from the news articles. This method searches for appropriate information from automatically parsed documents by referring to domain knowledge that is automatically extracted from a domain corpus. The details are described in Chapter 4.

Transitions between the modules are allowed as shown in Figure 1.8. The modules "greeting (GR)", "keep silence (KS)" and "confirmation (CO)" are also prepared. The GR module generates fixed greeting patterns by using regular expression matching. The CO module makes a confirmation if the system does not have certainty about the user query.

In terms of dialogue flow, these modules can be called at any time.

An expected example of the information navigation is shown in **Figure 1.9**. In this example, the system presents news of the day. The user is interested in the topic and makes a question about the news. The ASR module makes some errors, but the flexible matching of QA module retrieves a related information and presented it. The system detected that the user is interested in the current topic, thus, the system proactively presents information that is related to the dialogue history with PP module even if the user does not speak anything. When a new question of the user is invoked by the system presentation, and the system answers the question again. Even if the system cannot find the information that is exactly matched to the user query, the system presents relevant information based on partial matching of the QA module.

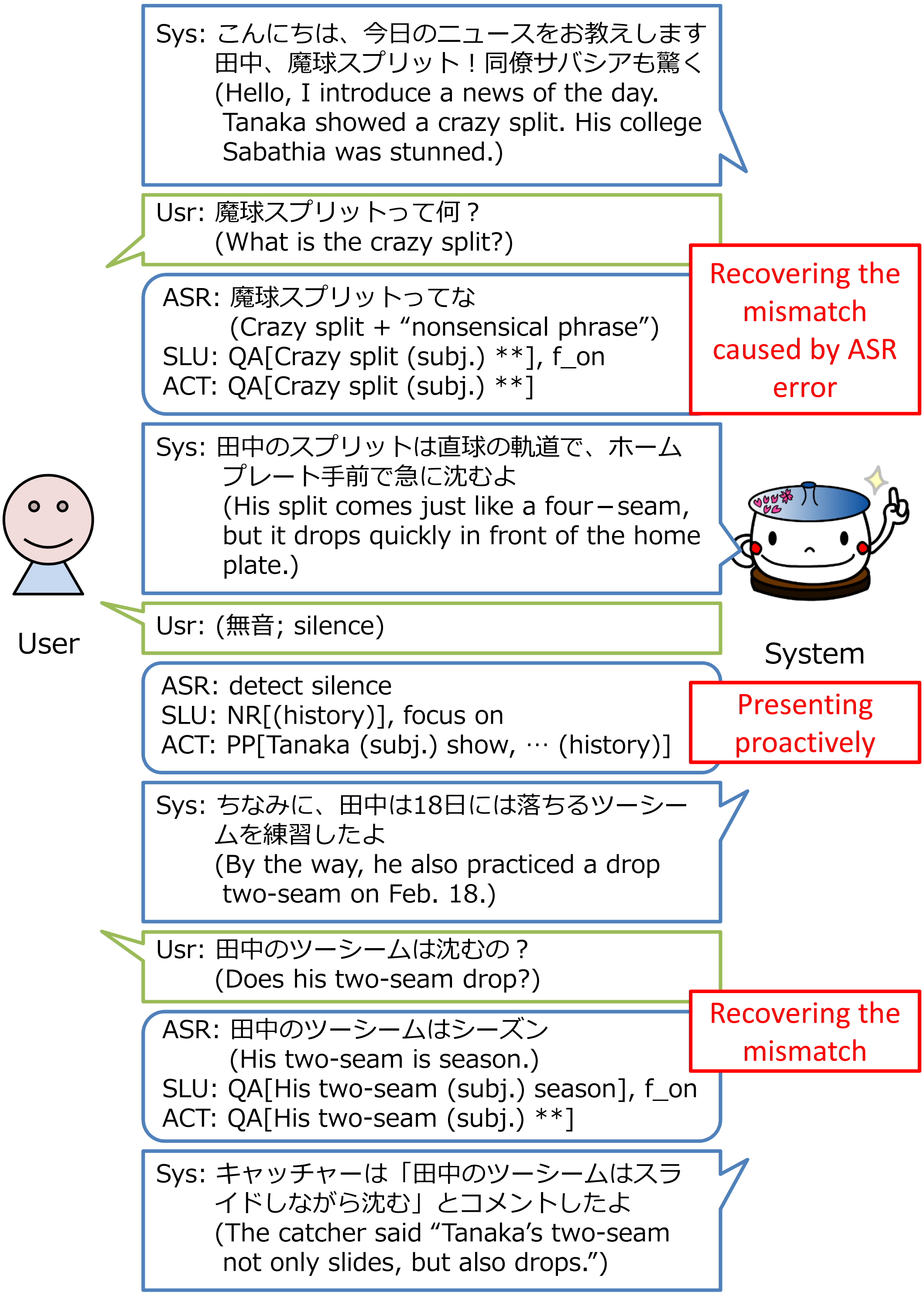
The proposed scheme enables the system to answer not only clear requests, but also ambiguous requests that do not have any specified goal. The system can respond with flexible matching between the user query and the back-end knowledge source by using the statistical learning result of the semantic P-A structure. As a result, the system has a capability to answer not only factoid questions, but also non-factoid questions such as "How was today's Ichiro?" or "How do you feel about the all-star game?". By responding to these questions with some specified news such as "Ichiro hit a home-run" or "28 members are selected for the all-star game", the user can know the outline of the news that he may be interested in, and some more specific questions are invoked.

The dialogue is generated based on the news articles in the knowledge source texts. All modules of the system are automatically trained from the knowledge source, and they are easily portable to different domains.

1.5 Outline of the Thesis

Figure 1.10 shows an outline of the remaining chapter of the thesis corresponding to key components of the information navigation system.

Unsupervised information extraction framework is proposed based on the automatically extracted P-A structure, which is the baisis of the system. The information extraction is realized without annotation of domain knowledge. Chapter 2 addresses the domain knowledge extraction.



Sys: こんにちは、今日のニュースをお教えします
田中、魔球スプリット！同僚サバシアも驚く
(Hello, I introduce a news of the day.
Tanaka showed a crazy split. His college
Sabathia was stunned.)

Usr: 魔球スプリットって何？
(What is the crazy split?)

ASR: 魔球スプリットってな
(Crazy split + "nonsensical phrase")
SLU: QA[Crazy split (subj.) **], f_on
ACT: QA[Crazy split (subj.) **]

Sys: 田中のスプリットは直球の軌道で、ホーム
プレート手前で急に沈むよ
(His split comes just like a four−seam,
but it drops quickly in front of the home
plate.)

Usr: (無音; silence)

ASR: detect silence
SLU: NR[(history)], focus on
ACT: PP[Tanaka (subj.) show, ⋯ (history)]

Sys: ちなみに、田中は18日には落ちるツーシー
ムを練習したよ
(By the way, he also practiced a drop
two-seam on Feb. 18.)

Usr: 田中のツーシームは沈むの？
(Does his two-seam drop?)

ASR: 田中のツーシームはシーズン
(His two-seam is season.)
SLU: QA[His two-seam (subj.) season], f_on
ACT: QA[His two-seam (subj.) **]

Sys: キャッチャーは「田中のツーシームはスラ
イドしながら沈む」とコメントしたよ
(The catcher said "Tanaka's two-seam
not only slides, but also drops.")

Figure 1.9: *A typical example of the proposed news navigation system.*

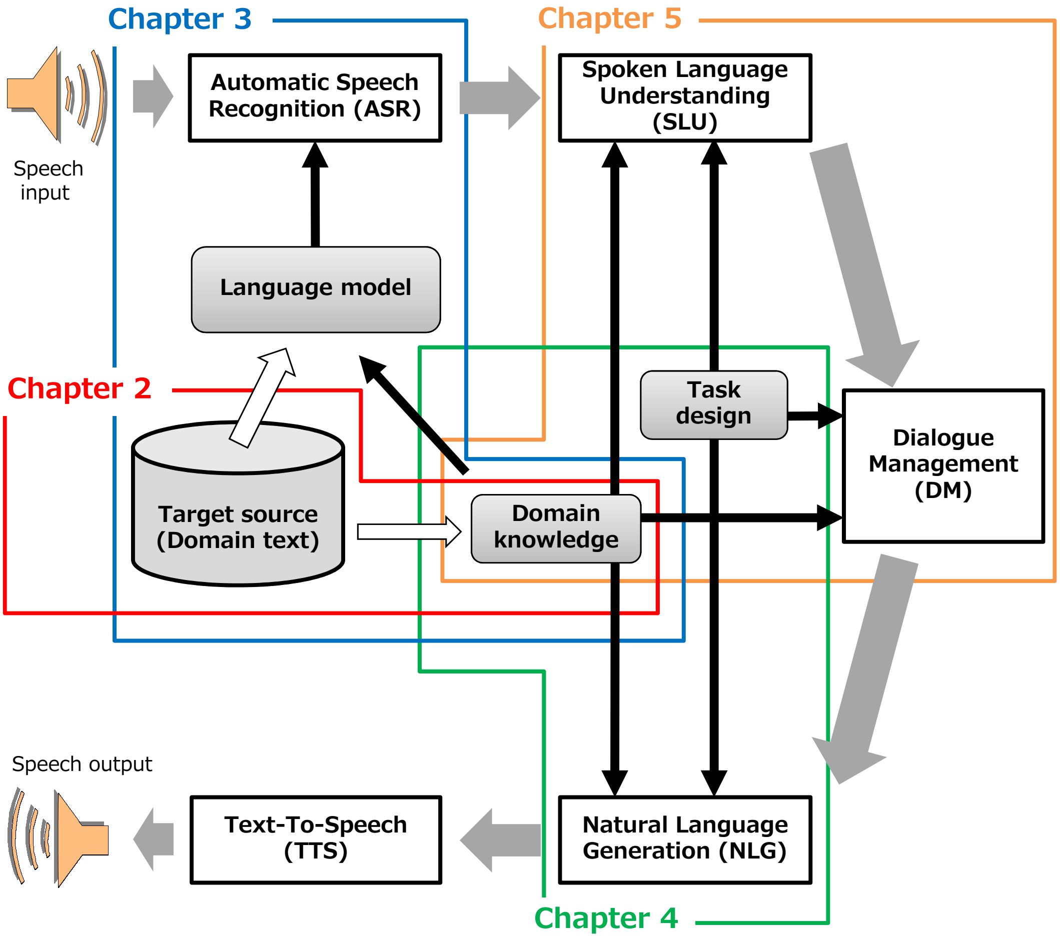


Figure 1.10: *An outline of the remaining chapter of the thesis.*

The system enhances the module of LM for ASR by selecting a training set of sentences from Web texts. This selection method is based on the automatically extracted semantic information (statistical measure of the P-A structure for a domain). The method improves not only word-based ASR accuracy, but also semantic-level accuracy. Chapter 3 describes this LM construction.

Dialogue modules based on the P-A structure are proposed. The proposed question answering module finds information that flexibly matches the user queries. If the module cannot find the exact information to the user query, the module extends the user query by relaxing elements of the P-A structure by using the relevance measure of the P-A structure and the significance score of the P-A structure. The proactive presentation module presents information by referring to dialogue history when the user does not utter. The module activates the potential information demand of the user. The modules are described in Chapter 4.

The DM of dialogue modules are controlled based on partially observable Markov decision process (POMDP) that tracks the user focus. Rewards for POMDP are defined by the appropriateness in the interaction, compared to the general POMDP that requires a clear task goal to define rewards. The proposed dialogue manager tracks the user focus to provide information that the user potentially demands. The new dialogue management is described in Chapter 5.

The proposed framework is designed to realize portability across domains. The modules are automatically constructed from the domain text that is collected from the Web, and they are designed independently of the domain. This thesis is concluded in Chapter 6.

Chapter 2

Statistical Learning of Domain-Dependent Semantic Structure

This chapter introduces a statistical learning method of domain knowledge based on semantic structure of the domain corpus, which plays an important role in the proposed system. The domain knowledge is based on predicate-argument (P-A) structure, which is one of the most fundamental information structures in a natural language text. The userful information structure depends on the domain. In order to automatically extract useful domain-dependent P-A structure, a statistical measure is introduced, resulting in a completely unsupervised learning of semantic information structure given a domain corpus.

2.1 Semantic Information Structure based on Predicate-Argument Structure

The P-A structure is used to define the domain-dependent semantic structure. P-A structure is generated by a parser as a baseline, but every P-A structure is not useful for information extraction and retrieval (Kiyota et al., 2002; Dzikovska et al., 2003; Harabagiu et al., 2005). Conventionally, the templates for information extraction were hand-crafted (Grishman, 2003; Ramshaw and Weischedel, 2005), but this heuristic process is so costly that it cannot be applied to a variety of domains on the Web. In this chapter, two scoring methods are prescribed to extract domain-dependent useful information patterns.

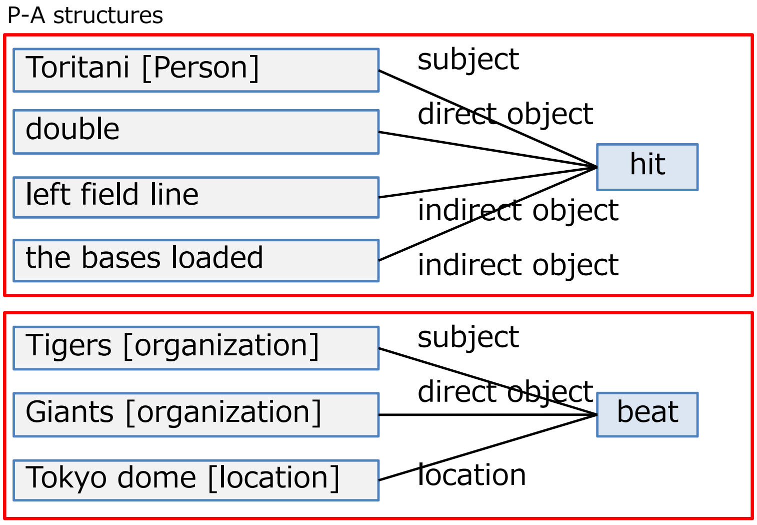


Figure 2.1: *An example of predicate-argument structures.*

2.1.1 Predicate-Argument Structures

The P-A structure represents a relationship of the semantic role between a predicate, a verb or an event noun, and arguments, which depends on the predicate (Fillmore, 1968). An example of the P-A structure is shown in **Figure 2.1**. There are some required semantic roles depending on the type of the predicate (verb or event-noun), and also arbitrary semantic roles like time, place, and other modifications. This structure is a classic concept in natural language processing, but recently automatic semantic parsing has reached a practical level thanks to corpus-based learning techniques (Kawahara and Kurohashi, 2006) and has been used for several large-scale tasks (Shen and Lapata, 2007; Wang and Zhang, 2009; Wu and Fung, 2009). We use KNP[1] (Kawahara and Kurohashi, 2006) as a syntactic and semantic parser.

2.2 Extraction of Domain-dependent P-A Patterns

The P-A structure automatically generated by the semantic parser provides useful information structure as a baseline. However, every P-A pair is not meaningful in information navigation; actually, only a fraction of the patterns is useful, and they are domain-dependent. For example, in the baseball domain, key patterns include “[A (subject) beat B (object)]”

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

and “[A (subject) hit B (object)]”, and in the business domain, “[A (subject) sell B (object)]” and “[A (subject) acquire B (object)]”. A method is proposed to automatically extract these useful domain-dependent patterns given a domain corpus. We assume each article in the newspaper corpus/websites is annotated with a domain such as sports-baseball and economy-stock.

The method is to filter P-A structure patterns (=templates) based on statistical measure (=score) which accounts for the domain. The filtering process is also expected to eliminate inappropriate patterns caused by parsing errors. Moreover, in spoken dialogue systems, errors in automatic speech recognition (ASR) may result in erroneous matching. By eliminating irrelevant patterns, we expect robust information extraction for spoken input. Specifically, the following two significance measures are investigated.

2.2.1 Significance Score based on TF-IDF Measure

First, we use the TF-IDF measure to evaluate the importance of word w_i in a particular document set of domain D.

$$\mathbf{TFIDF}(w_i) = \underbrace{P(w_i)}_{\mathbf{TF}} \log \underbrace{\frac{C(d)}{C(d: w_i \in d)}}_{\mathbf{IDF}}. \tag{2.1}$$

The TF term is the occurrence probability of word w_i, defined as:

$$\mathbf{TF} = P(w_i) \approx \frac{C(w_i) + \alpha}{\sum_j \big(C(w_j) + \alpha\big)}, \tag{2.2}$$

where $C(w_i)$ is the occurrence count of word w_i in the domain D in the corpus, and α is a smoothing factor of Dirichlet smoothing. The IDF term is the inverse document frequency that contains word w_i:

$$\mathbf{IDF} = \frac{C(d)}{C(d: w_i \in d)} \approx \frac{C(d) + \beta}{C(d: w_i \in d) + \beta}, \tag{2.3}$$

where $C(d)$ is the number of documents (=newspaper articles) in the corpus and $C(d: w_i \in d)$ is the number of documents which contain word w_i. β is a smoothing factor. The **IDF** term is not probability, but it needs smoothing to avoid the problem of zero division. In this work, α=1 and β=1.

2.2.2 Significance Score based on Naive Bayes (NB) Model

The second measure is based on the Naive Bayes model.

$$P(D|w_i) = \frac{C(w_i, D) + P(D)\gamma}{C(w_i) + \gamma}. \tag{2.4}$$

Here, $C(w_i, D)$ is a count of word w_i that appears in domain D, γ is a smoothing factor with the Dirichlet process prior (the detail is shown in the Appendix) and $P(D)$ is a normalization coefficient of the corpus size of the domain D.

$$P(D) = \frac{\sum_j C(w_j, D)}{\sum_k C(w_k)}. \tag{2.5}$$

The evaluation measure for a P-A pattern is obtained by taking a geometric mean of the component words.

2.2.3 Clustering of Named Entities

The statistical learning often falls into the data sparseness problem, especially for named entities (NEs; name of persons, organizations, locations). Moreover, there may be a mismatch in the set of NEs between the training corpus and the test phase. For robust estimation, NE classes are introduced. The example of automatically labeled NEs are shown in the Figure 2.1. Note that unifying all named entities in the corpus before computing the evaluation measure would weaken the significance of these entities. Thus, we compute statistics for every proper noun before clustering, and sum up values for the class afterwards.

Clustering is conducted to classify P-A structures which have the same triplet of predicate, semantic case and NE. The example is shown in **Figure 2.2**, two P-A structures which have the same NE P-A pairs are clustered to the same template. We extend the probability of argument w_a as

$$P(D|\mathrm{NE}_i) \quad = \sum_{k(w_k \in \mathrm{NE}_i)} P(D|w_k)P(w_k). \tag{2.6}$$

2.2.4 Evaluation of P-A Significance Scores

An experimental evaluation is performed to compare the effectiveness of the two significance measures (TF-IDF and Naive Bayes (NB)) in the Japanese professional baseball domain. The models are trained with the Mainichi Newspaper corpus 2000 – 2008. The clustering

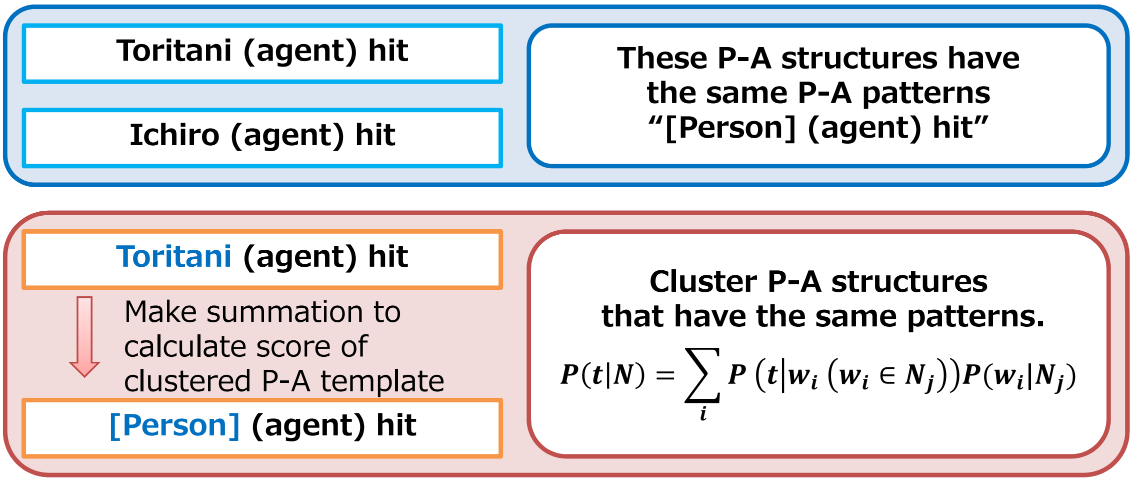

Toritani (agent) hit
Ichiro (agent) hit
These P-A structures have the same P-A patterns "[Person] (agent) hit"
Toritani (agent) hit
Make summation to calculate score of clustered P-A template
[Person] (agent) hit
Cluster P-A structures that have the same patterns.
$P(t|N) = \sum_i P(t|w_i(w_i \in N_j))P(w_i|N_j)$

Figure 2.2: *Clustering of named entities (NEs).*

of NEs is applied to both methods. The P-A templates having larger significance scores are selected. We determined a threshold for selecting templates using a development set which was held out from the test set by 10%. The test set was made from the Mainichi newspaper's website which talks about games played between April 21-23, 2010. Manual annotation was made on typical P-A patterns which can be used for question answering and proactive presentation. The filtering was performed on the test set by matching the patterns defined by each measure, and evaluated against the annotated answers in terms of recall, precision and F-measure (F). **Figure 2.3** lists the result for the baseline and two measures. The baseline is every P-A structure that is outputted by a parser. In PS+A, we calculated a geometric mean of scores of P-A structure elements; the score of the argument and the score of the pairs of the predicate and the semantic role. In PSA, we calculated the score to a triplet of P-A structure elements; argument, predicate and semantic role. **Figure 2.4** shows the precision-recall curve of the two measures.

In this result, the Naive Bayes (NB) model (PS+A) performed the best. The TF-IDF model can also remove some irrelevant patterns, but it is difficult to extract the meaningful patterns for the domain only from a domain text. Compared with the baseline without any filtering, the Naive Bayes methods significantly improved precision with some degradation of recall. This property is important in realizing informative response generation robust against ASR and parsing errors. In comparison between PS+A and PSA in the Naive Bayes

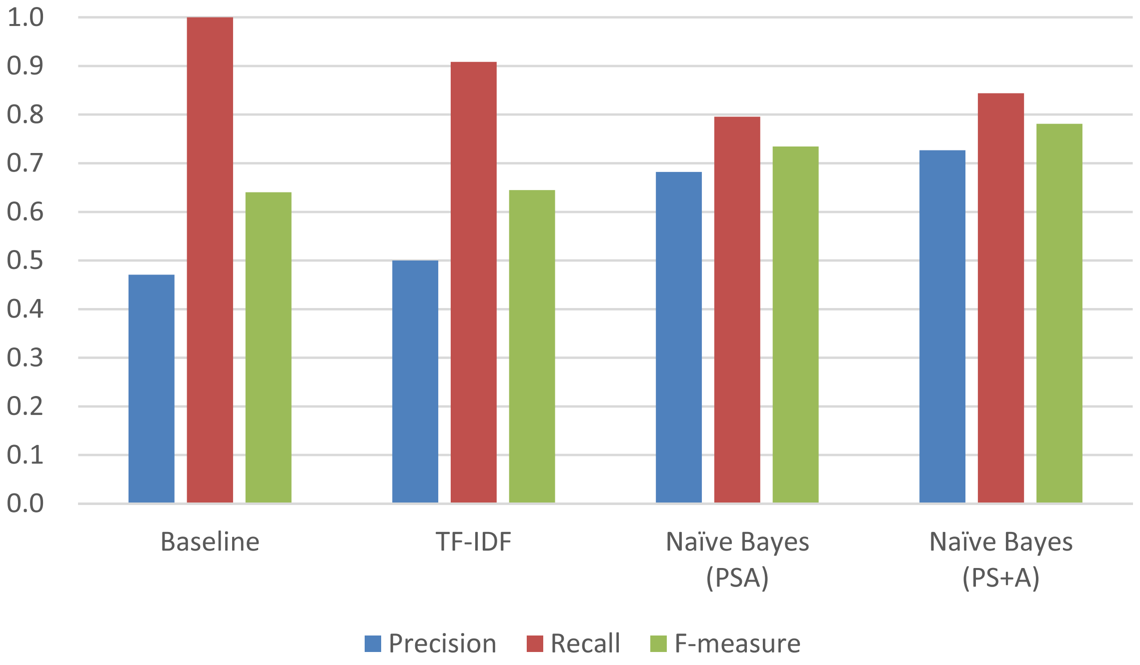


Figure 2.3: *Precision, recall, F-measure of each filtering method.*

model, the PS+A successfully extracted important patterns by mitigating the problem of data-sparseness. Among the selected templates, we can find typical and important patterns like “勝つ (have a win)”, “登板する (come into pitch)”, and “連勝する (make it consecutive wins)”. Most of recall errors are infrequent patterns, and majority of precision errors are those patterns that are frequently observed but not useful for presentation. For example, major precision errors are caused by predicates that are frequent verb in Japanese such as “する (do)” or “なる (become)”. Their verbs have a variety of meanings and appear on any domain document. The most typical recall error is “日本一 (ニ格) 輝く (won the championship)”, it is very important information pattern in baseball domain, but it also appears on other sports domains and it is infrequent in baseball domain, that happens once a year.

2.3 Conclusion

In this chapter, the statistical learning of semantic structures is formulated by defining the significance score of the domain-dependent P-A structure. The score based on the Naive Bayes is introduced to select useful templates in a given domain automatically. The

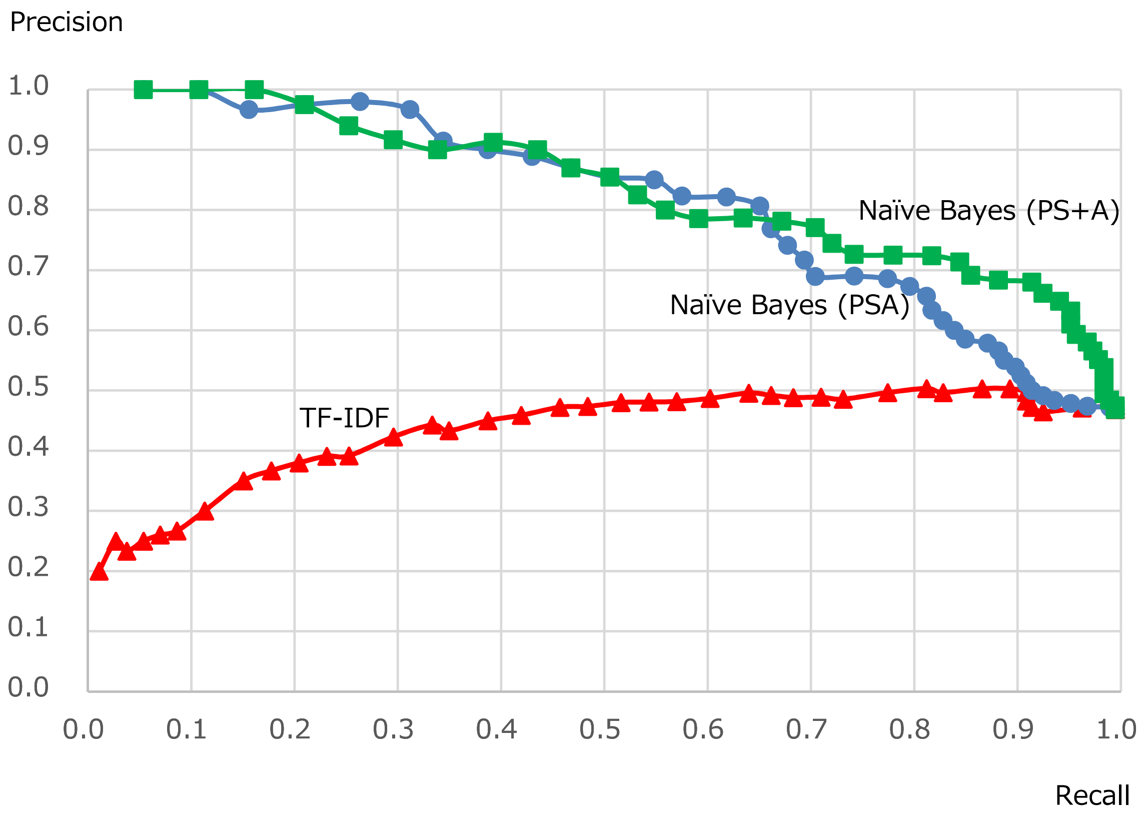


Figure 2.4: *Precision-recall curve of TF-IDF method and Naive Bayes method.*

experimental results show that the high scores are given to important patterns in the domain. The scoring method does not require any annotated data or thesaurus in the domain and it can be applied to a variety of domains. This extracted significance score is used as domain knowledge of the dialogue system through this thesis.

Chapter 3

Incorporating Semantic Similarity to the Training Data Selection of Language Model of Automatic Speech Recognition

This chapter addresses a text selection for training a language model (LM) text for automatic speech recognition (ASR) of spoken dialogue systems. ASR for spoken dialogue systems needs to cover more spoken-style inputs than typical ASR tasks such as dictation or query for Web search. However, it is difficult to collect spoken-style sentences for the exact domain because the back-end document or database of the dialogue systems consist of written-style sentences. In most of previous studies, spoken style sentences are collected from the Web and filtered to make an appropriate training set of LM. Conventionally, word surface-level selection criterion is widely used to measure the relevance between collecting sentences and the back-end knowledge base of dialogue systems. The proposed approach introduces a semantic-level relevance. The domain-dependent P-A structure introduced in the previous chapter is used to filter semantically relevant sentences in the domain. Several choices of the statistical measure and combination methods with the surface-level measure are investigated in this chapter.

3.1 Speech Recognition for Spoken Dialogue Systems

The ASR module for spoken dialogue systems needs an appropriate LM adapted to the task domain and style. Even an ASR system with a very large-vocabulary cannot cover all proper nouns or named entities (NEs), which are critical in information navigation. Ideally, an LM should be trained with a large-scale matched corpus, but in many cases, this is not realistic. Therefore, two approaches are commonly adopted. One approach involves mixing document text of the target domain with a dialogue corpus of spoken style expressions (Komatani et al., 2001). The other involves collecting relevant texts, possibly spoken-style sentences, from the Web (Bulyko et al., 2003). These approaches try to cover the target domain and style of speech in an indirect way, but the resultant model will inevitably contain a large amount of irrelevant texts.

In general, information navigation systems with speech interfaces have a set of back-end documents that the system talk about. The systems require matching between the back-end documents and the user utterances. Based on this assumption, a similarity measure between collected sentences from the Web (=expected user utterances) and the back-end documents is defined by using the significance score of the semantic structure that is proposed in Chapter 2, and select well-matched sentences for LM training.

3.1.1 Language Resources

The system has a back-end document set that will be used for retrieval of the reply to user queries. The documents are not a relational database but natural language text of particular domains. This work primarily uses a set of newspaper articles of the professional baseball domain as described in Chapter 2. Note that the documents are not directly used for LM training, because their style is much different from user queries and the majority of the content is not relevant to the query.

This work turns to a much larger Web resource of Yahoo! QA[1], a Web site of wisdom of crowds, in which people can ask questions according to the domain category. Note that the definition of domain categories does not match that of the spoken dialogue system, and moreover there are many irrelevant queries on the Web site.

3.1.2 Related Works

Use of Web resources for LM training has been investigated as the Web becomes prevailing. In the early years, (Zhu and Rosenfeld, 2001) enhanced trigram statistics with the frequen-

[1] This corpus is provided by Yahoo! JAPAN and National Institute of Informatics.

cies of the trigram in the Web, and (Nishimura et al., 2001) collected texts from the Web by manually specifying keywords in the task domains.

But the most standard approach is to use characteristic N-gram entries as search queries for the Web to collect relevant texts (Bulyko et al., 2003; Sarikaya et al., 2005; Ng et al., 2005; Wan and Hain, 2006; Tsiartas et al., 2010), but this requires a seed corpus to estimate a seed N-gram model. Several works used other resources for generating search queries, such as back-end documents (Misu and Kawahara, 2006), presentation slides in lectures (Munteanu et al., 2007; Kawahara et al., 2008), or initial ASR transcripts (Suzuki et al., 2006).

Selection of the collected Web texts has also been investigated. The majority of the previous studies adopted the perplexity measure by the seed LM (Bulyko et al., 2003; Bulyko et al., 2007; Ng et al., 2005; Misu and Kawahara, 2006), or its variants such as BLEU score (Sarikaya et al., 2005) and normalization by the back-end topic model (Sethy et al., 2005) or the self model (Moore and Lewis, 2010). (Masumura et al., 2011) introduced a Naive Bayes classifier for selecting spoken-style texts. But all the previous works do not consider semantic-level information.

An exception is the work by (Akbacak et al., 2005), which defined characteristic noun phrase and verbs to filter Web texts. However, their method was largely heuristic and did not define a statistical semantic relevance measure. This work is different in that the P-A structure is introduced to define a semantic relevance measure, which is used for selection of large-scale Web texts.

3.2 Selection based on Surface-level Relevance Measure on N-gram

This section defines the surface-level relevance measure on N-gram based on KL-divergence (Kullback and Leibler, 1951) to compare a sentence for training q and the back-end document D. KL-divergence is a non-symmetric distance measure of two probability distributions. The KL-divergence between a sentence q and the document set D is defined as,

$$KL(q||D) = \sum_i P_q(w_i) \log_2 \frac{P_q(w_i)}{P_D(w_i)}, \tag{3.1}$$

where w_i is a word in the sentence q, P_D and P_q is a probability of language model that is calculated from D and q based on N-gram (3-gram) model. If the N-grams in the sentence q is unique, any N-gram chain does not appear twice or more in q, and the probability of word w_i in the distribution of P_q, $P_q(w_i)$ becomes 1. This condition is true in most situations when we calculate the probabilities with 3-gram model because the q is a single sentence. Thus, the Eqn. (3.1) is approximated as,

$$
\begin{aligned}
KL(q||D) &\approx \sum_i \log_2 \frac{1}{P_D(w_i)} &(3.2)\\
&= -\sum_i \log_2 P_D(w_i). &(3.3)
\end{aligned}
$$

Previously, many studies have been conducted on selection of Web texts for LM training, but the majority of them adopt the perplexity criterion or its variants for selection. Many works assume a seed corpus to prepare a seed LM for generating a Web search query or computing perplexity. The test-set perplexity $PP(q,D)$ is defined as,

$$
\begin{aligned}
H(q,D) &= -\frac{1}{n}\sum_{i=1}^{n} \log_2 P_D(w_i). &(3.4)\\
PP(q,D) &= 2^{H(q,D)}. &(3.5)
\end{aligned}
$$

Here, $H(q,D)$ is entropy of sentence q given the back-end document D. From Eqn. (3.3) and Eqn. (3.5), the perplexity-based approach is equal to use the KL-divergence between the sentence and the back-end document.

3.3 Selection based on Semantic Relevance Measure on P-A Structure

The relevance measure in Section 3.2 evaluates the word surface-level relevance between a sentence q and the back-end document D. However, it is difficult to select sentences that has adequateness to the target back-end document not only on surface level but also on syntactic and semantic level. Semantic information is defined in many SLU modules of dialogue systems. In this chapter, a semantic relevance measure is introduced based on the P-A structure to select sentences that has adequateness to the target back-end document on semantic level. The surface-level relevance measure based on the KL divergence calculates the relevance for all of the words in the sentence. In contrast, the semantic relevance measure defined focuses on the elements of the P-A structure. The proposed relevance measure is based on the significance score based on the Naive Bayes model in Section 2.2.2. It can be explained as a discriminative model compared to the generative model of Section 3.2.

3.3.1 Definition of Semantic Relevance Measure

According to the Section 2.2.2, the relevance measure in semantic level based on the P-A structure. The probability of domain D, which is defined by the back-end document of the dialogue system, given a word w_i is defined as

$$P(D|w_i) = \frac{P(w_i|D) \times P(D)}{P(w_i)}, \tag{3.6}$$

with Bayes theory. It is approximated as

$$P(D|w_i) \simeq \frac{C(w_i, D) + P(D) \times \gamma}{C(w_i) + \gamma}, \tag{3.7}$$

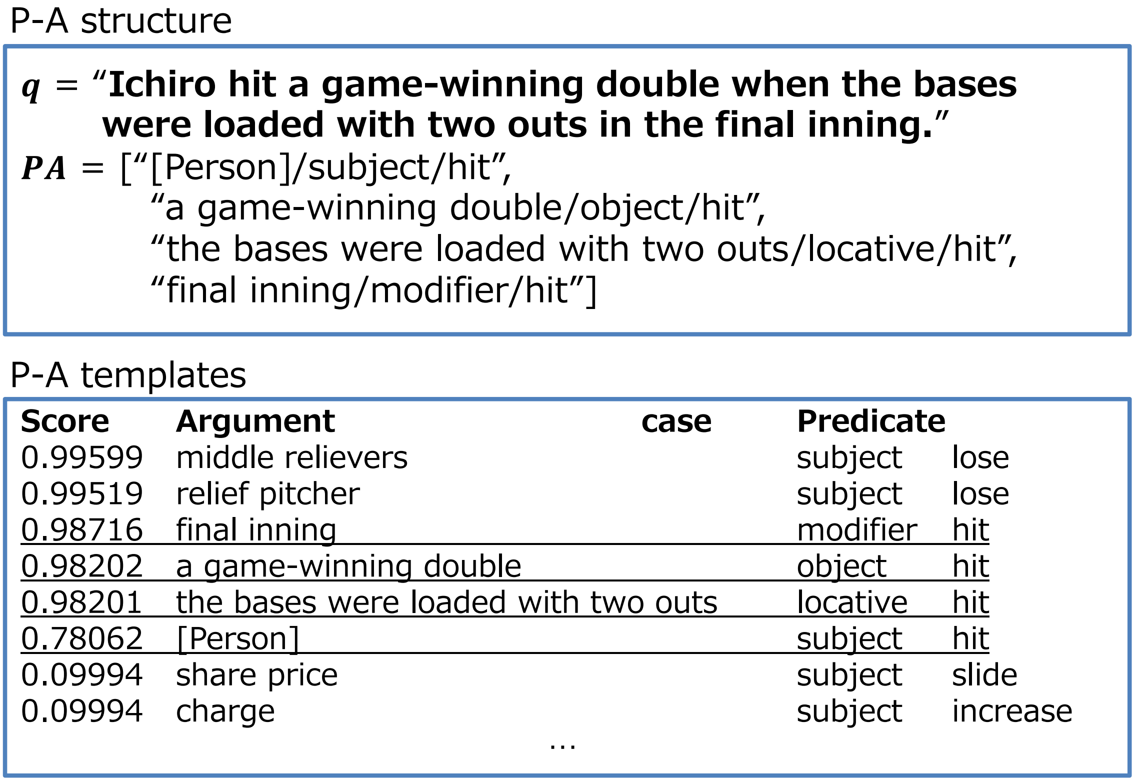
by Dirichlet smoothing based on Chinese restaurant process (the detail is shown in the Appendix). Here, $C(w_i)$ is a count of word w_i and $C(w_i, D)$ is a count of word w_i in the domain (document set) D. $P(D)$ is calculated from the proportion of the back-end document set D and out-of-domain document set $\bar{D}$. The above formula is a variation of unigram probability, but here P-A pairs of predicate $w_{i,p}$ and argument $w_{i,a}$, instead of dealing all words uniformly. The score of a pair of a predicate and an argument $PA(D|pa_i)$ is defined as a geometric mean of $P(D|w_{i,p})$ and $P(D|w_{i,a})$,

$$PA(D|pa_i) = \sqrt{P(D|w_{i,p}) \times P(D|w_{i,a})}. \tag{3.8}$$

Clustering of named entities is also conducted in Section 2.2.3. The relevance score for the P-A pair that has an entity class NE_i is rewritten as,

$$P(D|\mathrm{NE}_i) = \sum_{k(w_k \in \mathrm{NE}_i)} P(D|w_k)P(w_k|\mathrm{NE}_i). \tag{3.9}$$

For each sentence q, we compute a mean of $PA(D|pa_i)$ for P-A pairs included in the sentence, defined as $PA(D|q)$. An example of scoring is shown in **Figure 3.1**. There are four P-A pairs on the input sentence q; “final inning (modifier) hit”, “a game-winning double (object) hit”, “the bases were loaded with two outs (locative) hit”, and “[Person] (subject) hit”. Each P-A pair is scored by $PA(D|pa_i)$. Then, the average of these four scores is computed to defined $PA(D|q)$.



P-A structure

q = “**Ichiro hit a game-winning double when the bases were loaded with two outs in the final inning.**”
PA = [“[Person]/subject/hit”,
“a game-winning double/object/hit”,
“the bases were loaded with two outs/locative/hit”,
“final inning/modifier/hit”]

P-A templates

Score	Argument	case	Predicate
0.99599	middle relievers	subject	lose
0.99519	relief pitcher	subject	lose
0.98716	final inning	modifier	hit
0.98202	a game-winning double	object	hit
0.98201	the bases were loaded with two outs	locative	hit
0.78062	[Person]	subject	hit
0.09994	share price	subject	slide
0.09994	charge	subject	increase
	…		

Figure 3.1: *An example of score of* $PA(D|pa_i)$.

For the training of sentences that have high $PA(D|q)$ are selected as a training data of language model for ASR module. It is expected that the measurement can select sentences that match the user utterances of the dialogue system, and that it improves not only a word surface-level accuracy but also an accuracy of semantic structure extraction.

3.3.2 Combination with Surface-level Measure

Then, combination of the proposed semantic relevance measure with the perplexity measure is investigated, since they presumably model different aspects of the relevance with the target domain. A simple combination method is to use the ranks by the two measures. The sentences can be re-ordered based on the sum of them.

A score-based combination can also be defined. For this purpose, the perplexity $PP(q, D)$ is converted into a score dimension $[0; 1]$ via a sigmoid function,

$$PP'(q, D) = \frac{1}{1 + \exp(-PP(q, D))}, \tag{3.10}$$

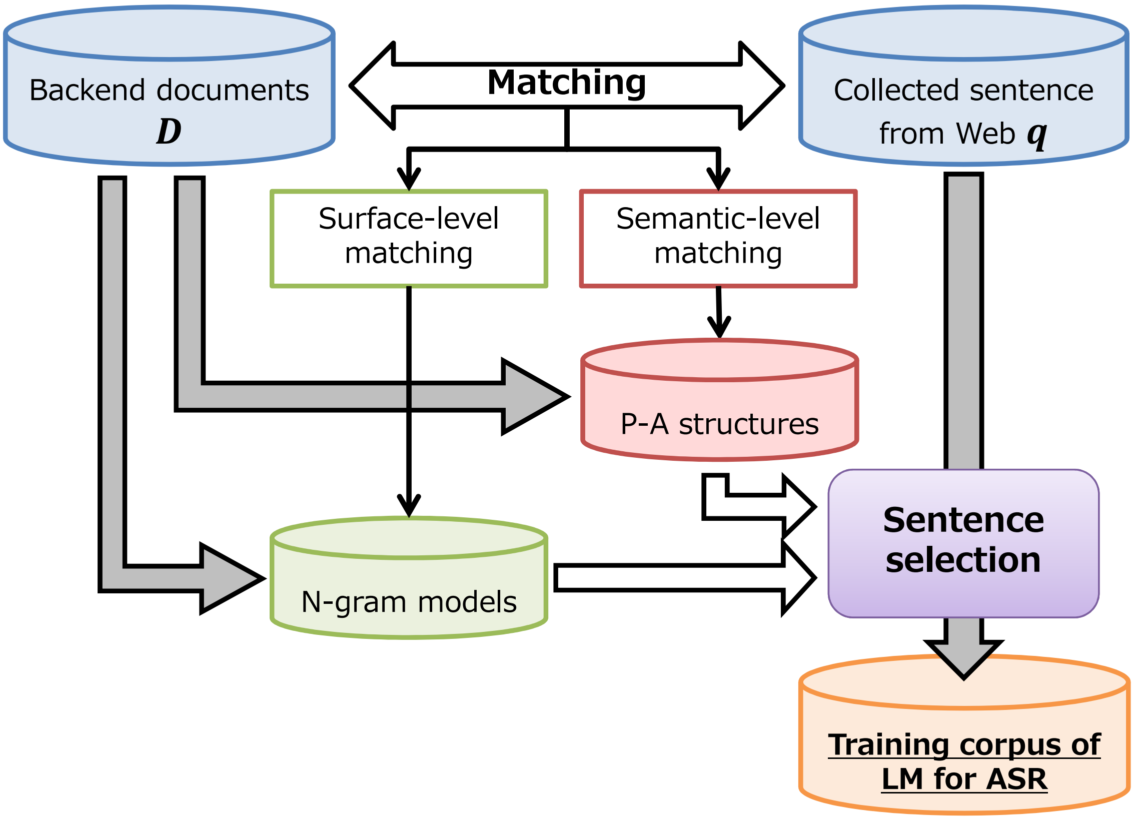


Figure 3.2: *Overview of the proposed selection method.*

which can be linearly-combined with the semantic relevance measure based on $PA(D|q)$ as,

$$PP + PA(q, D) = \alpha \times PA(D|q) + (1 - \alpha) \times PP'(q, D). \tag{3.11}$$

α is decided as $\alpha = 0.7$ in the preliminary evaluation.

The overall procedure is summarized in **Figure 3.2**, in which text selection is conducted based on the two relevance measures.

3.4 Experimental Evaluation

The proposed approach is evaluated in a speech-based information navigation system in the Japanese professional baseball domain and the Kyoto sightseeing domain (Misu and Kawahara, 2010). The system can answer user's question regarding the domain using the back-end documents of the respective domains. Evaluation is conducted by automatic speech recognition (ASR) with a measure of word error rate and semantic level accuracy.

Table 3.1: Details of training set.

Usage	Task	Corpus	Sentences
D	sightseeing	Wikipedia	35,641
	baseball	Mainichi News paper	176,852
q	sightseeing	Yahoo! QA: tourism - domestic	679,588
	baseball	Yahoo! QA: entertainment - baseball	403,602

Table 3.2: The details of test-set for evaluations.

Task	Speaker	Utterances
Kyoto sightseeing	4	219
Japanese professional baseball	10	2,747

3.4.1 Experimental Setting

The back-end document set D, is newspaper articles (Mainichi Newspaper Corpus) tagged with professional baseball and Wikipedia entries with a tag of Kyoto City for the respective domains.

The relevance measure described in Section 3.2 and 3.3 are trained with the document set D, and used for selecting query sentences collected in the Yahoo! QA Web site. Sentences in the "entertainment - baseball" category and "tourism - domestic" category are collected, respectively. The details of the training set, back-end document set D and query sentences q are shown in **Table 3.1**. The test set of user utterances was separately collected using the dialogue system and reading of collected sentences. The number of speakers and utterances of the test set are summarized in **Table 3.2**.

Word trigram LMs were trained with the texts selected based on the relevance measures described in Section 3.2 and 3.3. A variety of LMs are trained using the texts of different sizes relative to all available texts (3/10 through 10/10 where all texts are used) by changing the selection threshold.

3.4.2 Evaluation with ASR Accuracy and Perplexity

For evaluation of ASR accuracy, word error rate (WER) is computed for the test-set utterance. A speaker-independent triphone model for the acoustic model and the Julius decoder (Lee et al., 2001) are used. WER is plotted for LMs of different text sizes in **Figure 3.3** and

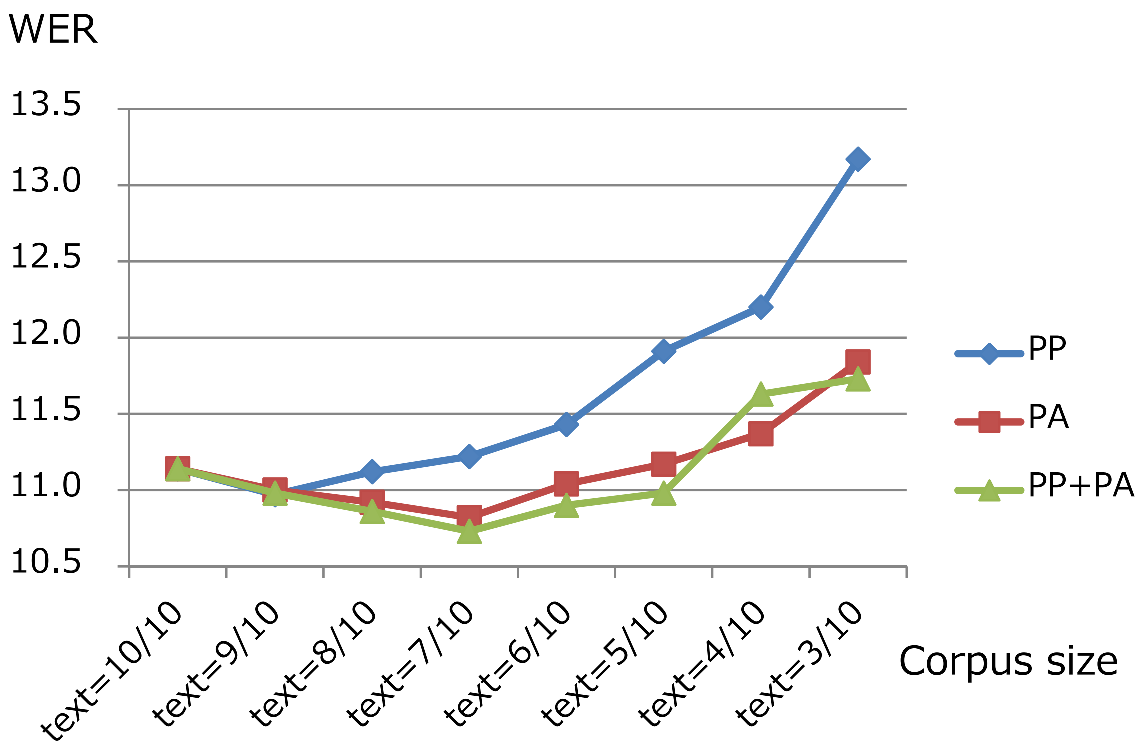


Figure 3.3: *Word error rate (*WER*) in baseball domain.*

3.4 for the baseball news domain and the Kyoto sightseeing domain, respectively. Adjusted test-set perplexity of each domain is plotted in **Figure 3.5** and **3.6** for reference. In the adjusted perplexity (APP), the probability of unknown words (<*UNK*>) is divided by the number of unseen lexical entries in the current training set.

PP is the result of the perplexity-based measure and PA is the result of the semantic relevance measure. In the preliminary evaluation, significant difference between the rank-based combination method and the score-based combination method is not observed. So, the results by the rank-based method is shown for PP+PA.

In the baseball news domain (**Figure 3.3**), it is shown that the proposed text selection based on semantic relevance (PA; text=7/10) results in significant WER reduction than the baseline method without any selection (text=10/10). The proposed semantic relevance measure (PA; text=7/10) and the combination of the two measures (PP+PA; text=7/10) is not different significantly. In the Kyoto sightseeing domain (**Figure 3.4**), the combination method works effectively.

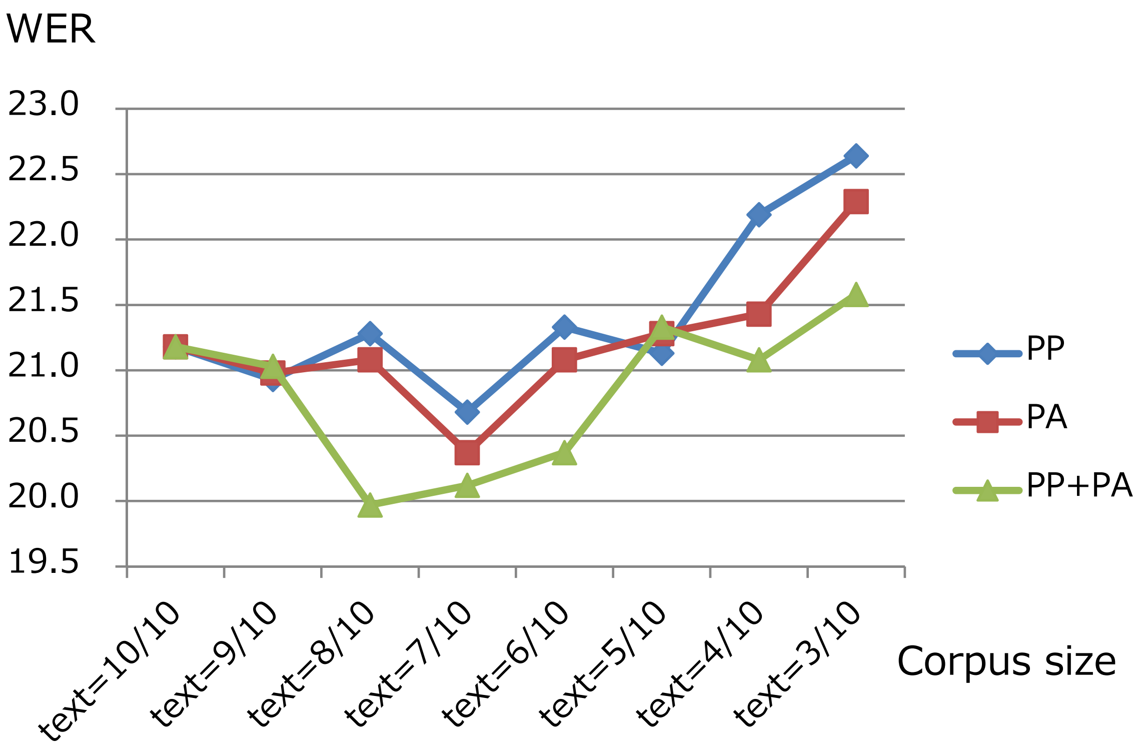


Figure 3.4: *Word error rate (*WER*) in sightseeing domain.*

The problem is how to decide the optimal point for selection in the graphs. When applying the method to a new task-domain, it is not possible to prepare a development set. The optimal point of the proposed method based on semantic relevance (PA) lies around text=7/10 in both domains, however, the optimal point of the perplexity-based method (PP) is different between two domains. If it is assumed that the Kyoto sightseeing domain is the development set for the baseball news domain, the optimal point is text=7/10 in both method (PP and PA). There is a significant difference between the two methods at this point in the baseball domain.

3.4.3 Evaluation with Semantic and Dialogue-level Accuracy

Next, evaluation is conducted with the semantic and dialogue-level accuracies, which are more related with the performance of the spoken dialogue system. Semantic accuracy (PAER) is measured by the error rate of P-A triplet, in which it is counted as correct if the triplet of the predicate, the argument and its semantic role are all correctly extracted. The automatically parsed ASR result and transcript are aligned to calculate the PAER. The

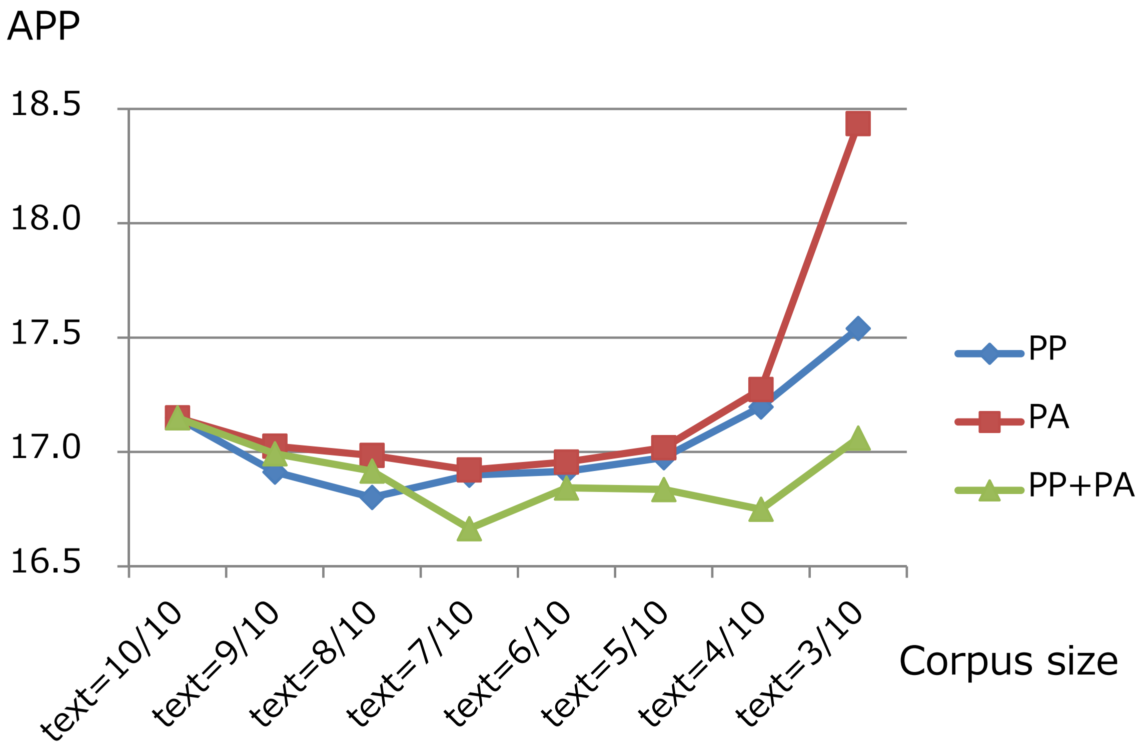


Figure 3.5: *Adjusted perplexity (*APP*) in baseball domain.*

test set of the Kyoto sightseeing task was too small to evaluate the PAER, thus, PAER for the baseball news domain is plotted in **Figure 3.7**. By using the LM selected (text=7/10) by the combination method (PP+PA), the PAER is reduced to 20.4% from the baseline 21.5% without text selection. It is not a significant difference because the number of the P-A triplet is small (2935 P-A triplets), but the result shows that the text selection based on deep semantic knowledge contributes to improvement of the semantic accuracy.

Dialogue-level accuracy is measured by the ratio of appropriate responses for all user queries. It is observed that an increase of the appropriate responses (by 0.8% absolute) as a result of the PAER improvement.

3.5 Conclusion

A novel text selection approach for training LMs for spoken dialogue systems is presented. Compared to the conventional perplexity criterion, the proposed approach introduces a semantic-level relevance measure with the back-end knowledge base used in the dialogue

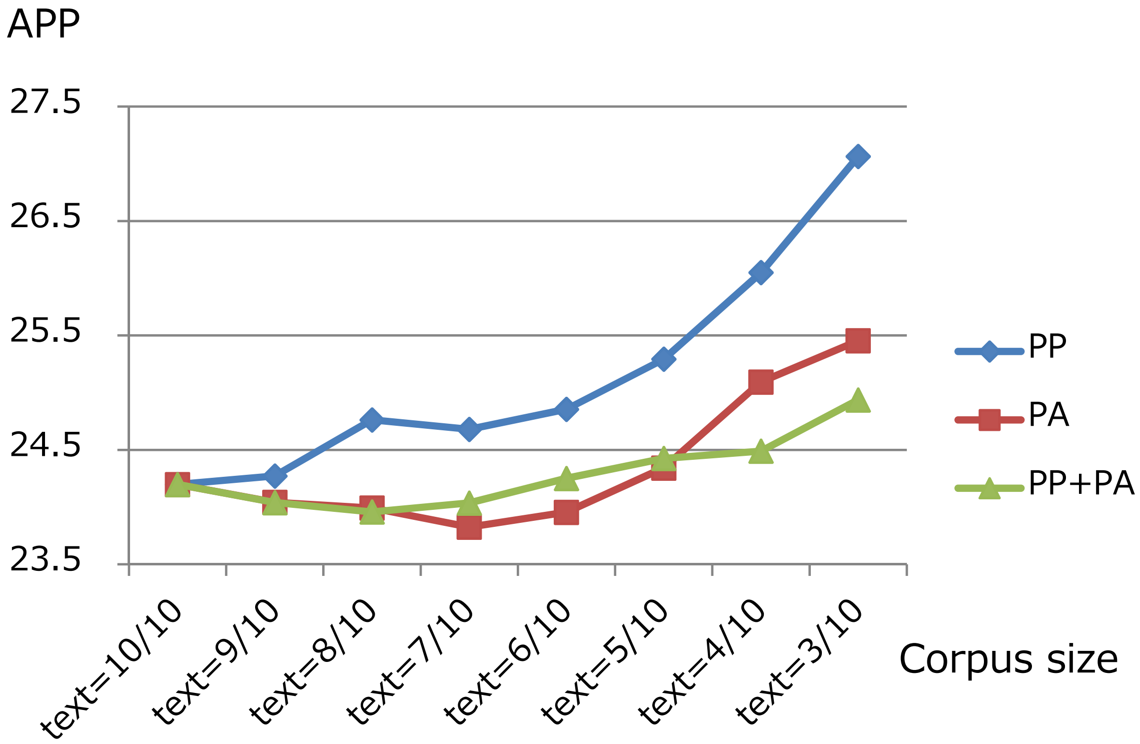


Figure 3.6: *Adjusted perplexity (*APP*) in sightseeing domain.*

system. Thus, it can effectively filter semantically relevant sentences for the task domain. It can also be combined with the perplexity measure for a synergistic effect. Experimental evaluations were conducted in two different domains, the proposed method demonstrated its effectiveness and generality. The combination method realized an improvement not only in WER but also in semantic and dialogue level accuracies (accuracy of the P-A structure analysis). The proposed approach only uses the texts in the back-end system, and does not require any "seed" corpus. Therefore, it can be used for building a spoken dialogue system of a particular domain from scratch.

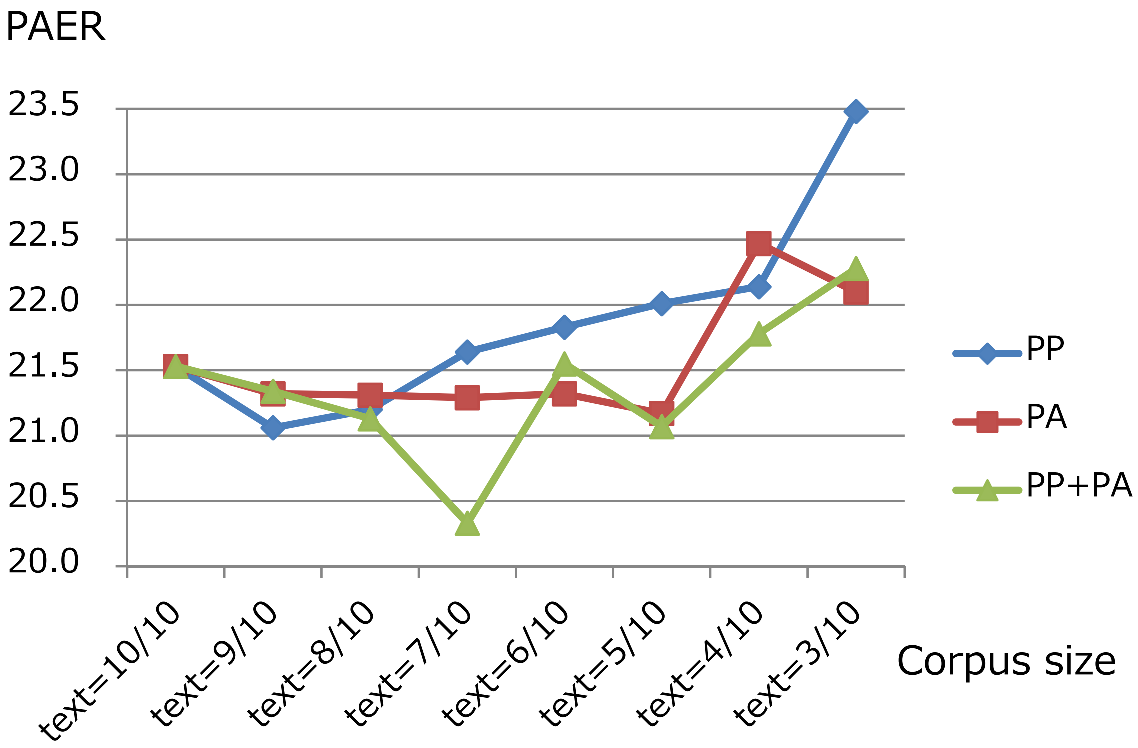


Figure 3.7: *P-A error rate (*PAER*) in baseball domain.*

Chapter 4

Presentation of Relevant Information based on Similarity of Semantic Structure

This chapter address a novel scheme of spoken dialogue modules which conducts information navigation based on the Web. The scheme is based on information extraction defined by the predicate-argument (P-A) structure and realized by semantic parsing. Based on the information structure, the dialogue system can perform question answering (QA) and proactive presentation (PP). In order to provide the most appropriate information to the user query, the modules use domain-dependent semantic structure of the P-A patterns defined in Chapter 2. Similarity measures of P-A structures are also introduced to select relevant information.

4.1 Information Retrieval for Spoken Dialogue Systems

Spoken dialogue systems that provide information to users are classified into two types as described in Chapter 1; using relational databases (RDB) (Dahl et al., 1994; Komatani et al., 2005; Raux et al., 2005) and using information retrieval techniques based on statistical document matching (Akiba and Abe, 2005; Misu and Kawahara, 2010). The document retrieval systems typically retrieves information by focusing on surface-level keywords, dependencies and question types of the queries. However, they do not consider semantic-level natural language processing result of the query or dialogue histories. As a result, they often output unnatural responses that are not related to the user demands or dialogue histories.

The proposed method focuses on the P-A structure, which is based on automatically defined domain knowledge. The system can find information that semantically matches the user query from a large document set on the Web. The P-A structure is widely used in the existing question answering systems (Kawahara et al., 2002; Dzikovska et al., 2003; Narayanan and Harabagiu, 2004; Harabagiu et al., 2005), but these approaches require a pre-defined P-A patterns for question answering. In this work, domain knowledge of the P-A structure is automatically extracted.

The information retrieval conducts information presentation that matches to potential demands of users. This concept is called "information concierge" (Hirayama et al., 2011), the system probes and clarifies the potential demands of users through a dialogue. The system presents partially-matched information even if there is no information exactly match to the user query. In the previous work of information concierge (Misu and Kawahara, 2010), the system presents information proactively if the user keeps silence. However, the system only presents a characteristic information in the document, which does not necessarily satisfy the user demand or dialogue history. In this work, the system searches information that matches to the P-A structure of a dialogue history, and presents information related to the user demands.

4.1.1 Proposed Information Retrieval based on P-A Structure

The architecture of the proposed spoken dialogue modules, question answering (QA), and proactive presentation (PP) is depicted in **Figure 4.2**. First, information extraction is conducted by parsing Web texts in advance. A user's query is also parsed to extract the same information structure, and the system matches the extracted information against the Web information. According to the matching result, the system either answers the user's question or makes proactive presentation of information which should be most relevant to the user's request.

4.1.2 Problems of Using P-A Structure

If the system finds some information which completely matches the user's query, the system makes a response using the corresponding Web text. When the system cannot find exact information, it searches for some information which matches partially. For example, in Figure 4.2, when a user asked "Did Ichiro hit a home-run?", the system cannot find exact

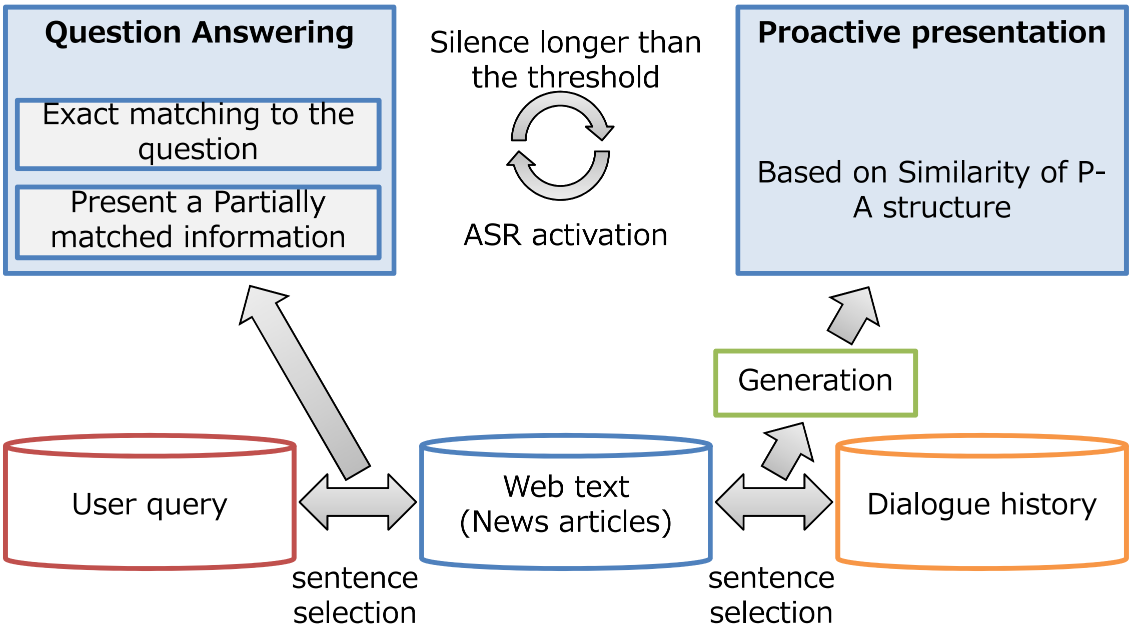


Figure 4.1: Architecture of the dialogue modules that retrieve related information to the user query.

information "[Ichiro (subject), home-run (object), hit]", but finds "[Ichiro (subject), double (object), hit]" which is partially matched and most relevant. This information is used to generate a relevant response that the user would expect.

In the conventional RDB-based dialogue scheme, the system hardly makes relevant responses if it finds no matched entries, thus usually replies "There is no matched entries". In the conventional question-answering scheme, the same situation often happens. Occasionally, a set of closely-matched answers may be found by statistical matching, but the found answers may not be relevant to the user's query. In the proposed scheme, it is guaranteed that the answer is at least partially matched to the user's query in terms of the semantic structure.

Flexible matching based on the P-A structure is critical, because the exact matching often fails and does not generate any outputs. In order to retrieve most relevant information, we define similarity measures of predicates and arguments, which are also learned from a domain corpus.

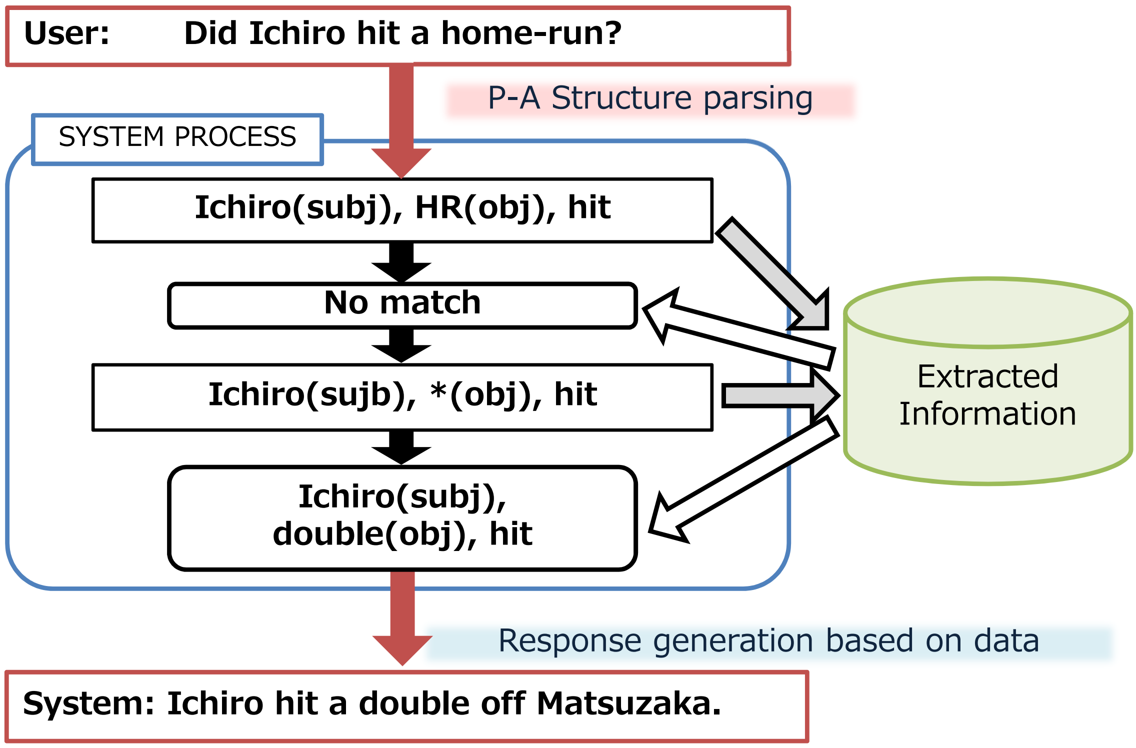

User: Did Ichiro hit a home-run?
P-A Structure parsing
SYSTEM PROCESS
Ichiro(subj), HR(obj), hit
No match
Ichiro(sujb), *(obj), hit
Ichiro(subj), double(obj), hit
Extracted Information
Response generation based on data
System: Ichiro hit a double off Matsuzaka.

Figure 4.2: An example of information retrieval based on the partial matching of P-A structures.

4.2 Presentation of Relevant Information using Similarity of P-A Structure (QA module)

The question answering (QA) module presents not only exact matched information but also proactively relevant information to the user query. When the system fails to find exact information that matches the user's query, the system tries to answer the question with proactive information presentation. It is based on the partially matched entries of the current or latest query. The fall-back is similar to collaborative response generation in the conventional spoken dialogue systems (Sadek, 1999), but it is intended for proactive information presentation using general documents.

4.2.1 Partial Matching based on P-A Significance Score

For preference among multiple components in the P-A pattern of the user query, the significance measure defined in Chapter 2 is used. Specifically, we relax (=ignore) the component

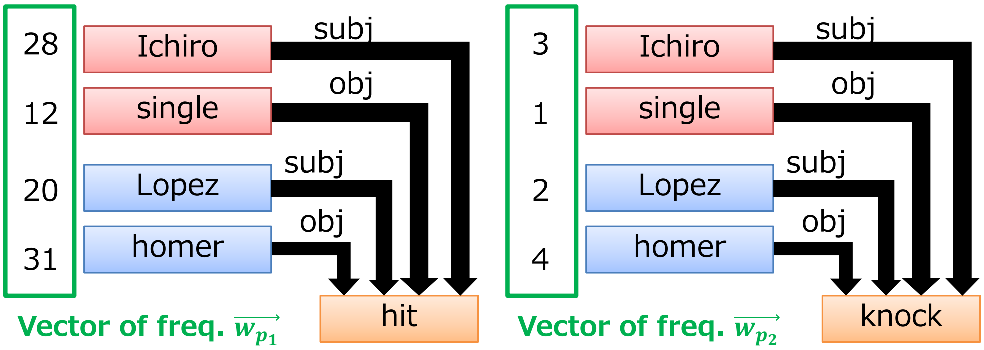


Figure 4.3: Examples of word vector of predicates.

of the least significance score, then search for relevant information. If any entry is not still matched, we relax the next less significant component. If multiple entries are found with this matching, we need to select the most relevant entry. Thus, we introduce two scores of relevance. The relevance measure is defined in different manners for predicates (=verbs or event nouns) and arguments. The measure for arguments is defined based on the co-occurrence statistics in the corpus. The measure for predicate is defined based on distributional analysis of arguments.

4.2.2 Relevance Score of Argument

The relevance of argument words (=nouns) w_i and w_j is defined as,

$$R_a(w_i, w_j) = \frac{\{C(w_i, w_j)\}^2}{C(w_i) \times C(w_j)}. \tag{4.1}$$

Here, w_i is in the original query, and relaxed (ignored) in the partial matching, and w_j of the best relevance score is retrieved for response generation. In the example of Figure 4.2, w_i is "home-run" and w_j is "double".

4.2.3 Relevance Score of Predicate

Distributional analysis (Harris, 1951; Lin, 1998) has been used to define similarity of words, assuming that similar words have similar contexts. Here, the distribution of arguments which have a modification relation to predicates (**Figure 4.3**) (Shibata et al., 2008; Pantel et al., 2009) is used. The relevance of predicate words w_{pre_i} and w_{pre_j} is defined as a cosine distance of occurrence vectors of the modifying arguments (Mitchell and Lapata, 2008; Thater et al., 2010). Here, argument entries are distinguished by their semantic roles such as subject and object, as shown in Figure 4.3. The relevance score of predicates is calculated by cosine similarity,

$$R_p(w_{p_i}, w_{p_j}) = \cos(\vec{w_{p_i}}, \vec{w_{p_j}}) = \frac{\vec{w_{p_i}} \cdot \vec{w_{p_j}}}{|\vec{w_{p_i}}||\vec{w_{p_j}}|},$$
$$\text{where } \vec{w_{p_i}} = (C(w_{s_1}, w_{a_1}), ..., C(w_{s_k}, w_{a_l})). \tag{4.2}$$

As the distribution of arguments is sparse and its reliable estimation is difficult, we introduce smoothing is introduced by using another distributional analysis of arguments. This relevance measure is introduced to compare arguments that appear in the same context. The measure is defined by cosine distance of word bi-gram vector: frequencies of antero-posterior position word. The vector of argument word w_{a_i} that has posterior position word list $\{w_{L1,a_i}, ..., w_{L_n,a_i}\}$ and prior words list $\{w_{R1,a_i}, ..., w_{R_n,a_i}\}$ is defined as,

$$\vec{w_{a_i}} = (C(w_{L1,a_i}), ..., C(w_{L_n,a_i}), C(w_{R1,a_i}), ..., C(w_{R_n,a_i})). \tag{4.3}$$

The cosine similarity to compare arguments is defined as,

$$\cos(\vec{w_{a_i}}, \vec{w_{a_j}}) = \frac{\vec{w_{a_i}} \cdot \vec{w_{a_j}}}{|\vec{w_{a_i}}||\vec{w_{a_j}}|}. \tag{4.4}$$

Here, we rewrite the count of a pair of semantic role and an argument $C(w_{s_k}, w_{a_l})$ that expresses the vector of predicates in Eqn. (4.2) as,

$$C'(w_{s_k}, w_{a_l}) = \delta \times C(w_{s_k}, w_{a_l}) + (1 - \delta) \sum_{j \ s.t. \ j \neq i} \frac{\cos(\vec{w_{a_i}}, \vec{w_{a_j}})}{\sum_{k \ s.t. \ k \neq i} \cos(\vec{w_{a_i}}, \vec{w_{a_k}})}. \tag{4.5}$$

The value of δ is decided as 0.6 in preliminary experiments. The Eqn. (4.5) reduces the problem of data sparseness by adding counts of similar argument to the target argument. Here, the smoothing count of argument that has high cosine similarity ($\cos(\vec{w_{a_i}}, \vec{w_{a_j}}) \geq 0.5$) is added to avoid the combinatorial explosion problem.

4.2.4 Back-off to Bag-Of-Words Model

If no entry is matched with all possible partial matching, we can resort to the naive "bag-of-words" (BOW) model, in which a sentence is represented with a vector of word occurrence and matching is done based on this vector. This method is widely used for document retrieval. We count only content words. The score is defined as,

$$R_{\text{BOW}}(s_i, s_j) = \cos(\vec{v_i}, \vec{v_j}) = \frac{\vec{v_i} \cdot \vec{v_j}}{|\vec{v_i}||\vec{v_j}|}, \tag{4.6}$$

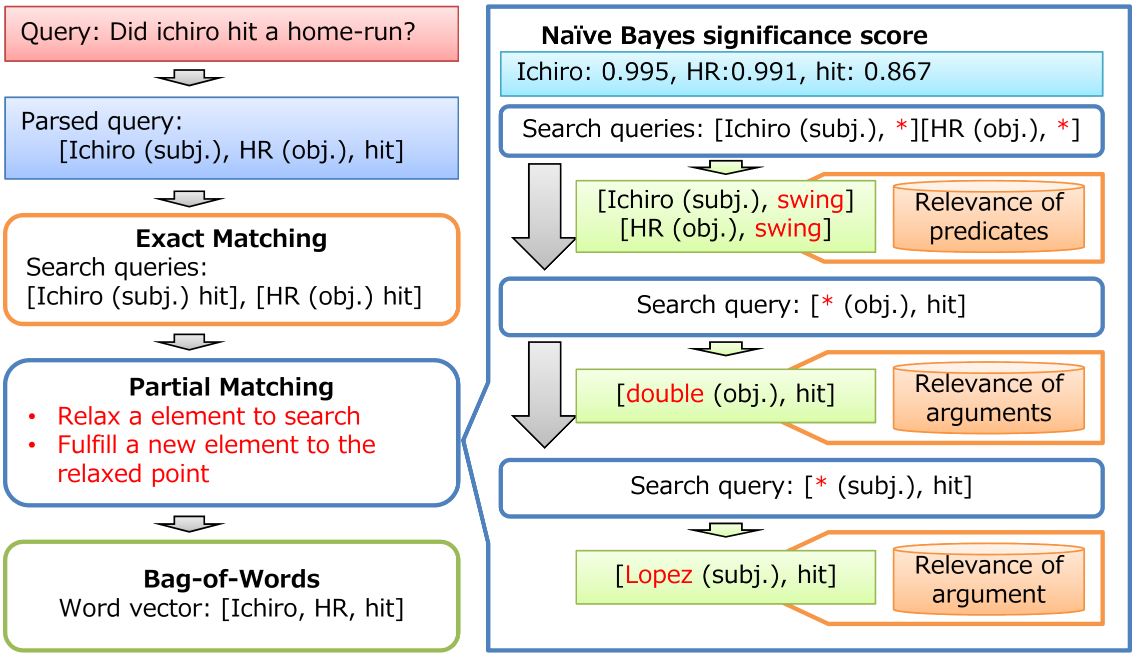

Query: Did ichiro hit a home-run?
Parsed query: [Ichiro (subj.), HR (obj.), hit]
Exact Matching
Search queries: [Ichiro (subj.) hit], [HR (obj.) hit]
Partial Matching
Relax a element to search
Fulfill a new element to the relaxed point
Bag-of-Words
Word vector: [Ichiro, HR, hit]
Naïve Bayes significance score
Ichiro: 0.995, HR:0.991, hit: 0.867
Search queries: [Ichiro (subj.), *][HR (obj.), *]
[Ichiro (subj.), swing] [HR (obj.), swing]
Relevance of predicates
Search query: [* (obj.), hit]
[double (obj.), hit]
Relevance of arguments
Search query: [* (subj.), hit]
[Lopez (subj.), hit]
Relevance of argument

Figure 4.4: The overall matching strategy of the proposed scheme.

by using cosine similarity of the count of content words $\vec{v_i} = \{C(w_{i_1}), ..., C(w_{i_n})\}$ in a sentence (or document) s_i. In this method, the significance score is used for preference of the words when multiple candidates are matched for a short query.

The overall matching strategy of the proposed scheme is summarized in **Figure 4.4**.

4.2.5 Selection of Relevant Information from a Sentence

Answer or information presentation is generated based on the matched sentence in a newspaper article. As a sentence is often complex or made of multiple predicates, simple presentation of the sentence would be redundant or even irrelevant. Therefore, the portion of the matched P-A structure need to be selected, to generate a concise response relevant to the user's query. For example, when a sentence "Ichiro hit a three-run homer in the seventh inning and Mariners won the game" is matched by the pattern "[Ichiro(subject), hit]", we select the former portion of the sentence which exactly answers the user's query, and generate a response "Ichiro hit a three-run homer in the seventh inning."

4.3 Proactive Presentation based on Similarity of P-A Structure (PP module)

The proactive presentation (PP) module presents information that is related to the dialogue history if the system detects a pause longer than a threshold. This function enables to present a new information that the user is interested in, even if the user does not make a question. The PP module uses the significance score and the relevance measure of the predicate and arguments. The relevance score is defined by the significance score and the relevance measure of the predicate and arguments.

4.3.1 Similarity of Sentences based on P-A Structure

For proactive presentation, a relevance score between sentences is defined based on the P-A structure. The relevance score is defined for a pair that have the same named entity as a subject to provide relevant information. The sentence-level relevance measure is defined as,

$$\begin{aligned} R_s(s_i, s_v) = & R_p(w_{p_i}, w_{p_v}) \times \frac{P(D|w_{p_i}) + P(D|w_{p_v})}{2} \\ & + \sum_j R_a(w_{a_j}, w_{a_v}) \times \frac{P(D|w_{a_j}) + P(D|w_{a_v})}{2}. \end{aligned} \tag{4.7}$$

Here, s_i is a sentence in the dialogue history and s_v is a candidate information to present. w_p and w_a are a predicate and arguments, included in u_i and u_v. $R_p(.)$ and $R_a(.)$ are relevance of a predicate and arguments that is defined in Section 4.2.2 and Section 4.2.3. The relevance score of predicates and arguments are weighted by the significance score of the domain D ($P(D|w_i)$). The system considers the dialogue history of both the user and the system, and calculates the dialogue-level relevance measure by using the latest h utterances,

$$R_{\mathsf{pp}} = \sum_{i \in h} R_s(u_i, u_v). \tag{4.8}$$

In this work, h is 2. The system presents information that has the highest score of Eqn. (4.8). The output of the PP module is generated from the sentence that has the highest score and the published date of the article.

Table 4.1: Evaluation of system response.

Input	Model	Correct	Ambiguous	Incorrect	No Answer
Text	Exact	0.299	0.005	0.015	0.681
	Exact+Partial	0.662	0.050	0.203	0.085
	Exact+Partial+BOW	**0.697**	0.050	0.253	0.000
	(cf) BOW	0.468	0.139	0.393	0.000
	(cf) SOW	0.542	0.114	0.343	0.000
Speech (ASR)	Exact	0.194	0.010	0.005	0.791
	Exact+Partial	0.572	0.060	0.189	0.179
	Exact+Partial+BOW	**0.641**	0.065	0.289	0.000
	(cf) BOW	0.398	0.094	0.488	0.000
	(cf) SOW	0.463	0.104	0.433	0.000

4.4 Evaluations

This section shows experimental results of the proposed two modules: question answering (QA) and proactive presentation (PP). The QA module is evaluated on the task of information retrieval from the target newspaper articles. The PP module is evaluated by subjective and objective experiments with users.

4.4.1 Evaluation of Presentation of Relevant Information (QA module)

The significance score (Naive Bayes model) and the relevance score were learned using the Mainichi Newspaper corpus of ten years (2000-2009). For evaluation of the system, we prepared 201 questions from news articles (September 19-26, 2010) seen at the website of Mainichi Newspaper[1]. Correct answers to the test queries were annotated manually. Evaluation was done with the text input as well as speech input. A word N-gram language model for ASR dedicated to the domain was trained using the relevant newspaper article corpus. The word error rate was approximately 24%.

The system responses for the test queries are categorized into one of the following four: correct answer only "Correct", case which includes the correct answer but also other redundant answers "Ambiguous", incorrect answer "Incorrect", and "No Answer". The ambiguous cases occur when multiple sentences or predicates are matched. Recall, precision and F-measure are also calculated by counting individual answers separately even when multiple answers are output. The results based on these evaluation measures are

[1] http://www.mainichi.jp

Table 4.2: Accuracy of system response.

Input	Model	Precision	Recall	F-measure
Text	Exact	0.938	0.303	0.458
	Exact+Partial	0.725	0.711	0.718
	Exact+Partial+BOW	0.701	0.746	0.723
	(cf) BOW	0.498	0.607	0.547
	(cf) SOW	0.552	0.656	0.600
Speech (ASR)	Exact	0.891	0.204	0.332
	Exact+Partial	0.658	0.632	0.645
	Exact+Partial+BOW	0.617	0.706	0.659
	(cf) BOW	0.429	0.493	0.459
	(cf) SOW	0.483	0.567	0.522

summarized in **Table 4.1** and **Table 4.2** for text input and speech input.

In the tables, the proposed method is broken down into three phases as shown in Figure 4.4: exact matching of P-A structure, incorporation of the partial matching, and back-off to the "bag-of-words" (BOW) model. For comparison, the BOW model and "sequence-of-words" (SOW) model are also tested. SOW model considers the sequence order in the BOW model. The exact matching assumes strong constraint of P-A patterns, so the generated answers are almost correct, but no answers are generated very often. By incorporating the partial matching and BOW model, the system can output more relevant answers. Compared with the BOW model, the proposed method achieves much higher ratio or precision of correct answers. F-measure is also higher by 17% absolute.

A similar tendency is observed for speech input, although the overall accuracy is degraded because of the ASR errors. However, degradation is relatively small considering the word accuracy of 76%. The partial matching works effectively even if the exact matching fails due to ASR errors. Moreover, the back-off to the BOW model is effective in ASR input.

The proposed method generates concise responses by selecting the relevant portion as described in Section 4.2.5, while the BOW method often generates long responses which includes many redundant portions. This property is particularly important in the speech interface.

Table 4.3: Objective evaluation by a majority.

Evaluation	Num.
(a) Correct.	32
(b) Correct, but it includes grammatical errors.	4
(c) Correct, but it is diffuse.	2
(d) Incorrect.	8

4.4.2 Evaluation of Proactive Presentation (PP) module

Subjective and objective experiments were conducted with examiners. Here, 140 pairs of questions and answers that are labeled as "Correct" in Section 4.4.1 are used as a history of dialogue ($h = 2$) of the proposed system, and 55 pairs outputted proactive presentations that have the same subjective named entity. The user question, system response and proactive presentation are given to examiners who evaluates the appropriateness of the proactive presentation in the context (=history). Three examiners evaluated on two criteria:

(1) Appropriateness of the presented information to the context (objective evaluation).

(2) How does the user feel if he or she was presented the information on the situation of proactive presentation (subjective evaluation)?

In the evaluation (1), examiners evaluated the output among four categories:

(a) Correct.

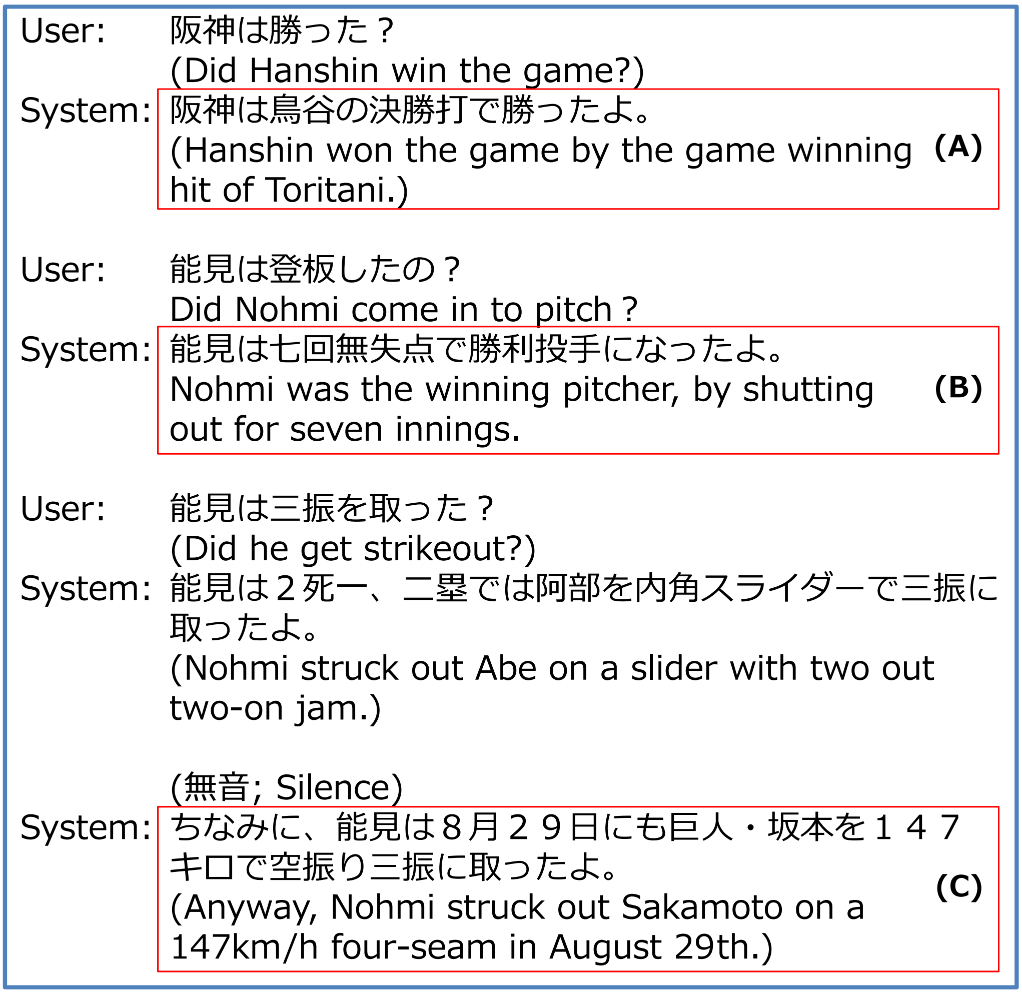
(b) Correct, but includes grammatical errors.

(c) Correct, but it is diffuse.

(d) Incorrect.

There are 9 presentations that the evaluation was split among examiners, and the evaluation of 46 presentations was decided by a majority. As shown in **Table 4.3**, number of (a) "correct" was 32. There are 38 "correct" including (b) "grammatical errors" or (c) "diffusion", and the total ratio of the correct responses was 82.6%.

In the evaluation (2), examiners evaluated the output in the 4-point scale. The average of the score was 3.33. It demonstrates that the system generally made meaningful proactive presentations.



User: 阪神は勝った？
(Did Hanshin win the game?)

System: 阪神は鳥谷の決勝打で勝ったよ。
(Hanshin won the game by the game winning hit of Toritani.) **(A)**

User: 能見は登板したの？
Did Nohmi come in to pitch？

System: 能見は七回無失点で勝利投手になったよ。
Nohmi was the winning pitcher, by shutting out for seven innings. **(B)**

User: 能見は三振を取った？
(Did he get strikeout?)

System: 能見は２死一、二塁では阿部を内角スライダーで三振に取ったよ。
(Nohmi struck out Abe on a slider with two out two-on jam.)

(無音; Silence)

System: ちなみに、能見は８月２９日にも巨人・坂本を１４７キロで空振り三振に取ったよ。
(Anyway, Nohmi struck out Sakamoto on a 147km/h four-seam in August 29th.) **(C)**

Figure 4.5: Dialogue example (original in Japanese, translated to English).

4.5 Dialogue Example

An example of dialogue that is based on the proposed modules (QA and PP) is shown in **Figure 4.5**. In the example, first, the system answers a user question with QA module with exactly matched information in (A). Next, the system cannot find the exact P-A triplet of “[Nohmi (subject), come into pitch]”, but the system can answer with partially matched information of “[Nohmi (subject), was the winning pitcher]” in (B). Finally, the system presents information that is related to the dialogue history in (C) with the PP module.

4.6 Conclusion

This chapter addressed a new scheme of spoken dialogue modules which can talk about Web texts in an interactive manner. The information extraction technique is adopted to conduct question answering as well as proactive information presentation. A statistical significance measure based on Naive Bayes works effectively for automatic selection of the important P-A structure in a given domain. Relevance measures are also defined for a predicate and arguments in order to retrieve relevant entries when the exact matching does not succeed. This information retrieval method enables the system to continue a dialogue by using information that matches the user query on the semantic-level. The proposed module improved the accuracy of information retrieval than conventional bag-of-words scheme. The improvement of F-measure was 17.6 points with a text input and 19.6 points with speech input.

The proactive presentation module is also proposed based on the similarity of P-A structure. The module presents a relevant information proactively by using the relevance measure of arguments and predicates. The experimental result shows that the proposed model can present meaningful information by following the dialogue context.

The proposed modules are based on an unsupervised framework from unannotated corpora, and the framework does not require a construction of any relational database or thesauruses. This feature enables easy adaptation of the system to a variety of domains.

Chapter 5

Statistical Learning of Dialogue Model by Tracking Dialogue State and User Focus

This chapter addresses a spoken dialogue management for the information navigation system. The proposed dialogue management is based on partially observable Markov decision process (POMDP) and conducts information navigation by selecting the most appropriate dialogue module to respond the user. The reward function (of POMDP) is defined by the quality of interaction to apply the POMDP to a task that does not have a clear task goal. POMDP also tracks a dialogue state and user's focus of attention to make appropriate actions to the user.

This chapter first introduces general statistical dialogue management based on reinforcement learning and then, the model of dialogue management of information navigation. Spoken language understanding (SLU) based on discriminative model provides an input of dialogue manager. POMDP updates the belief of user state by using the current SLU result and the previous belief that contracts the history of dialogue. The belief update allows for using a long context of dialogue. The framework is extended to track the user focus, attentional state of the user. The user focus is also detected by a discriminative model of SLU, and modeled as a stochastic variable in a belief. The belief of user focus is also updated to track the context of the dialogue.

5.1 Dialogue Management for Spoken Dialogue Systems

Dialogue management of spoken dialogue systems was usually made in a heuristic manner and often based on simple rules (Bohus and Rudnicky, 2003; Bratman et al., 1988; Lucas, 2000). In the past years, machine learning, particularly reinforcement learning (RL), have been investigated for dialogue management. Markov decision processes (MDPs) and partially observable Markov decision processes (POMDPs) are the most successful and now widely used to model and train dialogue managers (Roy et al., 2000; Levin et al., 2000; Williams and Young, 2007; Young et al., 2010). These approaches allow us to consider all possible future actions of a dialogue system, and thus to obtain a new optimal dialogue strategy which could not be anticipated in conventional hand-crafted dialogue systems.

5.1.1 Dialogue management based on POMDP

After several early successful applications of statistical approaches to various tasks, the focus of research in dialogue management is shifting towards two directions. The first is how to deploy such methods in realistic goal-oriented dialogue systems and the other is how to develop a non-goal-oriented dialogue system with the statistical approach. On the first viewpoint, unsupervised dialogue-act annotation (Lee et al., 2012) is one means of avoiding labeling costs when we can obtain actual dialogue data in a target domain. (Gašić et al., 2013) proposed a belief adaptation with Bayesian update (Thomson and Young, 2010) of dialogue states. This method adapts the belief of an existing POMDP-based dialogue manager to some other tasks by following Gaussian processes. The other approach is a *hybrid* of the statistical dialogue management based on POMDP and the conventional rule-based dialogue management. Hand-crafted rules or knowledge databases constructed for conventional rule-based dialogue system can improve the statistical dialogue management (Lemon et al., 2006; Williams, 2008; Young et al., 2010; Varges et al., 2011). One of the problems of such POMDP-based system optimization is that the task and the domain are mixed in existing systems. In this work, the task structure and the domain knowledge are separated so a flexible dialogue manger for a variety of domains can be designed in the task of information presentation.

Another problem of POMDP dialogue systems is that the parameter space is so large, that it is very difficult to find exact solutions. Typical approaches to the problem is to use approximations (Monahan, 1982) such as hypothesis cutting (e.g. by using task knowledge or parametric distribution selection) and to convert a feature vector into a small dimension (Spaan and Vlassis, 2005). These approaches eliminate superfluous hypotheses of the dialogue manager based on rules or knowledge databases.

In this work, we abstract the task structure of dialogue to a limited number of dialogue modules. Information of user focus is also compacted into a binary flag of focus existence that the user query has any user focus or not. This compaction is decided on the trade-off between performance and efficiency of POMDP, memory and also portability across domains.

5.1.2 Dialogue Management in Non-Goal-Oriented Systems

The conventional scheme for goal-oriented systems assumes that the task and dialogue goal are clearly defined and readily encoded in the RL reward function. This is not true in casual conversation or information navigation addressed in this work.

Some previous work has tackled with this problem. (Pan et al., 2012) designed a spoken document retrieval system whose goal is user's information need satisfaction, and defined rewards by using the structure of the target document set. This is possible only for well-defined document search problems. The strategy requires a structure of the document set and definition of user demand satisfaction. (Shibata et al., 2014) developed a conversational chatting system. It asks users to make evaluation at the end of each dialogue session to define rewards for reinforcement learning. (Meguro et al., 2010) proposed a listening dialogue system. In their work, levels of satisfaction were annotated in the logs of dialogue sessions to train a discriminative model. These approaches require costly input from users or developers, who provide evaluation and supervision labels. In this work, a framework in which reward is defined for the quality of system actions and also for encouraging long interactions is explored, in contrast to the previous approaches. Moreover, user focus is tracked to make appropriate actions, which are more rewarded.

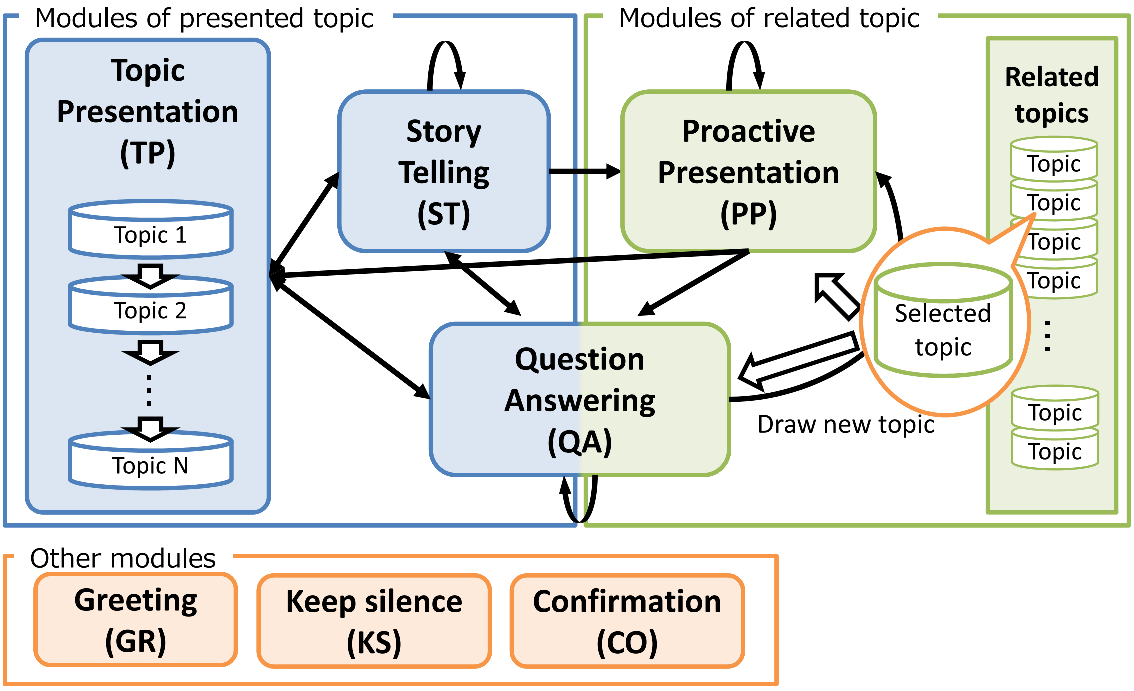


Figure 5.1: *Overview of the information navigation system.*

5.2 Dialogue Management for Information Navigation System

As described in Section 1.4, the task of information navigation is decoded as a module selection of seven dialogue modules: topic presentation (TP), story telling (ST), question answering (QA), proactive presentation (PP), greeting (GR), keep silence (KS) and confirmation (CO). The dialogue manager selects a module (action decision) based on an input of a user intention. A user intention is encoded as a request to the system, the user intention has six classes and each intention has a corresponding system action. The detail of user intention is described in Section 5.3.3. The system also detects a user focus that is tagged on at least one phrase in a user utterance. The existence of the user focus is used in manger for the belief update of POMDP framework. The aim of the dialogue management for information navigation is to decide the best system action module for the output of SLU modules: user intention analysis and user focus detection.

Again, the overview of the proposed system is shown in **Figure 5.1** (This is the same figure as Figure 1.8). The system starts a dialogue from TP module by presenting head-

```
SYS, tp, “Matsui, a part-time coach for the Giants, demonstrated 5 homeruns.
          The Giants president told him "Come back as a player!”
USR, st, 0, ” Give me the detail?”, ”Give:me/0 the:detail/0”
SYS, st, ”Matsui joined the extra batting training in the Giants' camp, and
          demonstrated his batting to young players.”
USR, qa, 1, “Will Matsui become the Giants' manager in the future?”,
             ”Matsui/0 will:become/0 the Giants' manager/1 in the future/0”
SYS, qa, ”The director of Giants told him "Come back as a player!””
USR, tp, 0, ”Next news, please.”, ”Next:news/0 please/0”
SYS, tp, ” To be a ace pitcher, has Fujinami improved from the rookie year?
…
```

Figure 5.2: An example of annotation for collected dialogue. System utterances have a tuple of three elements separated by a comma: utterer, called module and utterance. User utterances have a tuple of four elements: utterer, the module the user request falls in, binary information of user focus, utterance and user focus annotation on each phrase or P-A element.

lines of the news of the day. When the user is interested in the presented topic, the system describes the detail of topic with ST module. When the user asks a question, the system answers the question with QA module. PP module is activated to present a related information even if the user does not make a question. Transitions between the modules are allowed as shown in Figure 5.1, except GR, KS and CO modules, these modules can be called at any time.

5.3 Spoken Language Understanding (SLU)

In this section, the spoken language understanding (SLU) components of the system is presented. It detects the user's focus and intention and provides them to the dialogue manager. The SLU modules are formulated with a statistical discriminative model to give likelihoods which are used in POMDP.

5.3.1 Training Data and Annotation

We collected 606 utterances (from 10 users) with a rule-based dialogue system, and 312 utterances (from 8 users) with a preliminary statistical-based dialogue system which was constructed by using the data collected with the rule-based system. An example of annotation is shown in **Figure 5.2**. Annotation points are highlighted in the bold font.

To prepare the training data, each utterance was labeled with one of the six tags, indicating the best module to respond. In addition, each phrase or P-A elements is labeled whether it is the user's focus or not. The user focus is defined as "the main piece of information of interest to the user." The system reply should include the information that is a piece of interest to the user. For example, in the second user utterance in Figure 5.2, the user's focus is the phrase "the Giants' manager" and it should be included in the system response.

5.3.2 User Focus Detection based on CRF

To detect the user focus, a conditional random field (CRF) is used. The problem is defined as a sequential labeling of the focus labels to a sequence of the phrases of the user utterance. Features used are listed in **Table 5.1**. ORDER features are the order of the phrase in the sequence and in the P-A structure. These features are adopted because the user focus often appears in the first phrase of the user utterance. POS features are part-of-speech (POS) tags and their pairs in the phrase. P-A features are the semantic role of the P-A structure. The P-A significance score defined in Section 2.2.2 (P-A Score) is also incorporated. The score is discretized to 0.01, 0.02, 0.05, 0.1, 0.2, 0.5.

Table 5.2 shows the accuracy of user focus detection, which was conducted via five-fold cross-validation of the entire training data. CRFsuite (Okazaki, 2007) is used as a CRF classifier. "Phrase" is phrase-base accuracy and "sentence" indicates whether the presence of any user focus phrase was correctly detected (or not), regardless of whether the correct phrase was identified. This table shows that WORD features are effective for detecting the user focus, but they are not essential in the sentence-level accuracy. For portability across domains, we adopt the sentence-level features, and do not use the WORD features.

CRF gives a probability of whether any user focus is detected for each phrase h_l in the ASR result h. The sentence-level probability of focus existence is calculated by combining the sequence of probabilities of user focus label of being negative (0);

$$P(o_f|h) = 1 - \prod_l P(o_{f_l} = 0|h_l). \tag{5.1}$$

5.3.3 User Intention Analysis based on LR

The module classifies the six user intention categories from the user utterance.

Table 5.1: Features of user focus detection.

feature type	feature
ORDER	Rank in a sequence of phrases Rank in a sequence of elements of P-A
POS	POS tags in the phrase POS tag sequence
POSORDER	Pair of POS tag and its order in the phrase
P-A	Which semantic role the phrase has Which semantic roles exist on the utterance
P-AORDER	Pair of semantic role and its order in the utterance
P-A score	P-A significance score
WORD	Words in the phrase Pair of words in the phrase Pair of word and its order in the phrase

Table 5.2: Accuracy of user focus detection.

	Accuracy
phrase	78.5%
phrase + (WORD)	80.8%
sentence (focus exists or not)	99.9%
sentence (focus exists or not) + (WORD)	99.9%

- *TP*: request to the TP module.
- *ST*: request to the ST module.
- *QA*: request to the QA module.
- *GR*: greeting to the GR module.
- *NR*: silence longer than a threshold.
- *II*: irrelevant input due to ASR errors or noise.

Logistic regression (LR) is adopted for the dialogue act tagging (Tur et al., 2006). The probability of user intention o_s given an ASR result of the user utterance h is defined as,

$$P(o_s|h) = \frac{\exp(\omega \cdot \phi(h, o_s))}{\sum_i \exp(\omega \cdot \phi(h, o_{s,i}))}. \tag{5.2}$$

Here, $\phi(h, o_s)$ is a feature vector and ω is a feature weight. We use POS, P-A and P-A score as a feature set. In addition, we add a typical expression feature (TYPICAL) to classify *TP*, *ST* or *GR* tags. For example, typical expressions in conversations are “Hello” or “Go on,” and those in information navigation are “News of the day” or “Tell me in detail.” Features for the classifier are listed in **Table 5.3**.

Table 5.3: Features of user intention analysis.

feature type	feature
POS	Bag of POS tags Bag of POS bi-gram
P-A	Bag of semantic role labels Bag of semantic role labels bi-gram Pair of semantic role label and its rank
P-A score	P-A significance score
TYPICAL	Occurrence of typical expressions

Table 5.4: Accuracy of user intention analysis.

	All features	without TYPICAL
TP	98.7%	98.7%
ST	75.8%	65.3%
QA	95.1%	93.5%
GR	97.7%	97.7%
II	31.3%	31.3%
All	93.0%	92.2%

The classification accuracy in five-fold cross-validation is shown in **Table 5.4**. The TYPICAL feature improves the classification accuracy while keeping the domain portability.

5.3.4 SLU for ASR output

An overview of spoken language understanding that uses N-best list of ASR is depicted in **Figure 5.3**. ASR and intention analysis involves errors. Here, s is a true user intention and o_s is an observed intention. The observation model $P(o_s|s)$ is given by the likelihood of ASR result $P(h|u)$ (Komatani and Kawahara, 2000) and the likelihood of the intention analysis $P(o_s|h)$,

$$P(o_s|s) = \sum_h P(o_s, h|s) \tag{5.3}$$

$$\approx \sum_h P(o_s|h)P(h|u). \tag{5.4}$$

Here, u is an utterance of the user. We combine the N-best ($N = 5$) hypotheses of the ASR result h.

The user focus detection also needs to consider ASR errors. The probability of user focus is given by the likelihood of ASR result $P(h|u)$ and the likelihood of the user focus

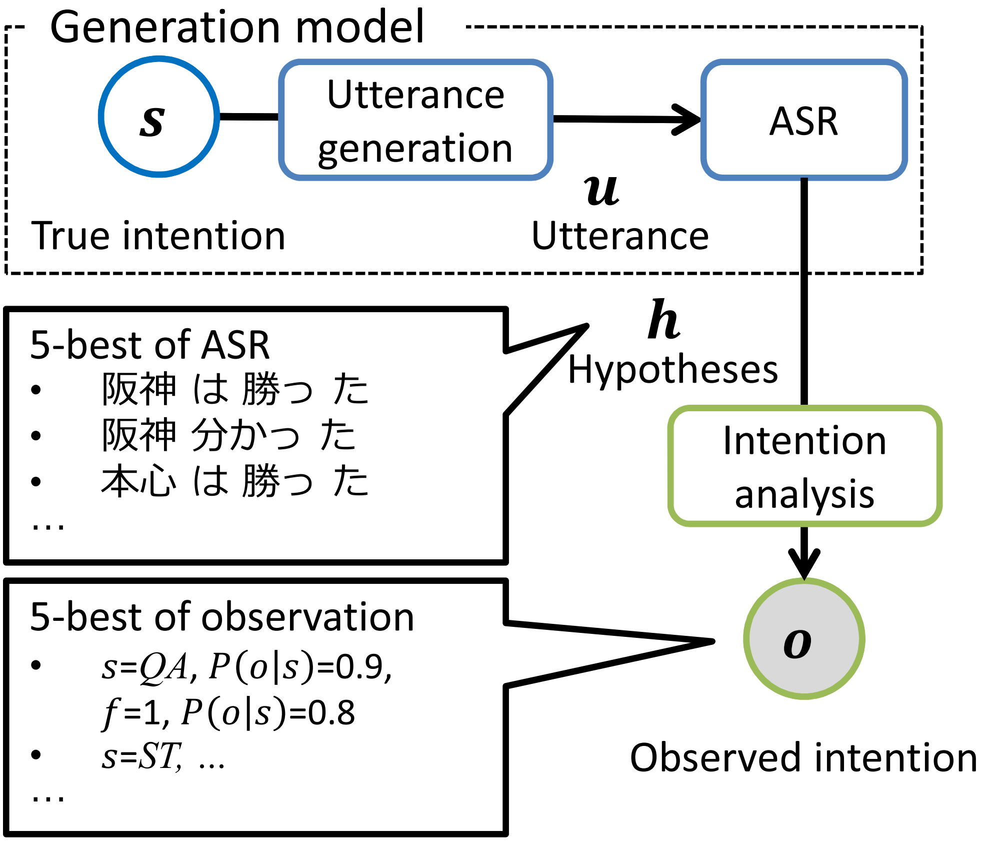


Figure 5.3: Overview of spoken language understanding (SLU).

detection $P(o_f|h)$,

$$P(o_f|f) = \sum_h P(o_f, h|f) \tag{5.5}$$

$$\approx \sum_h P(o_f|h)P(h|u). \tag{5.6}$$

5.4 Dialogue Management for Information Navigation

5.4.1 Dialogue Management based on POMDP

POMDP is a probabilistic extension of MDP that deals with the posterior probabilities of hidden user states (referred to as *beliefs*), which are recursively estimated from observation sequences. Belief updates are performed using the transition probabilities of the hidden states, combined with observation probabilities. The objective of POMDP optimization is to produce a policy that maps from beliefs to system actions such that the overall expected cost of the dialog is minimized. To calculate these observation and transition probabilities,

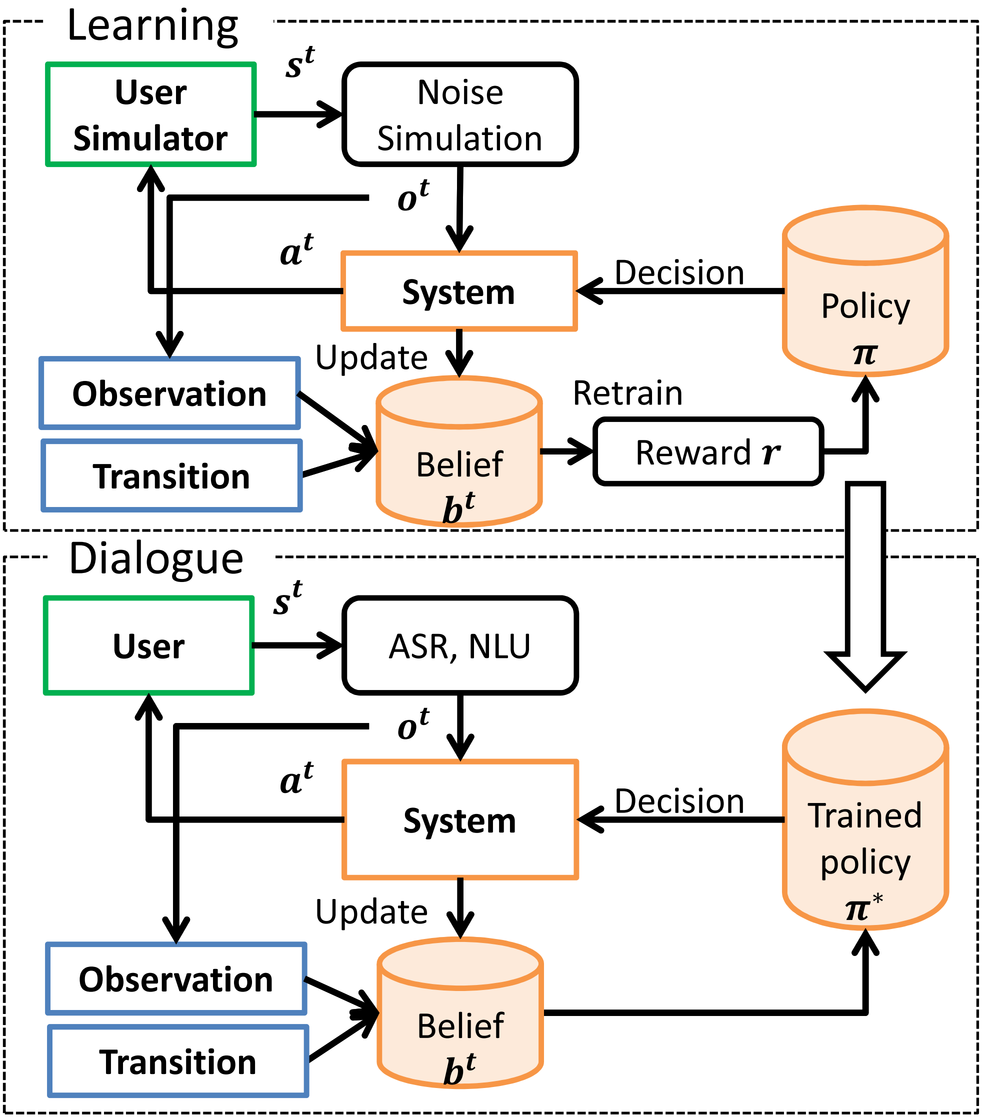


Figure 5.4: Overview of the proposed statistical dialogue management and learning scheme.

this optimization requires data from dialogue corpora, which are manually annotated with task-oriented dialogue act tags that correspond to the hidden user states.

Figure 5.4 shows an overview of the framework. There are two phases in the construction of a dialogue manager: learning and dialogue. In the learning phase, the dialogue manager learns the optimal policy function. In the dialogue phase, the dialogue manager responds to a user by following the trained policy and its belief update.

The random variables involved at a dialogue turn t are as follows:

- $s \in I_s$: user state

User intention.

- $a \in K$: system action

  Module that the system selects.

- $o \in I_s$: observation

  Observed user state, including ASR and intention analysis errors.

- $b_{s_i} = P(s_i|o^{1:t})$: belief

  Stochastic variable of the user state.

- π: policy function

  This function determines a system action a given a belief of user b. π^* is the optimal policy function that is acquired by the learning.

- r: reward function

  This function gives a reward to a pair of the user state s and the system action a.

The aim of the statistical dialogue management is to output an optimal system action $\hat{a}^t$ given a sequence of observation $o^{1:t}$ from 1 to t time-steps and a sequence of previous actions $a^{1:t-1}$ from 1 to $(t-1)$.

Next, we give the belief update that includes the observation and state transition function. The probability of the user state s_i given an observation sequence t $o^{1:t}$ from 1 to t with confidence $x^{1:t}$ and action sequence $a^{1:t-1}$ from 1 to $(t-1)$ is denoted by,

$$b^t_{s_i} = P(s_i|o^{1:t}; x^{1:t}, a^{1:t-1}), \tag{5.7}$$

and referred to as "belief". To avoid clutter, we will usually omit $x^{1:t}$. We can obtain the following update equation from $b^t_{s_i}$ to $b^{t+1}_{s'_j}$:

$$\begin{aligned}
b^{t+1}_{s'_j} &= P(s'_j|o^{1:t+1}, a^{1:t}) && (5.8)\\
&= \frac{P(s'_j, o^{t+1}, a^t|o^{1:t}, a^{1:t-1})}{P(o^{t+1}, a^t|o^{1:t}, a^{1:t-1})} && (5.9)\\
&\propto \sum_{s_i} P(o^{t+1}, s'_j, a^t|s_i) P(s_i|o^{1:t}, a^{1:t-1}) \\
&\propto \sum_{s_i} P(o^{t+1}, s'_j, a^t|s_i) b_{s_i}, && (5.10)
\end{aligned}$$

Then, by introducing the previous system action $a^t = a_k$ that is determined by a policy function, we can rewrite $P(o^{t+1}, s'_j, a^t|s_i)$ in Eqn. (5.10) as follows:

$$\begin{aligned} P(o^{t+1}, s'_j, a_k|s_i) &= P(o^{t+1}, s'_j|s_i, a_k)\underbrace{P(a_k|s_i)}_{\textbf{policy}} \\ &= P(o^{t+1}, s'_j|s_i, \hat{a_k}) \\ &= P(o^{t+1}|s'_j, s_i, \hat{a_k})P(s'_j|s_i, \hat{a_k}). \end{aligned} \tag{5.11}$$

where $P(a_k|s_i)$ is replaced by the policy function in the POMDP. As the observation model is independent from factors of previous time-step of s_i and $\hat{a_k}$, we can omit s_i and $\hat{a_k}$ in the first member and rewrite the Eqn. (5.11) as,

$$P(o^{t+1}|s'_j, s_i, \hat{a_k})P(s'_j|s_i, \hat{a_k}) \approx P(o^{t+1}|s'_j)P(s'_j|s_i, \hat{a_k}). \tag{5.12}$$

We rewrite Eqn. (5.10) with Eqn. (5.12) as follows:

$$b^{t+1}_{s'_j} \propto \underbrace{P(o^{t+1}|s'_j)}_{\textbf{Obs.1}}\sum_{s_i}\underbrace{P(s'_j|s_i, \hat{a_k})}_{\textbf{Trans.1}} b^t_{s_i}. \tag{5.13}$$

Obs.1 is an observation function which is defined in Equation (5.4) and (5.6), and **Trans.1** is a state transition probability of the user state. Once the system estimates the belief b^t, the policy function outputs the optimal action $\hat{a}$ as follows:

$$\hat{a} = \pi^*(b^t). \tag{5.14}$$

5.4.2 Training of POMDP

Q-learning (Monahan, 1982; Watkins and Dayan, 1992) is usually adopted to acquire the optimal policy π^*. Q-learning relies on the estimation of a Q-function, which maximizes the discounted sum of future rewards of the system action a^t at a dialogue turn t given the current belief b^t. Q-learning is performed by iterative updates on the training dialogue data:

$$Q(b^t, a^t) \Leftarrow (1-\varepsilon)Q(b^t, a^t) + \varepsilon[R(s^t, a^t) + \gamma \max_{a^{t+1}} Q(b^{t+1}, a^{t+1})], \tag{5.15}$$

where ε is a learning rate, γ is a discount factor of a future reward. We experimentally decided $\varepsilon = 0.01$ and $\gamma = 0.9$. The optimal policy given by the Q-function is determined as,

$$\pi^*(b^t) = \operatorname*{argmax}_{a^t} Q(b^t, a^t). \tag{5.16}$$

However, it is impossible to calculate the Q-function for all possible real values of belief b. Thus, we train a limited Q-function given by a Grid-based Value Iteration (Bonet, 2002). The belief is given by a function,

$$b_{s_i} = \begin{cases} \eta & \textbf{if} \quad s = i \\ \frac{1-\eta}{|I_s|} & \textbf{if} \quad s \neq i \end{cases}. \tag{5.17}$$

Here, η is a likelihood of $s = i$ that is output of SLU, and 11 discrete points from 0.0 to 1.0 by 0.1 are selected. The case of uniform distribution is also added.

5.4.3 Dialogue Management using User Focus

The proposed POMDP-based dialogue management refers two kinds of belief information: the user intention s and the user focus f (0 or 1). Accordingly, two observations come from SLU: result of user intention analysis o_s and result of user focus detection o_f.

The equation of the belief update (Eqn. (5.13)) is extended by introducing the previous focus f_l and current focus f'_m,

$$b^{t+1}_{s'_j,f'_m} = \underbrace{P(o^{t+1}_s, o^{t+1}_f | s'_j, f'_m)}_{\textbf{Obs.2}} \sum_i \sum_l \underbrace{P(s'_j, f'_m | s_i, f_l, \hat{a_k})}_{\textbf{Trans.2}} b^t_{s_i,f_l}. \tag{5.18}$$

The observation probability is approximated as,

$$\begin{aligned} \textbf{Obs.2} &= P(o^{t+1}_s | o^{t+1}_f, s'_j, f'_m) P(o^{t+1}_f | s'_j, f'_m) \\ &\approx P(o^{t+1}_s | s'_j) P(o^{t+1}_f | f'_m). \end{aligned} \tag{5.19}$$

Here, we assume that information of the user focus f'_m and o^{t+1}_f does not affect the observation of the user state o^{t+1}_s, and the user intention s'_j does not affect the observation of the user focus o^{t+1}_f. These probabilities are calculated from the probability of the user intention analysis (Eqn. (5.4)) and the probability of the user focus detection (Eqn. (5.6)). The resultant trained policy is,

$$\hat{a} = \pi^*(b^t) = \pi^*(\{b^t_{s_i,f_l}\}). \tag{5.20}$$

To train the policy, the equation of Q-learning (Eqn. (5.15)) is modified as,

$$Q(b^t, a^t) \Leftarrow (1-\varepsilon) Q(b^t, a^t) + \varepsilon [R(s^t, f^t, a^t) + \gamma \max_{a^{t+1}} Q(b^{t+1}, a^{t+1})]. \tag{5.21}$$

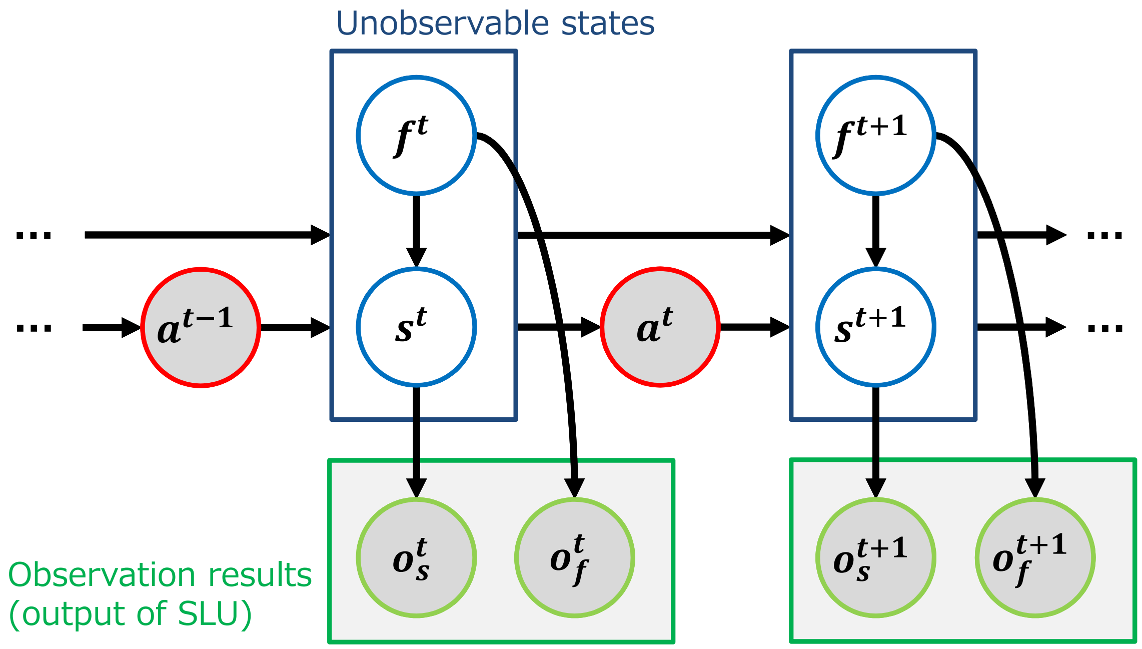


Figure 5.5: Graphical model of the proposed observation model and state transition model.

The state transition probability in Eqn. (5.18) is developed as,

$$\textbf{Trans.2} = \underbrace{P(s'_j, |f'_m, s_i, f_l, \hat{a}_k)}_{\textbf{intention model}} \underbrace{P(f'_m|s_i, f_l, \hat{a}_k)}_{\textbf{focus model}}. \tag{5.22}$$

Thus, the observations o_s and o_f are controlled by hidden states f and s that are determined by the state transition probabilities,

$$\textbf{focus model} = P(f^{t+1}|f^t, s^t, a^t), \tag{5.23}$$

$$\textbf{intention model} = P(s^{t+1}|f^{t+1}, f^t, s^t, a^t). \tag{5.24}$$

A graphical model of the proposed model is shown in **Figure 5.5**. A user simulator is constructed by using the annotated data described in Section 5.3.1.

5.4.4 Definition of Rewards

Definition of rewards is critical in the proposed system. **Table 5.5** defines a reward list at the end of each turn. A reward of +10 is given to appropriate actions, 0 to acceptable actions, and -10 to inappropriate actions. In Table 5.5, pairs of a state and its apparently corresponding action, TP and TP, ST and ST, QA and QA, GR and GR, and II and KS, have positive rewards. Rewards in bold fonts (**+10**) are defined for the following reasons.

Table 5.5: Rewards in each turn.

state	focus	action a						
s	f	TP	ST	QA	PP	GR	KS	CO
TP	0	+10	-10	-10	-10	-10	-10	0
	1							
ST	0	-10	+10	-10	0	-10	-10	0
	1							
QA	0	-10	**+10**	+10	-10	-10	-10	0
	1		-10	**+30**	**+10**			
GR	0	-10	-10	-10	-10	+10	-10	0
	1							
NR	0	**+10**	-10	-10	-10	-10	0	0
	1	-10			**+10**			
II	0	-10	-10	-10	-10	-10	+10	0
	1							

If a user asks a question (QA) without a focus (e.g. "What happened on the game?"), the system can continue by story telling (ST). But when the question has a focus, the system should answer the question (QA), which is highly rewarded (**+30**).

If the system cannot find an answer, it can present relevant information (PP). When the user says nothing (NR), the system action should be determined by considering the user focus; present a new topic if the user is not interested in the current topic (f=0), or present an article related to the dialogue history (f=1). Keeping silence (KS) is a safe action to the user silence (NR), thus, its reward is 0. However, we give 1 frustration point if the system selects KS in this case because the strategy conflicts with the concept of information navigation. Confirmation (CO) is a safe action to every user input, but it also frustrates the user. Thus, the reward of CO is defined as 0 for every intention, but 2 frustration points are given to the system. If the system selects an inappropriate action (action of $r = -10$), 2 frustration points are given to the system. If the frustration points accumulate more than 10, a large penalty -200 is given to the system and the dialogue is terminated. Reward of +200 is given if 20 turns are passed to reward a long continued dialogue.

5.5 Experimental Evaluation

The proposed system is evaluated with two experiments; dialogue state tracking with real users and simulation. For evaluation of the system, additional 626 utterances (12 users, 24

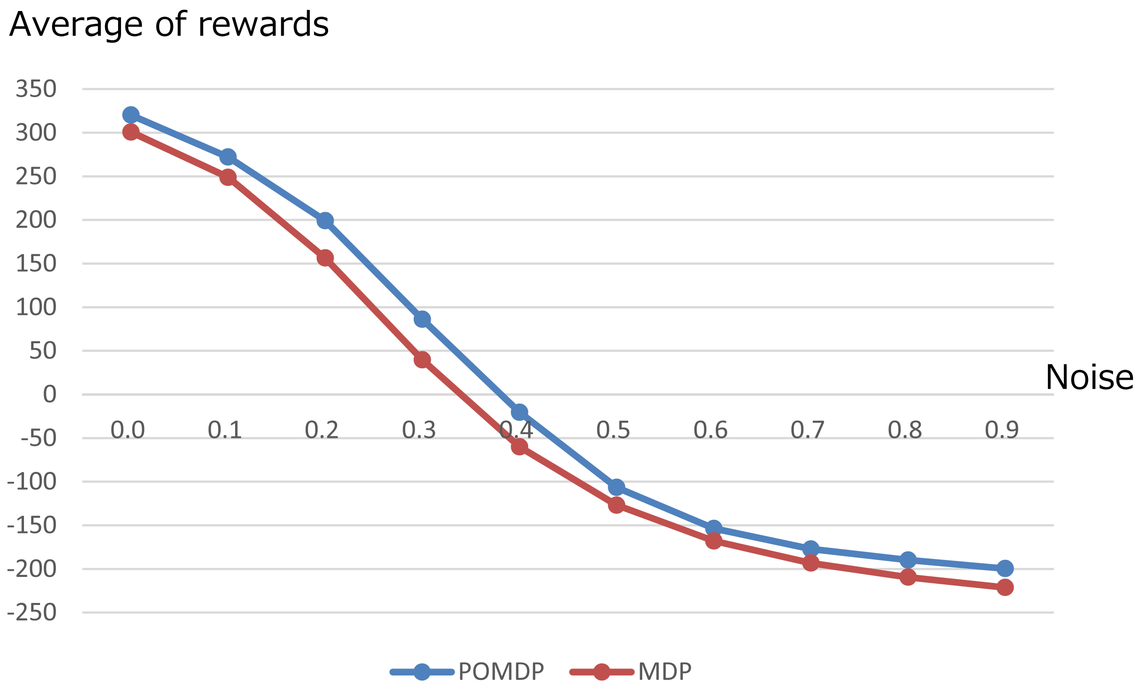


Figure 5.6: Effect of introduction of the stochastic variable (MDP vs. POMDP)

dialogues; 2 dialogues with each user) were collected with the proposed dialogue system. There are 58 cases regarded as no request (*NR*) when the user did not say anything for longer than 5 seconds.

5.5.1 Evaluation of Dialogue Manager with User Simulator

First, the dialogue manager is evaluated with a user simulator that was constructed from the training corpus (Section 5.3.1). In this evaluation, the system calculated average reward of 100,000 dialogue sessions between the system and the user simulator given a fixed noise rate. **Figure 5.6** compares MDP and POMDP. This result shows that the stochastic extension of reinforcement learning is effective for noisy input that includes ASR or NLU errors. **Figure 5.7** shows the effect of the user focus. By introducing the user focus, the system receives higher rewards than the model without the user focus. Especially, the proposed model is more robust with a noise level of 10–30% that spoken dialogue systems often encounter as described in Chapter 3.

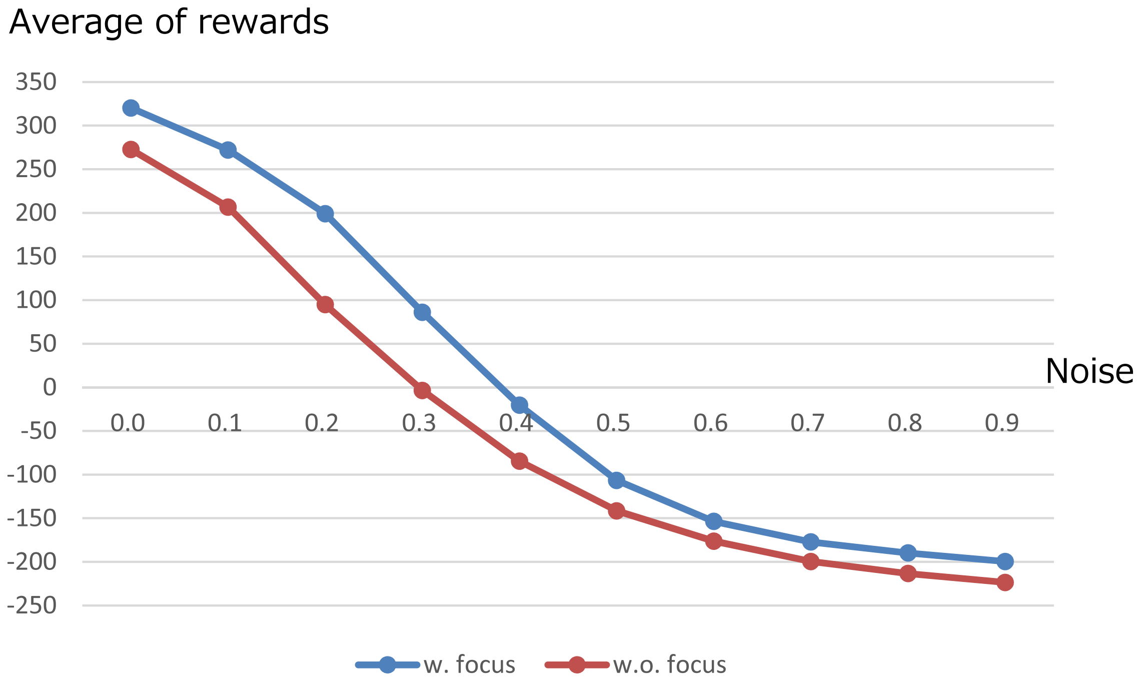


Figure 5.7: Effect of introduction of the user focus in simulation.

5.5.2 Evaluation of Dialogue State Tracking and Action Selection

Dialogue state tracking (DST) is a task of tracking the correct user state with a noisy input (e.g. ASR and SLU errors) (Williams et al., 2013). It tries to maximize the probability of the belief of the correct states, and the accuracy of the 1-best result of the belief update is evaluated. Accuracy of the system action and the average reward for dialogue sessions are also evaluated. The accuracy of the system action shows not only the effect of the belief update but also the effect of the trained policy. There are two baseline systems. The first baseline system is a rule-based dialogue manager that is operated by a score of the question-answering module using the P-A structure (Chapter 4) and regular expressions for TP and GR modules. The other baseline system is operated by the POMDP-based dialogue manger that does not refer to user focus (POMDP w.o. focus).

The DST accuracy, accuracy of action selection, and average reward are summarized in **Table 5.6**. This result shows that the proposed method tracks the dialogue state of the user with high accuracy, and responds with more appropriate modules. A breakdown of the DST accuracy is shown in **Table 5.7**. The table shows precision (P), recall (R) and F-measure (F) of each intention tag. The performance for greeting (*GR*) and irrelevant

Table 5.6: Summary of results comparing with rule-based system and POMDP without user focus.

	Rule	POMDP w.o. focus	POMDP proposed
Accuracy of DST (1-best)	0.786 (=492/626)	0.853 (=534/626)	**0.866** (=542/626)
Accuracy of Actions (1-best)	0.756 (=517/684)	0.763 (=522/684)	**0.860** (=588/684)
Average reward	213.333	242.917	**374.583**

Table 5.7: Performance of dialogue state tracking (DST) (precision, recall and F-measure).

tag	Rule			POMDP w.o. focus			POMDP proposed		
	P	R	F	P	R	F	P	R	F
TP	0.985	0.812	**0.890**	0.912	0.828	0.868	0.951	0.820	0.880
ST	0.500	0.012	0.023	0.906	0.585	0.711	0.887	0.768	**0.824**
QA	0.696	0.990	0.818	0.823	0.963	0.888	0.842	0.946	**0.891**

Table 5.8: SCount of tag of user intention s.

tag	count
TP	239
ST	82
QA	299
GR	2
II	4
All	626
NR	58
All (+*NR*)	684

Table 5.9: Count of tag of system action a.)

tag	count
TP	259
ST	90
QA	290
PP	38
GR	2
KS	4
All	684

input (*II*) is not shown because the number of these tags was very small (#*GR*=2, #*II*=4). The count of each tag of user intention s and system action a are shown in **Table 5.8**, **5.9**. The proposed framework improved the SLU accuracy and robustness against ASR errors,

Table 5.10: Performance of action selection (precision, recall and F-measure).

tag	Rule			POMDP w.o. focus			POMDP proposed		
	P	R	F	P	R	F	P	R	F
TP	0.884	0.822	0.852	0.917	0.764	0.834	0.959	0.803	**0.874**
ST	1.000	0.022	0.043	0.900	0.500	0.643	0.910	0.789	**0.845**
QA	0.678	0.993	0.806	0.797	0.962	0.872	0.843	0.945	**0.891**
PP	0.929	0.342	0.500	0.000	0.000	0.000	0.854	0.921	**0.886**

especially reducing misclassification of question answering (*QA*) that were actually topic presentation (*TP*) or story telling (*ST*). Moreover, the belief update can detect the *ST* state even if the SLU incorrectly predicts *QA* or *TP*.

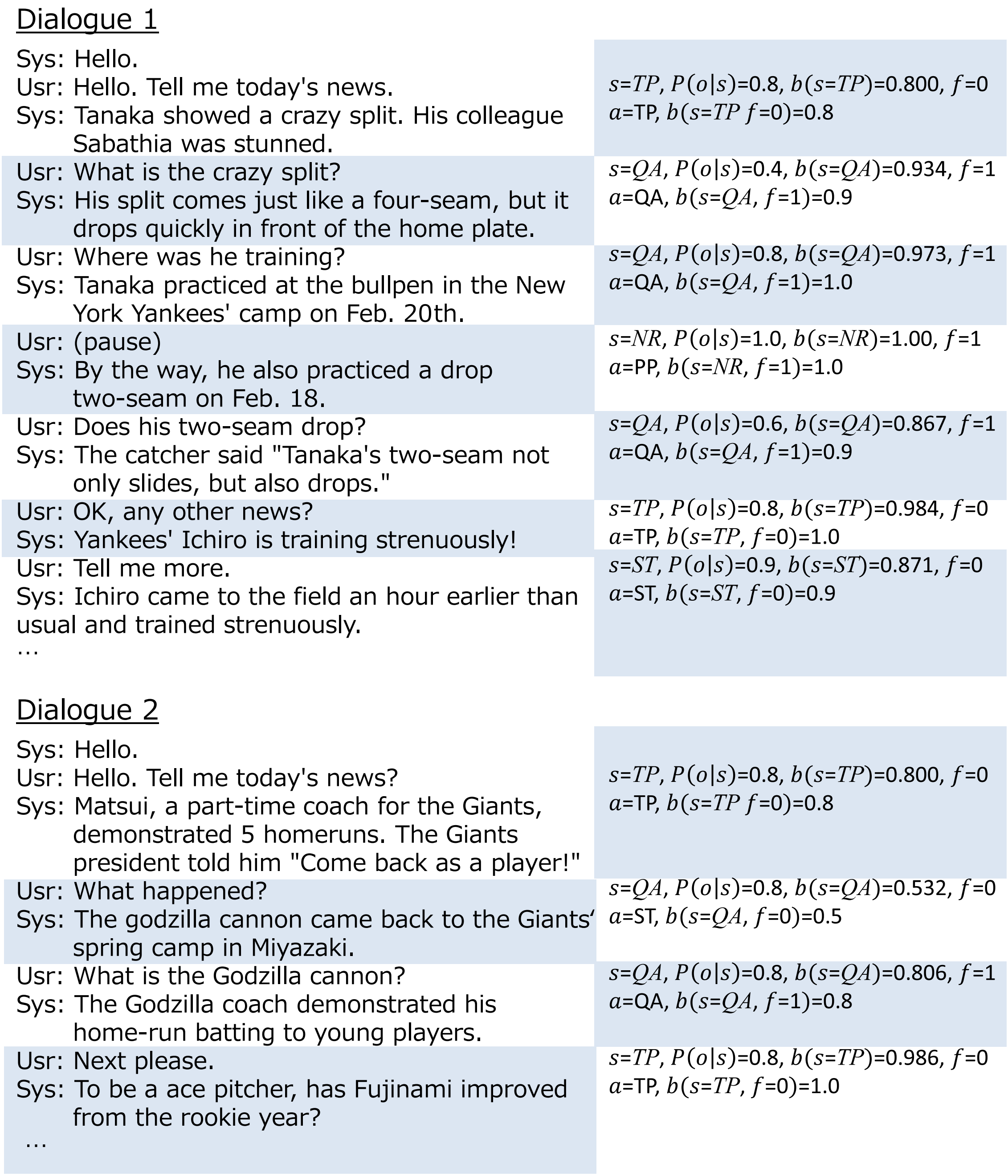
The proposed method also improved the accuracy of action selection. A breakdown is shown in **Table 5.10**. Here, the results of TP, ST, QA and PP are presented because the number of KS and GR was very small (#GR=2, #KS=4), and CO was not labeled as a correct action. The proposed method outperformed the baseline systems for all actions. The proposed method improved the accuracy for topic presentation (TP) and proactive presentation (PP) especially when the user intention was no request (*NR*). The POMDP without user focus always selected the keep silence (KS) module if the user said nothing (*NR*).

The proposed method also made more effective confirmations (CO) when the SLU result was not correct. It made confirmations (CO) 18 times, and 15 times of them was done when the SLU result was incorrect (15/18=83.3%). The POMDP without user focus made only two confirmations, when the detected user intention was correct (0/2=0.0%).

The proposed method made 35 proactive presentations (PP), and 17 times of them (17/35=48.6%) invoked new user questions. This result demonstrates that the proposed system encouraged interactions in news navigation.

5.5.3 Discussion of Trained Policy

An example dialogue is shown in **Figure 5.8**. In the example, the system selects appropriate actions even if the observation likelihood is low. At the 4th turn of Dialogue 1 in this example, the system with the user focus responds with an action of proactive presentation a=PP, but the system without the user focus responds with an action of topic presentation a=TP. At the 2nd turn of Dialogue 2, the user asks a question without a focus. The



Dialogue 1

Sys: Hello. Usr: Hello. Tell me today's news. Sys: Tanaka showed a crazy split. His colleague Sabathia was stunned.	s=TP, $P(o\|s)$=0.8, $b(s$=$TP)$=0.800, f=0 a=TP, $b(s$=$TP\ f$=0)=0.8
Usr: What is the crazy split? Sys: His split comes just like a four-seam, but it drops quickly in front of the home plate.	s=QA, $P(o\|s)$=0.4, $b(s$=$QA)$=0.934, f=1 a=QA, $b(s$=QA, f=1)=0.9
Usr: Where was he training? Sys: Tanaka practiced at the bullpen in the New York Yankees' camp on Feb. 20th.	s=QA, $P(o\|s)$=0.8, $b(s$=$QA)$=0.973, f=1 a=QA, $b(s$=QA, f=1)=1.0
Usr: (pause) Sys: By the way, he also practiced a drop two-seam on Feb. 18.	s=NR, $P(o\|s)$=1.0, $b(s$=$NR)$=1.00, f=1 a=PP, $b(s$=NR, f=1)=1.0
Usr: Does his two-seam drop? Sys: The catcher said "Tanaka's two-seam not only slides, but also drops."	s=QA, $P(o\|s)$=0.6, $b(s$=$QA)$=0.867, f=1 a=QA, $b(s$=QA, f=1)=0.9
Usr: OK, any other news? Sys: Yankees' Ichiro is training strenuously!	s=TP, $P(o\|s)$=0.8, $b(s$=$TP)$=0.984, f=0 a=TP, $b(s$=TP, f=0)=1.0
Usr: Tell me more. Sys: Ichiro came to the field an hour earlier than usual and trained strenuously. …	s=ST, $P(o\|s)$=0.9, $b(s$=$ST)$=0.871, f=0 a=ST, $b(s$=ST, f=0)=0.9

Dialogue 2

Sys: Hello. Usr: Hello. Tell me today's news? Sys: Matsui, a part-time coach for the Giants, demonstrated 5 homeruns. The Giants president told him "Come back as a player!"	s=TP, $P(o\|s)$=0.8, $b(s$=$TP)$=0.800, f=0 a=TP, $b(s$=$TP\ f$=0)=0.8
Usr: What happened? Sys: The godzilla cannon came back to the Giants' spring camp in Miyazaki.	s=QA, $P(o\|s)$=0.8, $b(s$=$QA)$=0.532, f=0 a=ST, $b(s$=QA, f=0)=0.5
Usr: What is the Godzilla cannon? Sys: The Godzilla coach demonstrated his home-run batting to young players.	s=QA, $P(o\|s)$=0.8, $b(s$=$QA)$=0.806, f=1 a=QA, $b(s$=QA, f=1)=0.8
Usr: Next please. Sys: To be a ace pitcher, has Fujinami improved from the rookie year? …	s=TP, $P(o\|s)$=0.8, $b(s$=$TP)$=0.986, f=0 a=TP, $b(s$=TP, f=0)=1.0

Figure 5.8: A dialogue example.

confidence of s=QA is lowered by the belief update, and the system selects the story telling module a=ST. These examples show that the trained policy reflects the design described in Section 5.4.4. It is better to make a proactive presentation when the user is interested in the topic.

5.6 Conclusions

This chapter addressed the dialogue management for information navigation of Web news articles updated day-by-day. The system presents relevant information according to the user's interest by tracking the user focus. A POMDP framework is extended to track the user focus to select the appropriate action module. In the experimental evaluations, the proposed dialogue management determines the state of the user more accurately than the existing rule-based system and the POMDP-based system without focus information. The proposed method also improved the accuracy of action selection.

Chapter 6

Conclusions

This thesis addressed a spoken dialogue system that navigates information. The task of information navigation is a direction of spoken dialogue systems from conventional task-oriented dialogue systems to general non-task-oriented dialogue systems. In information navigation, users are not forced to accommodate the task goal of the system. Instead, the user can make ambiguous queries. The system provides information that the user wants to know by probing and clarifying the potential demands of the user.

The proposed information navigation system can converse with users in a user-friendly manner. It does not respond "I can't answer the question", or turns to the Web search even if it cannot find exact information. The system responds with a partially matched information to the user query and proactively presents related information by following the dialogue context and attentional state of the user even if the user demand is not clear or the user query is ambiguous or the user does not ask a question.

The framework is based on tracking of the semantic and dialogue structure of the conversation that is statistically trained with machine learning. The learning is conducted in an unsupervised and domain-independent manner. The system does not require additional annotation of the data. The task structure and domain knowledge are separated. The domain knowledge is trained from unannotated data that only have a tag of domain.

First, semantic significance score is defined based on P-A structure with a Naive Bayes method. The method only requires the domain tag of text that is generally annotated to the newspaper articles, envisaged knowledge source to information navigation system. The significance score successfully extracted important P-A structure patterns in a domain, and the score is used as a domain knowledge through this thesis.

For the ASR module of the spoken dialogue system, an appropriate language model is constructed using this knowledge. As Web resources such as the wisdom of crowds often include inappropriate sentences. The proposed method selects well-matched sentence for the domain based on semantic relevance measure. The selected sentences by the proposed system match the user query of the information navigation, and the system improved not only word-level ASR accuracy but also semantic-level accuracy.

For response generation, a flexible information retrieval is designed. The module presents relevant information to the query even if the system cannot find exactly matched information to the user query based on the semantic relevance measure. The proactive presentation module is also designed. The module presents relevant information proactively even if the user does not express their information demands.

For the dialogue management, POMDP is extended to the task of information navigation. The dialogue manager tracks the user focus to probe and satisfy potential user demands. The user focus is incorporated into the belief update of POMDP, and resultant system successfully outputs appropriate actions for information navigation. The reward function of POMDP is defined in the quality of system actions and also for encouraging long interactions.

The organization of this thesis is summarized in **Figure 6.1**. The system assumes that the system can use domain text and users input with spoken languages. Domain dependent P-A patterns (significance score of domain) are extracted from target source based on the method proposed in Chapter 2. The patterns are used in the selection of training data for language model of ASR in Chapter 3. The patterns are also used for information retrieval of dialogue module that presents relevant information as designed in Chapter 4. The modules are controlled by the dialogue manager that tracks dialogue state and user focus as proposed in Chapter 5. Finally, the system responds the user with speech output that is generated from the dialogue module.

6.1 Domain Portability of the Proposed System

The proposed system has an architecture applicable to a variety of domains. The system can be adapted to not only domains in newspaper articles, but also domains in general Web sources such as Wikipedia. The requirement of the architecture is that the knowledge

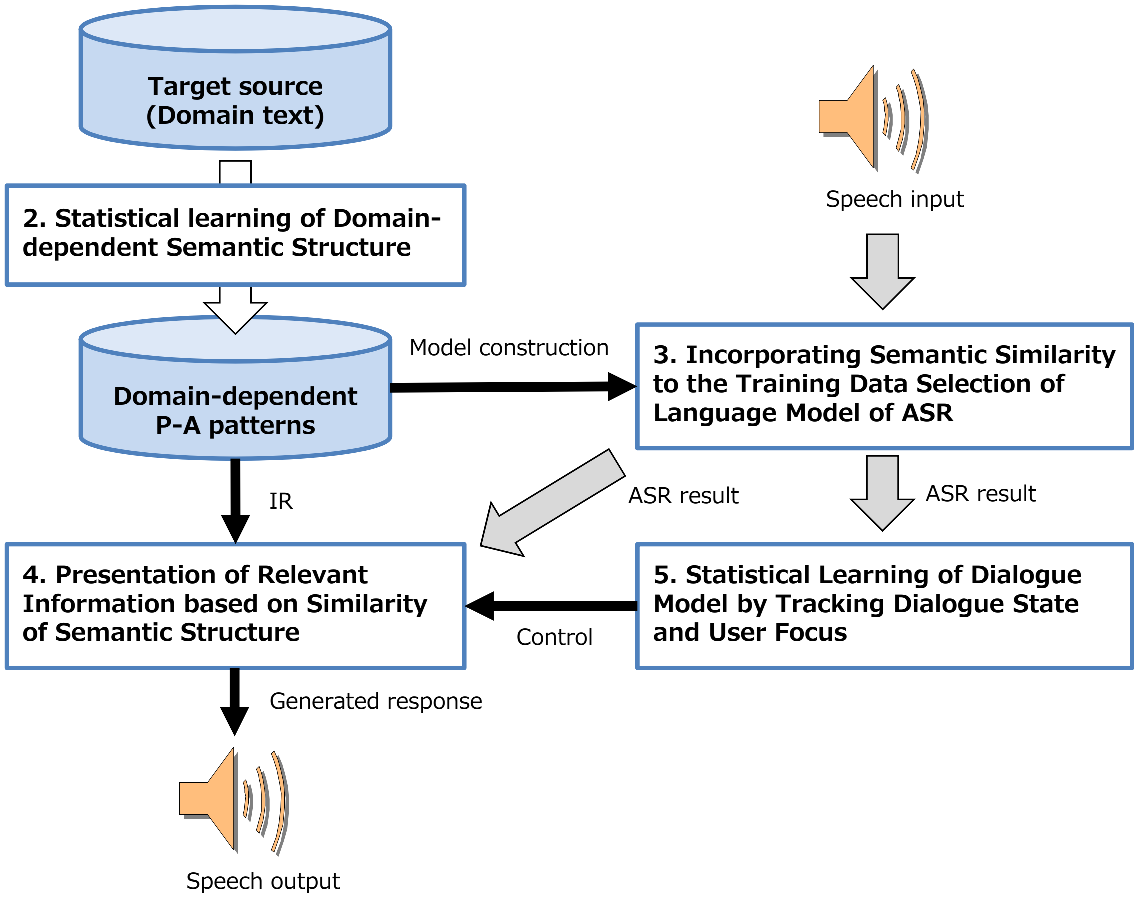

Target source
(Domain text)
2. Statistical learning of Domain-dependent Semantic Structure
Domain-dependent
P-A patterns
Model construction
Speech input
3. Incorporating Semantic Similarity to the Training Data Selection of Language Model of ASR
IR
ASR result
ASR result
4. Presentation of Relevant Information based on Similarity of Semantic Structure
5. Statistical Learning of Dialogue Model by Tracking Dialogue State and User Focus
Control
Generated response
Speech output

Figure 6.1: *Organization of this thesis.*

base text has a tag of a domain or any classified tag. Then, necessary domain knowledge of semantic and dialogue structure is learned in an unsupervised manner.

The framework is based on a statistical learning manner to enhance a portability of the spoken dialogue system. Every module is designed in an unsupervised manner and domain-independent concept. However, the performance of domain adaptation depends on the performance of basic natural language processing such as morphological analysis, dependency parsing and P-A structure analysis. Especially, the problem of out-of-vocabulary of named entities degrades the accuracy of P-A structure analysis. To cope with this problem, a dictionary of named entities of the domain is required. This kind of dictionary can be easily collected from Web resources.

6.2 Future Directions

There are some future directions for suitable information navigation. Incorporating non-verbal information is one of the important features in information navigation (Nakano and Ishii, 2010; Kimura et al., 2013). When we reflect conversations with human concierges, they understand the potential demands of the customer from not only verbal information but also other behaviors such as back-channels, gazes or other physical behaviors. The use of multi-modal information will improve the capacity of the system. The proposed system can be extended to use such variety of features.

Another direction is personalization. Adapting to the personal preference is one of the easiest ways for the system to be user friendly. Recently, rapid policy optimization is widely applied to the POMDP-based dialogue manager to adapt the system to a personality (Paquet et al., 2005; Gašić et al., 2010; Jurcicek et al., 2010; Daubigney et al., 2012). On the other hand, content-based personalization is widely investigated in the area of information retrieval. These techniques will improve the information navigation of spoken dialogue systems and enhance the information access of a variety of users.

Bibliography

Murat Akbacak, Yuqing Gao, Liang Gu, and Hong-Kwang Jeff Kuo. 2005. Rapid transition to new spoken dialogue domains: Language model training using knowledge from previous domain applications and web text resources. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2005, pages 1873–1876.

Tomoyosi Akiba and Hiroyuki Abe. 2005. Exploiting passage retrieval for n-best rescoring of spoken questions. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2005, pages 65–68.

Jacob Aron. 2011. How innovative is apple's new voice assistant, Siri? *New Scientist*, 212(2836):24.

Harald Aust, Martin Oerder, Frank Seide, and Volker Steinbiss. 1995. The philips automatic train timetable information system. *Speech Communication*, 17(3-4):249–262. Interactive Voice Technology for Telecommunication Applications.

Dan Bohus and Alexander I. Rudnicky. 2003. Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Proceedings of European Conference on Speech Communication and Technology*, EUROSPEECH 2003, pages 597–600.

Blai Bonet. 2002. An e-optimal grid-based algorithm for partially observable Markov decision processes. In *Proceedings of International Conference on Machine Learning*, ICML 2002, pages 51–58.

Michael E. Bratman, David J. Israel, and Martha E. Pollack. 1988. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4(3):349–355.

Ivan Bulyko, Mari Ostendorf, and Andreas Stolcke. 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, HLT-NAACL 2003, pages 7–9.

Ivan Bulyko, Mari Ostendorf, Manhung Siu, Tim Ng, Andreas Stolcke, and Özgür Çetin. 2007. Web resources for language modeling in conversational speech recognition. *ACM Transaction on Speech and Language Processing*, 5(1):1–25.

Robin D Burke, Kristian J Hammond, Vladimir Kulyukin, Steven L Lytinen, Noriko Tomuro, and Scott Schoenberg. 1997. Question answering from frequently asked question files: Experiences with the faq finder system. *AI Magazine*, 18(2):57–66.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. In *Proceedings of the Workshop on Human Language Technology*, HLT 1994, pages 43–48.

Lucie Daubigney, Matthieu Geist, and Olivier Pietquin. 2012. Off-policy learning in large-scale POMDP-based dialogue systems. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2012, pages 4989–4992.

Myroslava O Dzikovska, James F Allen, and Mary D Swift. 2003. Integrating linguistic and domain knowledge for spoken dialogue systems in multiple domains. In *Proceedings of International Joint Conference on Artificial Intelligence Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, IJCAI-WS 2003.

Charles J. Fillmore. 1968. The case for case. In *Universals in Linguistic Theory.*

Alexander Mark Franz, Monika H Henzinger, Sergey Brin, and Brian Christopher Milch. 2006. Voice interface for a search engine.

Milica Gašić, Filip Jurčíček, Simon Keizer, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Gaussian processes for fast policy optimisation of POMDP-based dialogue managers. In *Proceedings of Annual SIGdial Meeting on Discourse and Dialogue*, SIGDIAL 2010, pages 201–204.

Milica Gašić, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013. POMDP-based dialogue manager adaptation to extended domains. In *Proceedings of Annual SIGdial Meeting on Discourse and Dialogue*, SIGDIAL 2013, pages 214–222.

Ralph Grishman. 2003. Discovery methods for information extraction. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, ISCA-WS 2003, pages 243–247.

Sanda Harabagiu, Andrew Hickl, John Lehmann, and Dan Moldovan. 2005. Experiments with interactive question-answering. In *Proceedings of Annual Meeting on Association for Computational Linguistics*, ACL 2005, pages 205–214.

Zellig S Harris. 1951. *Methods in structural linguistics.*

Ryuichiro Higashinaka, Katsuhito Sudoh, and Mikio Nakano. 2006. Incorporating discourse features into confidence scoring of intention recognition results in spoken dialogue systems. *Speech Communication*, 48(3):417–436.

Takatsugu Hirayama, Yasuyuki Sumi, Tatsuya Kawahara, and Takashi Matsuyama. 2011. Info-concierge: Proactive multi-modal interaction through mind probing. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, APSIPA 2011.

Lichan Hong, Shigeru Muraki, Arie Kaufman, Dirk Bartz, and Taosong He. 1997. Virtual voyage: Interactive navigation in the human colon. In *Proceedings of Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH 1997, pages 27–34.

Filip Jurcicek, Blaise Thomson, Simon Keizer, Francois Mairesse, Milica Gašić, Kai Yu, and Steve Young. 2010. Natural belief-critic: a reinforcement algorithm for parameter estimation in statistical spoken dialogue systems. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2010.

Boris Katz and Jimmy Lin. 2002. Annotating the semantic web using natural language. In *Proceedings of Workshop on NLP and XML - Volume 17*, NLPXML 2002, pages 1–8.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL 2006, pages 176–183.

Daisuke Kawahara, Nobuhiro Kaji, and Sadao Kurohashi. 2002. Question and answering system based on predicate-argument matching. In *Proceedings of NTCIR Workshop on Research Information Retrieval, Automatic Text Summarization andQuestion Answering*, NTCIR 2002.

Tatsuya Kawahara, Yusuke Nemoto, and Yuya Akita. 2008. Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2008, pages 4929–4932.

Tatsuya Kawahara. 2009. New perspectives on spoken language understanding: Does machine need to fully understand speech? In *IEEE Workshop on Automatic Speech Recognition & Understanding*, ASRU 2009, pages 46–50.

Akisato Kimura, Ryo Yonetani, and Takatsugu Hirayama. 2013. Computational models of human visual attention and their implementations: A survey. *IEICE Transactions on Information and Systems*, 96(3):562–578.

Yoji Kiyota, Sadao Kurohashi, and Fuyuko Kido. 2002. "dialog navigator" : A question answering system based on large text knowledge base. In *Proceedings of Proceedings of International Conference on Computational Linguistics*, COLING 2002, pages 460–466.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for n-gram language modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 1995, pages 181–184.

Kazunori Komatani and Tatsuya Kawahara. 2000. Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output. In *Proceedings of International Conference on Computational Linguistics*, COLING 2000, pages 467–473.

Kazunori Komatani, Katsuaki Tanaka, Hiroaki Kashima, and Tatsuya Kawahara. 2001. Domain-independent spoken dialogue platform using key-phrase spotting based on combined language model. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2001, pages 1319–1322.

Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G Okuno. 2005. User modeling in spoken dialogue systems to generate flexible guidance. *User Modeling and User-Adapted Interaction*, 15(1-2):169–183.

Kazunori Komatani, Naoyuki Kanda, Mikio Nakano, Kazuhiro Nakadai, Hiroshi Tsujino, Tetsuya Ogata, and Hiroshi G. Okuno. 2006. Multi-domain spoken dialogue system with extensibility and robustness against speech recognition errors. In *Proceedings of SIGdial Workshop on Discourse and Dialogue*, SIGDIAL 2009, pages 9–17.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Julian Kupiec. 1993. Murax: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 1993, pages 181–190.

Lori Lamel, Samir Bennacef, Jean-Luc Gauvain, Herve Dartigues, and Jean-Noel Temem. 2002. User evaluation of the mask kiosk. *Speech Communication*, 38(1-2):131–139.

Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. 2001. Julius–an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology*, EUROSPEECH, pages 1691–1694.

Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. 2009. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, 51(5):466–484.

Donghyeon Lee, Minwoo Jeong, Kyungduk Kim, and Gary Geunbae Lee. 2012. Unsupervised modeling of user actions in a dialog corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2012, pages 5061–5064.

Oliver Lemon, Xingkun Liu, Daniel Shapiro, and Carl Tollander. 2006. Hierarchical reinforcement learning of dialogue policies in a development environment for dialogue systems: Reall-dude. In *Proceedings of Workshop on the Semantics and Pragmatics of Dialogue*, SEMDIAL 2006, pages 185–186.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1):11–23.

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. 2006.

Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18:1138–1150.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, ACL-COLING 1998, pages 768–774.

Bruce Lucas. 2000. Voicexml. *Communications of the ACM*, 43(9):53.

Ryo Masumura, Seongjun Hahm, and Akinori Ito. 2011. Training a language model using webdata for large vocabulary japanese spontaneous speech recognition. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2011, pages 1465–1468.

Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. 2010. Controlling listening-oriented dialogue using partially observable markov decision processes. In *Proceedings of International Conference on Computational Linguistics*, COLING 2010, pages 761–769.

Teruhisa Misu and Hideki Kashioka. 2012. Simultaneous feature selection and parameter optimization for training of dialog policy by reinforcement learning. In *IEEE International Workshop Spoken Language Technology Workshop*, IWSLT 2012, pages 1–6.

Teruhisa Misu and Tatsuya Kawahara. 2006. A bootstrapping approach for developing language model of new spoken dialogue system by selecting web texts. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2006, pages 9–13.

Teruhisa Misu and Tatsuya Kawahara. 2010. Bayes risk-based dialogue management for document retrieval system with speech interface. *Speech Communication*, 52(1):61–71.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, ACL 2008, pages 236–244.

George E. Monahan. 1982. State of the art? a survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):1–16.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, ACL 2010, pages 220–224.

Cosmin Munteanu, Gerald Penn, and Ron Baecker. 2007. Web-based language modelling for automatic lecture transcription. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2007, pages 2353–2356.

Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of International Conference on Intelligent User Interfaces*, IUI 2010, pages 139–148.

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of Proceedings of International Conference on Computational Linguistics*, COLING 2004, pages 693–701.

Tim Ng, Mari Ostendorf, Mei-Yuh Hwang, Man-Hung Siu, Ivan Bulyko, and Xin Lei. 2005. Web-data augmented language models for mandarin conversational speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2005, pages 589–592.

Ryuichi Nishimura, Kumiko Komatsu, Yuka Kuroda, Kentaro Nagatomo, Akinobu Lee, Hiroshi Saruwatari, and Kiyohiro Shikano. 2001. Automatic n-gram language model creation from web resources. In *Proceedings of European Conference on Speech Communication and Technology*, EUROSPEECH 2001, pages 5181–5184.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Yi-Cheng Pan, Hung yi Lee, and Lin shan Lee. 2012. Interactive spoken document retrieval with suggested key terms ranked by a Markov decision process. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):632–645.

Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, EMNLP 2009, pages 938–947.

Sébastien Paquet, Ludovic Tobin, and Brahim Chaib-draa. 2005. An online POMDP algorithm for complex multiagent environments. In *Proceedings of International Joint*

Conference on Autonomous Agents and Multiagent Systems, AAMAS 2005, pages 970–977.

Jim Pitman. 1995. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158.

Lance A. Ramshaw and Ralph M. Weischedel. 2005. Information extraction. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5 of *ICASSP 2005*, pages 969–972.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let' s go public! taking a spoken dialog system to the real world. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2005.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, pages 583–593.

Nicholas Roy, Joelle Pineau, and Sebastian Thrun. 2000. Spoken dialogue management using probabilistic reasoning. In *Proceedings of Annual Meeting on Association for Computational Linguistics*, ACL 2000, pages 93–100.

David Sadek. 1999. Design considerations on dialogue systems: From theory to technology-the case of Artimis. In *Proceedings of ESCA Tutorial and Research Workshop on Interactive Dialogue in Multi-Modal Systems*, ETRW-IDS 1999, pages 173–187.

Ruhi Sarikaya, Agustin Gravano, and Yuqing Gao. 2005. Rapid language model development using external resources for new spoken dialog domains. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2005, pages I–573–I–576.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

Stephanie Seneff, Edward Hurley, Raymond Lau, Christine Pao, Philipp Schmid, and Victor Zue. 1998. Galaxy-ii: a reference architecture for conversational system development. In *Proceedings of International Conference on Spoken Language Processing*, ICSLP 1998, pages 931–934.

Abhinav Sethy, Panayiotis G. Georgiou, and Shrikanth Narayanan. 2005. Building topic specific language models from webdata using competitive models. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2005, pages 1293–1296.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL 2007, pages 12–21.

Tomohide Shibata, Michitaka Odani, Jun Harashima, Takashi Oonishi, and Sadao Kurohashi. 2008. Syngraph: A flexible matching method based on synonymous expression extraction from an ordinary dictionary and a web corpus. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, IJCNLP 2008, pages 787–792.

Tomohide Shibata, Yusuke Egashira, and Sadao Kurohashi. 2014. Chat-like conversational system based on selection of reply generating module with reinforcement learning. In *Proceedings of International Workshop Series on Spoken Dialog Systems*, IWSDS 2014, pages 124–129.

Matthijs TJ Spaan and Nikos Vlassis. 2005. Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research*, 24(1):195–220.

Motoyuki Suzuki, Yasutomo Kajiura, Akinori Ito, and Shozo Makino. 2006. Unsupervised language model adaptation based on automatic text collection from WWW. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2006, pages 2202–2205.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476).

Yee Whye Teh. 2006a. A bayesian interpretation of interpolated kneser-ney.

Yee Whye Teh. 2006b. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*, ACL-COLING 2006, pages 985–992.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, ACL 2010, pages 948–957.

Blaise Thomson and Steve Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588.

Andreas Tsiartas, Panayiotis Georgiou, and Shrikanth Narayanan. 2010. Language model adaptation using www documents obtained by utterance-based queries. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2010, pages 5406–5409.

Gokhan Tur, Umit Guz, and Dilek Hakkani-Tur. 2006. Model adaptation for dialog act tagging. In *Proceedings of IEEE workshop on Spoken Language Technology*, IWSLT 2006, pages 94–97.

Sebastian Varges, Giuseppe Riccardi, Silvia Quarteroni, and Alexei V Ivanov. 2011. POMDP concept policies and task structures for hybrid dialog management. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2011, pages 5592–5595.

Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of Annual Meeting on Association for Computational Linguistics*, ACL 2001, pages 515–522.

Vincent Wan and Thomas Hain. 2006. Strategies for language model web-data collection. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2006, pages 1069–1072.

Rui Wang and Yi Zhang. 2009. Recognizing textual relatedness with predicate-argument structure. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, EMNLP 2009, pages 784–792.

Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning*, 8(3):279–292.

Graham Wilcock and Kristiina Jokinen. 2013. Wikitalk human-robot interactions. In *Proceedings of ACM on International Conference on Multimodal Interaction*, ICMI 2013, pages 73–74.

Graham Wilcock. 2012. Wikitalk: A spoken wikipedia-based open-domain knowledge access system. In *COLING-2012 Workshop on Question Answering for Complex Domains*, COLING-WS 2012, pages 57–69.

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

Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of Annual SIGdial Meeting on Discourse and Dialogue*, SIGDIAL 2013, pages 404–413.

Jason D. Williams. 2008. The best of both worlds: Unifying conventional dialog systems and POMDPs. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH 2008, pages 1173–1176.

Dekai Wu and Pascale Fung. 2009. Can semantic role labeling improve SMT? In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, EAMT 2009, pages 218–225.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Xiaojin Zhu and Ronald Rosenfeld. 2001. Improving trigram language modeling with the world wide web. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2001, pages 533–536.

Victor Zue, Stephanie Seneff, James R. Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen, and Lee Hetherington. 2000. Jupiter: a telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96.

Appendix A

Dirichlet Smoothing based on Hierarchical Dirichlet Process

The appendix shows the detail of Dirichlet smoothing based on Hierarchical Dirichlet Process that is used in Chapter 2. First, the appendix introduces the aspect of Dirichlet smoothing based on the Dirichlet distribution that is a prior distribution of Multinomial distribution. Second, the appendix extends the Dirichlet distribution to the Dirichlet process by introducing the Chinese Restaurant Process (CRP), a typical expression of Dirichlet process and its hierarchization.

Dirichlet Smoothing

Dirichlet smoothing is for multinomial distribution based on Dirichlet distribution. The Dirichlet distribution is a prior distribution of multinominal distribution that has the same size of dimensions. A large number of probability calculations in natural language processing are based on the multinomial distribution. A Multinomial distribution consists of K-dimensional M from m_1 to m_K. When the observed counts of them are n_1 to n_K, the probability density function of the multinomial distribution is defined as,

$$P(n|\theta) = \frac{\Gamma(\sum_{i=1}^{K} n_i + 1)}{\prod_{i=1}^{K} \Gamma(n_i + 1)} \prod_{i=1}^{K} \theta_i^{n_i}, \tag{A.1}$$

where the parameters θ_i (probability of m_i) satisfy the constraint,

$$\sum_{i=1}^{K} \theta_i = 1. \tag{A.2}$$

At maximum a posteriori (MAP) estimation, θ is estimated by,

$$P(\theta|n) \propto P(n|\theta)P(\theta), \tag{A.3}$$

with the Bayes theory. $P(n|\theta)$ is a likelihood that is estimated by the maximum likelihood estimation and $P(\theta)$ is a prior distribution of probability. By introducing the Dirichlet distribution, a conjugate prior to multinomial distribution, we rewrite the Eqn. (A.3) as,

$$P(\theta|n,\alpha) \propto P(n|\theta)P(\theta|\alpha). \tag{A.4}$$

Here, α (set of $\alpha_{1:K}$) are hyper parameters. Usually, they take the same value. The prior Dirichlet distribution is given as,

$$\begin{aligned} P(\theta|\alpha) &= \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} && \text{(A.5)}\\ &= \frac{1}{Z}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}. && \text{(A.6)} \end{aligned}$$

α_k is a parameter of the prior probability of m_k and Z is a normalization that is calculated from only α. Eqn. (A.1) and Eqn. (A.6) are conjugate. By Eqn. (A.1) and Eqn. (A.6), Eqn. (A.4) is rewritten as,

$$\begin{aligned} P(n|\theta)P(\theta|\alpha) &= \frac{\Gamma(\sum_{k=1}^{K}n_k+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)}\prod_{k=1}^{K}\theta_k^{n_k}\frac{1}{Z}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} && \text{(A.7)}\\ &= \underbrace{\frac{\Gamma(\sum_{k=1}^{K}n_k+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)}}_{\text{combination of } n}\underbrace{\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}}_{\text{normalization } (Z)}\underbrace{\prod_{k=1}^{K}\theta_k^{n_k+\alpha_k-1}}_{\text{parameter}}. && \text{(A.8)} \end{aligned}$$

The log-likelihood of the Eqn. (A.8) is,

$$\begin{aligned} \log P(n|\theta)P(\theta|\alpha) =& \log\Gamma(\sum_{k=1}^{K}n_k+1) - \sum_{k=1}^{K}\log\Gamma(n_k+1) + \sum_{k=1}^{K}\log n_k\theta_k\\ &- \log Z + \sum_{k=1}^{K}(\alpha_k-1)\log\theta_k. && \text{(A.9)} \end{aligned}$$

We can estimate θ_k by using constraint of Eqn. (A.2) and also the Lagrange multiplier. The optimal parameter $\hat{\theta_k}$ is

$$\hat{\theta_k} = \frac{n_k+\alpha_k-1}{\sum_{k=1}^{K}(n_k+\alpha_k-1)}. \tag{A.10}$$

Next, we develop the probability density function with Bayesian inference. In Bayesian inference, the probability distribution of $P(n|\alpha)$ is,

$$
\begin{aligned}
P(n|\alpha) &= \int_\theta P(n|\theta)P(\theta|\alpha)d\theta & \text{(A.11)}\\
&= \int_\theta \frac{\Gamma(\sum_{k=1}^K n_k+1)}{\prod_{k=1}^K \Gamma(n_k+1)} \prod_{k=1}^K \theta_k^{n_k} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k-1} d\theta & \text{(A.12)}\\
&= \frac{\Gamma(\sum_{k=1}^K n_k+1)}{\prod_{k=1}^K \Gamma(n_k+1)} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int_\theta \prod_{k=1}^K \theta_k^{n_k+\alpha_k-1} d\theta & \text{(A.13)}\\
&= \underbrace{\frac{\Gamma(\sum_{k=1}^K n_k+1)}{\prod_{k=1}^K \Gamma(n_k+1)}}_{\text{combination of } n} \underbrace{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)}}_{\text{normalization}} \underbrace{\frac{\prod_{k=1}^K \Gamma(n_k+\alpha_k)}{\Gamma(\sum_{k=1}^K (n_k+\alpha_k))}}_{\text{parameter}}. & \text{(A.14)}
\end{aligned}
$$

By comparing Eqn. (A.8) and Eqn. (A.14), the second factor is concluded as a normalization term of the Dirichlet distribution and the third factor is the parameter of the distribution. Thus, the Dirichlet distribution $P(X|\alpha)$ that output K-dimensional N numbers $X{=}(x_1, ..., x_N)$ can be calculated as,

$$
P(X|\alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k=1}^K \Gamma(n_k+\alpha_k)}{\Gamma(\sum_{k=1}^K (n_k+\alpha_k))}. \tag{A.15}
$$

The resultant generative probability of new instance x_i given the previous observations X is,

$$
\begin{aligned}
P(x_i|X,\alpha) =& \frac{P(x_i, X|\alpha)}{P(X|\alpha)} & \text{(A.16)}\\
=& \frac{\prod_{k=1}^K \Gamma(n_k+1+\alpha_k)\, \Gamma(\sum_{k=1}^K (n_k+\alpha_k))}{\Gamma(\sum_{k=1}^K (n_k+\alpha_k)+1) \prod_{k=1}^K \Gamma(n_k+\alpha_k)} & \text{(A.17)}\\
=& \frac{n_k+\alpha_k}{\sum_{k=1}^K (n_k+\alpha_k)}. & \text{(A.18)}
\end{aligned}
$$

Looking back to Eqn.(A.10) and Eqn.(A.18), we have an additional term of count $\alpha_k - 1$ or α_k to calculate the probability. These makes a smoothing factor of Dirichlet smoothing in Eqn. (2.2).

Dirichlet Process and Chinese Restaurant Process (CRP)

The Dirichlet process extends the dimensional number K, that is specified in the Dirichlet distribution, to the infinite number. In the Dirichlet process, α_k is rewritten as,

$$
\alpha_k = \alpha_0 G_0(x_i = k), \tag{A.19}
$$

where G_0 is a base measure and α_0 is a parameter given by $\sum_k \alpha_k$. G_0 follows the Poisson distribution parameterized by λ as,

$$G_0 = \frac{(\lambda - 1)^{k-1}}{\Gamma(k)} e^{1-\lambda}. \tag{A.20}$$

Chinese restaurant process (CRP) is a representation of the Dirichlet process (Pitman, 1995). In CRP, there are an infinite number of tables. Customers visit the restaurant one by one and sit at either of existing tables that already have at least one customer, or a new table that does not have any customer. The probability of sitting at any existing table $T=(t_1, ..., t_J)$ is,

$$P(t_j|T, \alpha) = \frac{n_{t_j}}{\alpha_0 + \sum_{j=1}^{J} n_{t_j}}. \tag{A.21}$$

The probability of selecting a new table $(t_J + 1)$ is,

$$P(t_j|T, \alpha) = \frac{\alpha_0}{\alpha_0 + \sum_{j=1}^{J} n_{t_j}}. \tag{A.22}$$

Here, n_{t_j} is the number of tables. If the customer sits at a new table, new dish D_j is assigned to the table. The probability of the new table resolves the problem of unknown words that does not occur in the training set.

Estimation of Hyper-parameter γ

The generative probability of observation sequence $X=(x_1, ..., x_N)$ is defined as,

$$P(X|\alpha_0, G_0) = \prod_{i=1}^{N} P(x_i|x_{1:i-1}, \alpha_0, G_0). \tag{A.23}$$

The Dirichlet process is extended to hierarchical Dirichlet process (HDP) (Teh et al., 2006). In HDP, the generative probability of domain D given word w_i is calculated by pairs of a domain D and a word w_i. The HDP makes a smoothing by using the unigram probability of domain $P(D)$ to avoid the zero-count problem of the pair of a domain D and a word w_i. This formulation is known as a generalization of Kneser-Ney smoothing (Kneser and Ney, 1995) in the non-parametric Bayesian manner (Teh, 2006a; Teh, 2006b). The hyper-parameter γ in Eqn (2.4) is equal to α_0 that is weighted by the unigram probability of the domain. The unigram probability is calculated by the maximum likelihood estimation as the unobserved domain need not to be considered. The optimal hyper-parameter γ can be calculated by maximizing the generative probability of sequence of domain $C(D_1), ..., C(D_N)$ given a particular word w_i as,

$$P(C(D_1, w_i), ..., C(D_N, w_i)|w_i, \gamma) = \frac{\prod_{l=1}^{N} \prod_{j=0}^{C(D_l, w_i)-1} (j + P(D_l)\gamma)}{\prod_{j=0}^{\sum_{l=1}^{N}(C(D_l, w_i))-1} (j + \gamma)}. \tag{A.24}$$

Here, $(C(D_1, w_i), ..., C(D_N, w_i))$ is an N dimensional count of the domain D_l given a word w_i. According to the process, the generative probability of the Naive Bayes model Eqn. (2.4) is calculated by,

$$P(C(D_1, w_1), ..., C(D_N, w_K)|w_1, ..., w_K, \gamma) = \prod_{k=1}^{K} \frac{\prod_{l=1}^{N} \prod_{j=0}^{C(D_l, w_k)-1} (j + P(D_l)\gamma)}{\prod_{j=0}^{\sum_{l=1}^{N}(C(D_l, w_k))-1} (j + \gamma)}. \tag{A.25}$$

The log-likelihood is,

$$L_{\mathrm{NB}} = \sum_{k=1}^{K} \sum_{l=1}^{N} \sum_{j=0}^{C(D_l, w_k)-1} (j + P(D_l)\gamma) - \sum_{k=1}^{K} \sum_{j=0}^{\sum_{l=1}^{N}(C(D_l, w_k))-1} (j + \gamma). \tag{A.26}$$

The optimal hyper-parameter γ that maximizes the log-likelihood is estimated by using the Newton's method.

Authored Works

Chapter 2 and 4

Refereed

吉野幸一郎, 森信介, 河原達也. 述語項の類似度に基づく情報抽出・推薦を行う音声対話システム. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3386–3397, 2011.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In *Proceedings of Annual SIGdial Meeting on Discourse and Dialogue*, SIGDIAL 2011, pp. 59–66, Portland, Oregon, June 2011.

Koichiro Yoshino and Tatsuya Kawahara. Spoken dialogue system based on information extraction from web text. In *Proceedings of Second International Workshop on Spoken Dialogue Systems Technology*, IWSDS 2010, pp. 196–197, Gotenba, Japan, 2010.

Unrefereed

吉野幸一郎, 森信介, 河原達也. 述語項の類似度に基づいてニュース記事の案内を行う音声対話システム. 人工知能学会研究会資料, 2011-SLUD63, 東京, October 2011.

吉野幸一郎, 森信介, 河原達也. 述語項の類似度に基づく情報推薦を行う音声対話システム. 情報処理学会研究報告, 2011-SLP87, 札幌, July 2011.

吉野幸一郎, 森信介, 河原達也. 情報抽出と述語項の類似度を利用した音声対話システム. 言語処理学会年次大会論文集, 2011-D1-6, 東京, March 2011.

吉野幸一郎, 河原達也. Web からの情報抽出を用いた対話システムの評価. 人工知能学会研究会資料, 2010-SLUD60, 東京, October 2010.

吉野幸一郎, 河原達也. Web からの情報抽出を用いた音声対話システム. 情報処理学会研究報告, 2010-SLP82, 仙台, July 2010.

吉野幸一郎, 河原達也. Web からの情報抽出に基づく雑談的な対話の生成. 言語処理学会年次大会論文集, 2010-C2-2, 東京, March 2010.

Chapter 3

Refereed

吉野幸一郎, 森信介, 河原達也. 述語項構造を介した文の選択に基づく音声対話用言語モデルの構築. 人工知能学会論文誌, Vol. 29, No. 1, pp. 53–59, 2014.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. Incorporating semantic information to selection of web texts for language model of spoken dialogue system. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2013, pp. 8252–8256, Vancouver, Canada, May 2013.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. Language modeling for spoken dialogue system based on sentence transformation and filtering using predicate-argument structures. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, APSIPA 2012, Hollywood, California, USA, December 2012.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. Language modeling for spoken dialogue system based on filtering using predicate-argument structures. In *Proceedings of International Conference on Computational Linguistics*, COLING 2012, pp. 2993–3002, Mumbai, India, December 2012.

Unrefereed

吉野幸一郎, 森信介, 河原達也. 述語項構造を介した web テキストからの文選択に基づく言語モデルの評価. 情報処理学会研究報告, 2013-SLP97, 仙台, July 2013.

吉野幸一郎, 森信介, 河原達也. 述語項を介した文の変換と選択に基づく音声対話用言語モデルの構築. 情報処理学会研究報告, 2012-NL206/SLP91, 東京, May 2012. 2012 年度情報処理学会自然言語処理研究会・音声言語情報処理研究会学生奨励賞.

吉野幸一郎, 森信介, 河原達也. 述語項構造を用いた文変換とフィルタリングに基づく音声対話用言語モデル. 言語処理学会年次大会論文集, 2012-D3-2, 広島, March 2012.

Chapter 5

Refereed

Koichiro Yoshino and Tatsuya Kawahara. Conversational system for information navigation based on POMDP with user focus tracking. *Computer Speech & Language*, Submitted.

Koichiro Yoshino and Tatsuya Kawahara. Information navigation system based on POMDP that tracks user focus. In *Proceedings of Annual SIGdial Meeting on Discourse and Dialogue*, SIGDIAL 2014, pp. 32–40, Philadelphia, Pennsylvania, June 2014.

Koichiro Yoshino, Shinji Watanabe, Jonathan Le Roux, and John R. Hershey. Statistical dialogue management using intention dependency graph. In *Proceedings of International Joint Conference on Natural Language Processing*, IJCNLP 2013, pp. 962–966, Nagoya, Japan, October 2013.

Unrefereed

吉野幸一郎, 河原達也. ユーザの焦点に適応的な雑談型音声情報案内システム. 言語処理学会年次大会論文集, 2014-C5-4, 札幌, March 2014.

吉野幸一郎, 河原達也. ユーザの焦点に適応的な雑談型音声情報案内システム. 人工知能学会研究会資料, 2014-SLUD70, 小金井, March 2014. 2013 年度人工知能学会研究会優秀賞.

吉野幸一郎. ユーザの焦点に適応的な音声によるニュース案内システム. 情報処理学会研究報告, 2013-SLP100, 伊豆, January 2014.

Others

Refereed

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. Predicate-argument structure analysis using partially annotated corpora. In *Proceedings of International Joint Conference on Natural Language Processing*, IJCNLP 2013, pp. 957–961, Nagoya, Japan, October 2013.

Unrefereed

吉野幸一郎, 森信介, 河原達也. 点予測による述語項構造解析. 情報処理学会研究報告, 2012-NL209, 京都, November 2012.

Co-authored Works

Refereed

平山直樹, 吉野幸一郎, 糸山克寿, 森信介, 奥乃博. 入力方言の自動推定による複数方言音声認識システムの構築. 情報処理学会論文誌, Vol. 55, No. 7, pp. 1681–1694, 2014.

Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyama, Shinsuke Mori, and Hiroshi G. Okuno. Automatic estimation of dialect mixing ratio for dialect speech recognition. In *Proceedings of Annual Conference on the International Speech Communication Association*, INTERSPEECH, Lyon, France, August 2013.

Shinsuke Mori, Tetsuro Sasada, Yoko Yamakata, and Koichiro Yoshino. A machine learning approach to recipe text processing. In *Proceedings of Cooking with Computers workshop*, CWC 2012, Lyon, France, August 2012.

Shinsuke Mori, Hirokuni Maeta, Tetsuro Sasada, Koichiro Yoshino, Atsushi Hashimoto, Takuya Funatomi, and Yoko Yamakata. Flowgraph2text: Automatic sentence skeleton compilation for procedural text generation. In *Proceedings of International Natural Language Generation Conference*, INLG 2014, Philadelphia, Pennsylvania, June 2014.

Unefereed

平山直樹, 吉野幸一郎, 糸山克寿, 森信介, 奥乃博. 混合方言言語モデルと混合比推定による方言音声認識システム. 情報処理学会全国大会論文集, 2014-4S-6, 仙台, March 2014. 2013 年度情報処理学会全国大会学生奨励賞.

上里美樹, 吉野幸一郎, 高梨克也, 河原達也. 傾聴対話における相槌の韻律的特徴の同調傾向の分析. 人工知能学会研究会資料, 2014-SLUD70, 小金井, March 2014.