NAIST-IS-DD0761020

Doctoral Dissertation

An Analysis of Non-Task-Oriented Dialogs and a Computational Model of Generating Affective Utterances

Ryoko TOKUHISA

August 20, 2009

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

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

Ryoko TOKUHISA

Thesis Committee: Professor Yuji Matsumoto (Supervisor)
Professor Kiyohiro Shikano (Co-supervisor)
Associate Professor Kentaro Inui (Co-supervisor)

An Analysis of Non-Task-Oriented Dialogs and a Computational Model of Generating Affective Utterances *

Ryoko TOKUHISA

Abstract

Previous research into human-computer interaction has mostly focused on task-oriented dialogs, where the goal is to achieve a given task as precisely and efficiently as possible by exchanging information required for the task through dialog. On the other hand, the necessity of non-task-oriented dialog systems has been raising in recent years. Specially, non-task-oriented dialog systems are needed for home robots. We aim at realization of the verbal communication with robots. This thesis takes up the problem of the non-task-oriented dialog by two steps: the first step is conversation analysis and the second step is its engineering implementation.

In the conversation analysis step, the following two issues are investigated: a) What makes a non-task oriented human-to-human conversation to be an enthusiastic one; b) What is the difference between task-oriented and non-task-oriented dialogs. The first issue is explained by investigating what type of utterances contributes to enthusiasm in a non-task-oriented human-to-human dialog. For this end, we first create a non-task-oriented human-to-human dialog corpus. We then analyze the relationship between utterances and enthusiasm by studying the instances by studying the instances in the corpus. As a result, we found that "affective utterance" and "cooperative utterance" were related to enthusiasm. On the other hand, concerning the second issue, we investigate what type of utterances appear saliently in non-task-oriented dialog. We first create two types of human-to-human dialog corpora: a task-oriented dialog corpus and a non-task-oriented dialog corpus. We investigate what are the discriminating characteristics that differentiate them. We found that initiation/response utterance appeared

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0761020, August 20, 2009.

more frequently in non-task-oriented dialog. This is because participants often try to take a lead to continue the exchange smoothly in a non-task-oriented dialog. In addition, we also found that most indirect response and clarification requests work as initiation/response utterance in a non-task-oriented dialog.

Next, in the implementation step, we propose a method to generate affective utterances that express sympathetic emotion to the partner. We first automatically collect a huge collection of emotion-provoking event instances from the World Wide Web. We classify the emotion-provoking events in terms of the emotion types and their polarity. So, the task in decomposed into the following two sub tasks: sentiment polarity (positive and negative) classification and emotion (e.g. happiness, sadness, fear) classification. The results of the experiments showed that our method significantly outperformed the baseline method.

Keywords:

Non-task-oriented dialog, Conversation analysis, Enthusiasm, Emotion classification, Emotion-provoking event corpus

雑談対話の分析と感情応答生成*

徳久 良子

内容梗概

従来の対話研究の多くは，対話を通して何らかのタスクを遂行する課題遂行対話を対象としてきた．一方，ホームロボットのようなアプリケーションでは，課題遂行対話だけでなく話すこと自体を目的とした非課題遂行対話も求められる．このような背景から，本論では，人と話すこと自体を目的としたロボットの実現に向けて雑談対話の課題に取り組む．

まず，会話分析的な立場から「いかにして盛り上がる雑談を実現するか（雑談の盛り上がりに寄与する発話は何か）」および「そもそも雑談とは何か（雑談に特徴的な発話は何か）」というふたつの課題を明らかにする．ひとつめの課題に対しては，人間同士の雑談対話を収集し，対話の盛り上がりと発話との関係を調べた．その結果,「感情を表現する発話」や「協調的な発話」などが盛り上がりと関係することが分かった．また，ふたつめの課題に対しては，人間同士の課題遂行対話と雑談対話を収集し，両者における発話のやりとりを比較することで雑談対話に特徴的に出現する発話を解明した．その結果，雑談対話では「働きかけ/ 応答」の発話が多く出現することが分かった．雑談対話では，話者が自然に発話のやりとりを継続しようとすることが原因と考えられる．また,「間接応答」および「問い返し」のほとんどが「働きかけ/応答」発話になることも明らかとなった．

次に，工学的な立場から，会話分析から雑談に重要と分かった「感情を表現する発話」を対象として，ユーザ発話の意味する感情を推定する手法を提案する．具体的には，感情極性（ポジティブ/ネガティブ）を推定した後に，感情（嬉しい，悲しいなど 10 種類）を推定する手法を提案した．また，感情極性や感情を推定する際には，World Wide Web から自動獲得した「感情生起要因コーパス（感情生起の要因となる事例集）」を用いた．評価実験の結果，我々の提案手法により，従来手法より有意に高い精度で感情推定できることが確認された．

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文, NAIST-IS-DD0761020, 2009 年 8 月 20 日.

キーワード

雑談対話、会話分析、盛り上がり、感情推定、感情生起要因コーパス

Acknowledgements

指導教官の松本裕治教授には，私の研究テーマに対して適宜的確なコメントをいただきました．仕事で大学に足を運ぶことが少ない私に，研究会やメールなどさまざまな形で真摯に御指導くださいました．ありがとうございました．鹿野清宏教授には学内で発表を聞いていただく度に音声対話システム実現の観点から意義のある指摘を多数いただきました．ありがとうございました．乾健太郎准教授には，私の研究テーマに多くの助言をいただいただけでなく，自然言語処理研究のおもしろさを教えていただきました．九州工業大学を卒業してから研究職についたのも，卒業後に再び博士後期過程に入学したのも，乾准教授が日頃から研究に対する熱心な態度を見せてくださったおかげです．本当にありがとうございます．

本論文や諸発表に関して細部に渡り助言をくださった浅原正幸さん，研究会でいつも意義のあるコメントをしてくださった飯田龍さん，勉強会でさまざまな議論してくださった水野淳太さん，清水友裕さん，対話勉強会のメンバーに心から感謝致します．その他の研究室のみなさまにも遠隔で学生生活を送る私をさまざまな形でサポートしていただきました．ありがとうございました．

本論の実験では小林のぞみさんが整備してくださった辞書，河原大輔さんが収集してくださったコーパス，および，工藤拓さんが開発してくださったツールを利用させていただきました．有益な言語資源やツールを提供してくださった彼らに深く感謝致します．また，九州工業大学時代からさまざまな形で助言をくださった乾孝司さん，藤田篤さんに感謝致します．

寺嶌立太氏は，本研究テーマの初期から一緒に問題に取り組み，さまざまな形で助言をくださいました．彼なしには本論文は成立しませんでした．ここに深く感謝致します．日頃から研究に関して多くの議論をしてくれた下岡和也氏，星野博之氏に深く感謝致します．また，対話の分析に不可欠なタグ付与作業を手伝ってくださった作業者のみなさまに深く感謝致します．

最後に，いつも私の研究を応援してくれる父・弘，母・淳子，いつもそばで支えてくれる夫・堀田隆介，娘・明里に心から感謝したいと思います．本当にありがとう．

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Since the performance in human language technologies such as Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) have been improved, this advance has made it possible to develop practical dialog systems.

Let's Go Lab

Let's Go Lab is a spoken dialog system for bus information [17]. This in turn is based on Olympus, a dialog system architecture born out of Carnegie Mellon's existing expertise in speech recognition, speech synthesis and spoken dialog systems. With around 60 calls a day Let's Go Lab offers a unique experimental environment for evaluation of advances in spoken dialog systems. Initially, Let's Go was a typical research mixed-initiative system used for in-lab experiments [42]. However, in order to serve the diverse user population of the Port Authority, the interaction was constrained to make it easier for novice users. Much effort was spent on the efficiency and robustness that is required for a continuous public service. Let's Go Lab went "live" in March 2005 [41], providing bus schedule information daily from early evening through early morning, when the Port Authority offices are closed.

Kyoto Bus Information System

Komatani et al. have developed the Kyoto City Bus Information System [26]. The system locates a bus the user wants to catch and tells him/her how long it will be before

the bus arrives. The system can be accessed by telephone. Users are required to input their boarding stop, destination or bus route number by voice. As a result, users obtain appropriate bus information.

Commute UX

Commute UX (Commute User Experience) is a telephone dialog system that provides location-based information to in-car commuters [51]. The system provides information about traffic, gas prices, and weather, based on real-time data obtained via web services. It was built as a telephone dialog system to enable any user with an ordinary cell phone to access these services without the need for a data plan or addi-tional hardware or software installed in the car.

As we reviewed above, previous research into human-computer interaction has mostly focused on task-oriented dialogs, where the goal is to achieve a given task as precisely and efficiently as possible by exchanging information required for the task through dialog [25, 32, 58]. In these researches, various issues have been addressed: domain-adaptive large vocabulary speech recognition, efficient dialog management, corpus-based utterance generation, etc.

On the other hand, the necessity of non-task-oriented dialog systems has been raising in recent years. Specially, non-task-oriented dialog systems are needed for home robots [70, 68]. These robots are called communication robot.

Previous research on communication robots has mostly focused on non-verbal communications, where nodding, prosody, and facial expressions are most concerned topics. For example, Breazeal et al. develop a sociable robot Kismet [8]. Kismet can determine whether it is praised, prohibited, soothed, or is given an attentional bid, by analyzing the prosody of a user's speech. They reported that the robot was able to robustly classify the four affective intents in multi-lingual experiments with naive female subjects. In addition, the subjects intuitively inferred when their intent had been properly understood by Kismet's expressive feedback. Matsusaka et al. developed a robot which can participate in a group conversation[37]. Their robot can recognize the conversational situation: who is speaking, to whom he is speaking, to whom the other participants pay attention; using image and acoustic features.

In contrast, only a few studies have been reported about non-task-oriented dialog

despite the necessity of verbal communication in communication robots. In this dissertation, we therefore aim at realization of the verbal communication with robots, particularly focusing on a non-task-oriented dialog.

1.2 Objective and Goal

In this dissertation, we will take two steps for achievement of non-task-oriented dialog: the first step is a conversation analysis and the second step is its engineering implementation. In the conversation analysis step, we investigate what type of utterances appear saliently in non-task-oriented dialog. In the implementation step, we propose a method to generate the utterance. We think that we can focus on utterance generation with a high contribution for a non-task-oriented dialog system because our research is based on the result of an analysis of conversation analysis approach.

In our conversation analysis, the following two issues are addressed.

a) What makes a non-task-oriented human-to-human enthusiastic dialog

b) What are the differences between task-oriented dialogs and non-task-oriented dialog

In task-oriented dialog systems, efficient dialog strategies have been explored to achieve a given task efficiently without failure [25, 32, 58]. However, for a non-task-oriented dialog system (e.g. home robots), other factors such as enjoyability and enthusiasm need to be taken into consideration. The first issue is addressed by investigating what type of utterances contributes to enthusiasm in non-task-oriented human-to-human dialogs. For this end, we first create a non-task-oriented human-to-human dialog corpus. We then analyze the relationship between utterances and enthusiasm in the corpus.

Many task-oriented dialog systems have been developed in previous work. On the other hand, only a few studies have been presented about non-task-oriented dialog. It is expected that there are some differences between task-oriented dialog and non-task-oriented dialog, but it is not clear what difference actually exists. For the second issue, we first create two types of human-to-human dialog corpora: a task-oriented dialog corpus and non-task-oriented dialog corpus. We then investigate what type of

utterances appears remarkably in non-task-oriented dialog by making a comparison between task-oriented dialog and non-task-oriented dialog.

Next, in the implementation step, we propose a method to generate affective utterances. There are two advantages in our approach. First, we focus on an important utterance in non-task-oriented human-to-human dialog. We think that we can focus on utterance generation with a high contribution for a non-task-oriented dialog system because our research is based on the result of an analysis of conversation analysis approach. Second, we use huge examples extracted automatically from the World Wide Web. A lot of previous dialog systems used a knowledge manually described. But an enormous cost were needed for such an approach. It is expected that appropriate system responses would be generated because large-scale examples are used for utterance generation in our approach.

The rest of this thesis is organized as follows. Firstly, in Chapter 2, we review previous studies related to dialog systems. In Chapter 3, we analyze a non-task-oriented dialog by conversation analysis approach. In Chapter 4, we propose a method for emotion classification using huge examples extracted from the web. Finally, Chapter 5 summarizes the thesis and ends with conclusions and future work.

Chapter 2

Previous work on dialog systems

We can categorize dialogs into two kinds: one is a task-oriented dialog and the other is a non-task-oriented dialog. The goal of a task-oriented dialog is to achieve a given task by exchanging information required for the task through dialog. On the other hand, the goal of a non-task-oriented dialog is to enjoy conversations. In the case of conversational robots, robots try to understand user's requests accurately in a task-oriented dialog, on the other hand, robots try to allow users to enjoy conversations in a non-task-oriented dialog. This chapter reviews previous work on dialog systems.

2.1 Findings from conversation analysis

Previous dialog systems have been developed based on the findings from conversation analysis.

Conversation analysis was developed in the late 1960s and early 1970s principally by a sociologist, Harvey Sacks. It aims to capture the basis definition to describe the structure and sequential patterns of interaction. For this purpose, the literature has developed such notions as adjacency pairs and dialog acts. An adjacency pair is a pair of conversational turns by two different speakers. For example, a question, such as *what is your name?*, requires the addressee to provide an answer in the next conversational turn. Many actions in dialog are accomplished through adjacency pair sequences, for example: greeting-greeting, request-acceptance/denial, offer-acceptance/rejection. These actions are called *speech acts* in conversation analysis.

Dialog Acts (DAs) tags are commonly used as a simple representation of the action or function of an utterance in dialog. DAs are basically based on speech acts, but the

original repertoire [46, 55] has been gradually enriched with other possible functions. Allen et al. defined the DA tag set called DAMSL (Dialog Act Markup in Several Layers), which was initially designed as a shared resource with a focus primarily on task-oriented dialogs [3]. The DAMSL annotation schema has been used to annotate many dialog corpora: the TRAINS corpus, TRIPS corpus, Monroe corpus, etc. DAMSL has four main layers: Communicative status, Forward communicative functions, Backward communicative functions, and Information level. Communicative status denotes an utterance whether it was interpretable: Uninterpretable, Self-talk, etc. The Forward communicative functions consist of a taxonomy in a similar style as the actions of traditional speech act theory: statement, Info-request, etc. The Backward communicative functions indicate how the current utterance relates to the previous dialog, such as accepting a proposal, confirming understanding, or answering a question. Information Level encodes whether the utterance deals with the dialog task, the communication process, or metalevel discussion about the task. DAMSL not only is applied to various corpora but also is derivative to other DA tagging schemata. The application of DAMSL to the Switchboard corpus lead to SWBD-DAMSL [22]. The Switchboard corpus is a collection of two-party telephone conversations. SWBD-DAMSL was adapted DAMSL to non-task-oriented dialogs because the designed DAMSL was designed for task-oriented dialogs. The MRDA was defined for the dialogue act annotation of data from the Meeting Recorder project at ICSI [14]. The tagset basically uses the SWBD-DAMSL tags, but allows the combination of several tags into a label for an utterance. The tagset also extends SWBD-DAMSL with disruption marks such as “interrupted”, “abandoned”.

The notion of dialog act proposed in conversation analysis is being widely adopted by researchers in the NLP community because it is useful for developing dialog systems. Corpora annotated with dialog acts are widely used for developing task-oriented dialog systems. Previous work on task-oriented dialog systems are reviewed in the next section.

2.2 Task-oriented dialog systems

Previous research into human-computer interaction has mostly focused on task-oriented dialogs, where the goal is to achieve a given task as precisely and efficiently as possible by exchanging information required for the task through dialog [25, 32, 58]. In these researches, various issues have been addressed: domain-adaptive large vocab-

ulary speech recognition, efficient dialog management, corpus-based utterance generation, etc.

The TRAINS project is one of the early projects aiming at building a dialog system. The goal of this project is to build a computerized planning assistant that can interact conversationally with its user [1]. The TRAINS dialog system allowed the system and user to work together to try to route trains to make deliveries across the eastern United States. A key part of the TRAINS project is the construction of the Trains system, which provides the research platform for a wide range of issues in natural language understanding, mixed-initiative planning systems, and representing and reasoning about time, actions and events.

TRIPS is the next generation of the TRAINS project. The goal of the TRIPS project is to build an intelligent planning assistant using natural language and graphical display [2]. TRIPS is a domain independent, mixed-initiative dialog system core, which can be easily ported to new domains. It has successfully been ported to such domains as emergency evacuation, disaster relief, and military resource allocation.

The DARPA Communicator project is mainly focused on a speech recognition and dialog management. The common task of the The DARPA Communicator project is a mixed-initiative dialog over the telephone, in which the user plans a multi-city trip by airplane, including all flights, hotels, and rental cars, all in conversational English over the telephone [59, 60]. The DARPA Communicator project describes a hub-and-spoke architecture for the design and development of natural language understanding systems. The system combines speech recognition, natural language understanding, dialog management, database access, language generation and speech synthesis to perform the desired task, which at present can be described as an automated travel agent that helps callers make airline reservations [30, 45].

Recently, the research about task-oriented dialog systems are applied to robots.

Roy et al. develop an autonomous wheelchair that can learn all about the locations in a given building, and then take its occupant to a given place in response to a verbal command [21, 15]. Just by saying *take me to the cafeteria* or *go to my room*, the wheelchair user would be able to avoid the need for controlling every twist and turn of the route and could simply sit back and relax as the chair moves from one place to another based on a map stored in its memory. The system also can learn locations by user's verbal guidance. For example, as the wheelchair is pushed around a nursing home for the first time, the patient or a caregiver would say: *this is my room* or *here we are in the nurse's station*. The system then learn about its environment by being taken

on a guided tour.

Lopes et al. develop CARL (Communication, Action, Reasoning and Learning in Robotics) [34]. CARL is a mobile intelligent robot which can achieve some tasks: spreading a table, etc. CARL can flexibly accomplish tasks because it works based on a KKR (knowledge representation and reasoning) module. The KRR module supports the integration of information coming from different interlocutors and is capable of handling contradictory facts. CARL then works using the hypothesis that a combination of reactivity with reasoning is more likely to produce useful results in a relatively near future than the purely reactive or behavior-based approaches. This is especially true for robots that are expected to perform complex tasks requiring decision-making.

Kanda et al. develop a robot called Robovie [70]. Robovie achieves natural communications like human-to-human communication using non-verbal-information: gaze, gesture, etc. They report that such a physical expressions are useful for communication, specially in a navigation task.

2.3 Non-task-oriented dialog systems

As we mentioned above, this dissertation focuses on non-task-oriented dialog. This section describes the previous research on non-task-oriented dialog systems.

Robots reviewed in section 2.2 are designed to achieve a given task by interacting with a user. In that sense, those robots are considered task-oriented. In contrast, non-task-oriented robots are called communication robot. Previous research on communication robots has mostly focused on non-verbal communications, where nodding, prosody, and facial expressions are most concerned topics. For example, Breazeal et al. develop a sociable robot Kismet [8]. Kismet can determine whether it is praised, prohibited, soothed, or is given an attentional bid, by analyzing the prosody of a user's speech. They reported that the robot was able to robustly classify the four affective intents in multi-lingual experiments with naive female subjects. In addition, the subjects intuitively inferred when their intent had been properly understood by Kismet's expressive feedback. Matsusaka et al. developed a robot which can participate in a group conversation[37]. Their robot can recognize the conversational situation: who is speaking, to whom he is speaking, to whom the other participants pay attention; using image and acoustic features.

On the other hand, some research groups have reported an interactive robot which mentally supports aged people. Initially, interaction with animals has long been known

to be emotionally beneficial to people [6, 18, 29]. In recent years, the effects of animals on humans have been studied and proved scientifically [47, 57, 65]. Shibata et al. develop a seal robot called Paro [47, 57]. Paro has five kinds of sensors: tactile, light, audition, temperature, and posture sensors, with which it can perceive people and its environment. With the light sensor, Paro can recognize light and dark. He feels being stroked and beaten by tactile sensor, or being held by the posture sensor. Paro can also recognize the direction of voice and words such as its name, greetings, and praise with its audio sensor. By interaction with people, Paro responds as if it is alive, moving its head and legs, making sounds. Elderly subjects with cognition disorders were asked to interact with Paro. The subjects' brain waves were measured before and after the interaction for analysis. Out of the 14 people who participated in the study, 7 (50%) experienced an improvement in brain function.

A chatbot (also chatterbot, chatterbox) is a dialog system which attempts to maintain a non-task-oriented dialog with a user. The classic early chatbots are ELIZA and PARRY. More recent chatbots are Albert One, A.L.I.C.E., etc [62, 13, 61].

ELIZA is a computer program by Joseph Weizenbaum, designed in 1966 [62]. It parodied a Rogerian therapist, largely by rephrasing many of the patient's statements as questions and posing them to the patient. For example, the response to "My mother hates me" might be "Who else in your family hates you?". The response to "My head hurts" might be "Why do you say your head hurts?". Weizenbaum reported that many patients felt that Eliza could understand their utterance.

PARRY was written in 1972 by psychiatrist Kenneth Colby [13]. While ELIZA was a tongue-in-cheek simulation of a Rogerian therapist, PARRY attempted to simulate a paranoid schizophrenic. It appeared to be a paranoid patient in a Veterans' hospital:

User: Have you been hospitalized before?

PARRY: No, this is the first time.

PARRY knew little of the world, had no syntax analysis and just worked by a large set (about 6000) of patterns with which it matched any input.

Albert One won the Loebner Prize Contest, a competition between chatbots, in both 1998 and 1999 [16]. The Albert One system was comprised of several subsystems. Among those were a version of Eliza, Elivs, another Eliza-like systems, and several other helper applications working together in a hierarchical arrangement. As a continuation of the stimulus-response library, various other database queries and assertions were tested to arrive at each of Albert's responses.

A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) has won the Loebner Prize three times (in 2000, 2001 and 2004) [61]. The following is part of a transcript recording the conversation between a user and A.L.I.C.E..

User: Hello, Alice.

ALICE: Hi there!

User: Tell me about yourself.

ALICE: I am an artificial linguistic entity. I was activated at Bethlehem, Pennsylvania, on November 23, 1995. My instructor was Dr. Richard S. Wallace. He taught me to sing a song. Would you like me to sing it for you?.

User: No, thank you.

ALICE: Why not?

User: I'm in a hurry.

ALICE: Do you have another appointment now?

User: Yes. Bye.

ALICE: See you later.

A.L.I.C.E has handwritten a database of thousands of possible conversational gambits. Type a utterance to ALICE, and it checks the phrase and its keywords for a response coded to those words.

As we described above, most chatbots simply scan for keywords within the input and pull a reply with the most matching keywords or the most similar wording pattern from a local database. But it is not clear basic issues: "how to accomplish a more "humanlike enthusiasm" for a non-task-oriented dialog" and "what is a non-task-oriented dialog".

Chapter 3

Conversation analysis as a basis of developing non-task-oriented dialog systems

3.1 Introduction

In this chapter, we investigate what type of utterances characterize in non-task-oriented dialogs. To this end, we analyze non-task-oriented dialogs from a point of view of conversation analysis. In the literature of task-oriented dialog systems, efficient dialog strategies have been explored to achieve a given task efficiently without failure [25, 32, 58]. However, to develop non-task-oriented dialog systems (e.g. home robots), other factors such as enjoyability and enthusiasm need to be taken into consideration.

In section 3.2, we investigate what makes a dialog enthusiastic by analyzing non-task-oriented human-to-human dialogs. More specifically, we investigate what type of utterances contributes to enthusiasm by creating non-task-oriented human-to-human dialog corpus and analyzing the correlation between utterance types and enthusiasm in the corpus. For utterance types, we use Dialog Acts and Rhetorical Relations, which are both well-established schemata for annotating utterances or sentences in dialogs.

While, section 3.2 focuses on enthusiasm in dialogs, in section 3.3, we more our focus on to the continuity of each dialogs exchange. Task-oriented dialogs tend to be simple and short because the goal is to achieve a given task as efficiently as possible. On the other hand, in non-task-oriented dialogs, the participants often make a particular kind of efforts to continue the current exchange because their aim is in conversing with

each other. We therefore segment each dialog into exchanges and investigate what type of utterances contribute to continuing an exchange.

3.2 What sort of utterances create "Enthusiasm" in non-task-oriented dialogs

In task-oriented dialog systems, efficient dialog strategies have been explored to achieve a given task efficiently without failure [25, 32, 58]. However, for a non-task-oriented dialog system (e.g. home robots), other factors such as enjoyability and enthusiasm need to be taken into consideration. Much research has been done into efficient dialog strategies, but it has not been clarified what makes a non-task-oriented human-to-human enthusiastic dialog. In this section, we analyze a non-task-oriented human-to-human dialog and investigate what type of utterances contributes to enthusiasm.

3.2.1 Corpus collection

Several non-task-oriented dialog corpora have been created under various settings: face-to-face or non-face-to-face, familiar or unfamiliar, and two-party or multi-party [19, 53]. The largest non-task-oriented dialog corpus is the Switchboard Corpus, which consists of about 2400 conversational English dialogs between two unfamiliar speakers over the telephone on one of 70 topics (e.g. pets, family life, education, gun control, etc.).

Our corpus was collected from face-to-face interaction between two unfamiliar speakers. The reasons for our choosing this are 1) face-to-face interaction increases the number of enthusiastic utterances, relatively to limited conversational channel interaction such as over the telephone; 2) the interaction between unfamiliar speakers reduces the enthusiasm resulting from unobserved reasons during the recording; 3) the exchange in a two-party dialog is expected to be simpler than that of a multi-party dialog.

We created a corpus containing ten non-task-oriented dialogs that were spoken by an operator (thirties, female) and one of ten subjects (twenties to sixties, equal numbers of males and females). Before beginning the recording session, the subject chose three cards from fifteen cards on the following topics:

Table 3.1: Dialog Act Definition

SWBD-DAMSL/MRDA	Our DAs	Definition
Statement non opinion	inform objective fact	inform non opinion
Statement opinion	inform subjective element	inform opinion
Wh-Question	request objective fact	request non opinion
Yes-No-question	request agreement	request agreement opinion
Open-Question	request disagreement	request disagreement opinion
Or-Question	confirm objective fact	confirm non opinion
	confirm agreement	confirm agreement opinion
	confirm disagreement	confirm disagreement opinion
Accept	accept	accept non opinion
	agree	accept opinion
Reject	denial	denial non opinion
	disagree	denial opinion
not marked	express admiration	inform admiration
Summary	DEL. (mark as RR)	————

Food, Travel, Sport, Hobbies, Movies, Prizes, TV Programs, Family,
Books, School, Music, Pets, Shopping, Recent Purchases, Celebrities

Straying from the selected topic was permitted because these topic cards were only ever intended as a prompt to start the dialog. Thus, we collected ten dialogs, each about 20 minutes long. For convenience, in this section, we refer to the operator as **speaker1**, and the subject as **speaker2**.

3.2.2 Annotation of dialog acts and rhetorical relations

3.2.2.1 Definition of tagging schema

To investigate what type of utterances contributes to enthusiasm, we annotate an utterance in non-task-oriented human-to-human dialogs. Dialog Acts (DAs) and Rhetorical Relations (RRs) are well-known tagging schemata for annotating an utterance or sentence. DAs are tags that pertain to the function of an utterance itself, while RRs indicate the relationship between sentences or utterances. We adopted both tags to allow us to analyze the aspects of utterances in various ways, but adapted them slightly for our particular needs.

Dialog Acts (DAs)

The DA annotations were based on SWBD-DAMSL [22] and MRDA [14]. The SWBD-DAMSL is the DA tagset for labeling a non-task-oriented dialog. The Switchboard Corpus mentioned above was annotated with SWBD-DAMSL. On the other hand, the MRDA is the DA tagset for labeling the dialog of a meeting between multiple participants. Table 3.1 shows the correspondence between SWBD-DAMSL/MRDA and our DAs[1]. We describe some of the major adaptations below.

The tags pertaining to questions:

In SWBD-DAMSL and MRDA, the tags pertaining to questions were classified by the type of their form (e.g. *Wh-question*). We re-categorized them into request and confirm in terms of the "act".

The tags pertaining to responses:

We subdivided *Accept* and *Reject* into objective responses (*accept*, *denial*) and subjective responses (*agree*, *disagree*).

The emotional tags:

From previous experience, we believed that it was significant for enthusiastic dialog to convey admiration or interest. We therefore added tags that indicate the expression of *admiration* and *interest*.

The overlap tags with the RRs definition:

We deleted any tags (e.g. *Summary*), that overlapped the RR definition.

Consequently, we defined 47 DAs for analyzing a non-task-oriented dialog.

Rhetorical Relations (RRs)

The RR annotations were based on the rhetorical relation defined in Rhetorical Structure Theory (RST) to impose a discourse structure on a multi-sentential text [35, 49]. RST acknowledges that there are two types of relations between discourse elements, and distinguishes between subject matter and presentational relation. The subject matter relation are informational; the presentational relation are intentional. Our RR definition was based only on informational level relation because we annotated the intentional level with DAs. Table 3.2 shows the correspondence between the informational relation of RST and our RRs. We describe some of the major adaptations below.

[1]The tags listed in *italics* are based on SWBD-DAMSL while those in **boldface** are based on MRDA.

Table 3.2: Rhetorical Relation Definition

RST	Our RRs	definition
Evaluation	evaluation (positive)	U2 is a positive evaluation about U1
	evaluation (negative)	U2 is a negative evaluation about U1
	evaluation (neutral)	U2 is neutral evaluation about U1
Volitional cause	volitional cause-effect	U2 is a volitional action, and U1 cause U2
Volitional result		
No Definition	addition	U2 consists of a part of U1

Subdivide evaluation:

The evaluation reflects the degree of enthusiasm in the dialog, so we divided the *Evaluation* into three types of *evaluation (positive/ negative/ neutral)*.

Integrate the causal relations:

We use a directed graph representation for RR annotations, so that we integrate *Non-volitional cause* and *Non-volitional result* into *non-volitional cause-effect*, and *Volitional cause* and *Volitional result* into *volitional cause-effect*.

Add addition relation:

The RRs initially represent the structure of the written text, segmented into clause-like units. Therefore, they do not cover those cases in which one clause is uttered by one speaker, but communicatively completed by another. So, we added an *addition* to our RRs. The following is an example of *addition*.

speaker A: the lunch in our company cafeteria

speaker B: is good value for money

Integrate Contrast and Otherwise:

According to the analysis of a corpus annotated in a trial, we found that it is difficult to distinguish between *Contrast* and *Otherwise*. So we integrated them into *antithesis*.

We defined 16 RRs as a result of these adaptations.

3.2.2.2 Dialog acts and rhetorical relations annotation

DAs and RRs are annotated using the MMAX2 Annotation Tool [2] [38]. This supports multilevel annotation and the creation of a relationship between words. Figure 3.1 shows our corpus annotated with DAs and RRs. The ⟨ ⟩ symbol in Figure 3.1 indicates a DA, while the [ ] symbol indicates an RR. Below, we describe our annotation process for DAs and RRs.

Step 1. Utterance Segmentation

All the utterances in the dialog are segmented into DA segments, each of which we define as an *utterance*. For example, ”yes, he’s really handy to have around” on line 9 is segmented into ”yes” and ”he’s really handy to have around”. In Figure 3.1, the utterance is surrounded with a square. In this step, we also eliminated backchannels from the exchange. We chose to do this because it is difficult to identify the function of backchannels that are usually made in the background by a speaker who does not have an initiative.

Step 2. Annotation of DAs

DAs are annotated to all utterances. In those cases in which one DA alone cannot represent an utterance, two or more DAs are used. For example, the utterance ”so many?” on line 4 indicates *understanding* of and *exclamation* at the previous utterance ”about 2 or 3 movies per week”, as well as indicating confirmation of the *objective fact* that leads to the following utterance, ”we sometimes watch many more”.

Step 3. Annotation of Adjacency Pairs

Adjacency pairs (APs) are labeled. An AP consists of two utterances where each part is produced by a different speaker. In Figure 3.1, the solid and dotted lines correspond to links between the APs.

Step 4. Annotation of RRs

RRs on APs are labeled, and have the relation listed in Table 3.2. A solid line indicates an AP that is labeled with RRs, while a dotted line indicates an AP that is not labeled with RRs. In those cases in which a single RR cannot represent the type of the relationship, RRs are used.

[2] http://www.eml-research.de/english/research/nlp/down- load/mmax.php

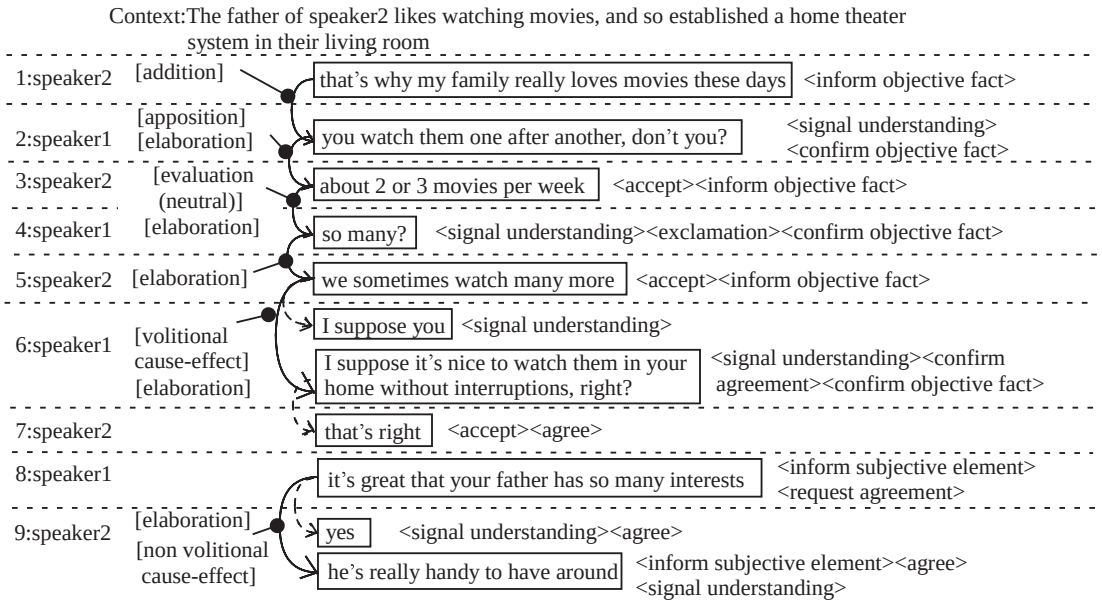


Figure 3.1: Example of Dialog annotated with DAs and RRs (Originally in Japanese)

3.2.3 Annotation of Enthusiasm

3.2.3.1 Related work on Enthusiasm

Wrede et al. annotated *Involvement* to the ICSI Meeting Recorder Corpus, which was motivated by a desire to summarize a meeting [63, 64]. Their annotation schema consisted of two steps. In the first step, utterances are labeled by a rater with respect to the perceived involvement while listening to the whole meeting. Examples of deep involvement will be detected throughout this step. In the second step, a rater judges *involvement (agreement, disagreement, other)* or *Not especially involved* or *Don't Know*, by listening to each utterance without the context of the dialog. A rater can listen to previous examples of deep involvement while performing this rating.

In the experiment, nine raters provided ratings on 45 utterances. Inter-rater agreement between *Involved* and *Not especially involved* yielded a Kappa of κ=.59 ($p<.01$), but 13 of the 45 utterances (28.9%) were rated as *Don't Know* by at least one of the raters. For automatic detection, it is certainly effective to rate involvement without context. However, the results indicate that it is quite difficult to recognize involvement from a single utterance. Moreover, the fluctuation of involvement can not be recognized by this method because Involvement is categorized into five categories only.

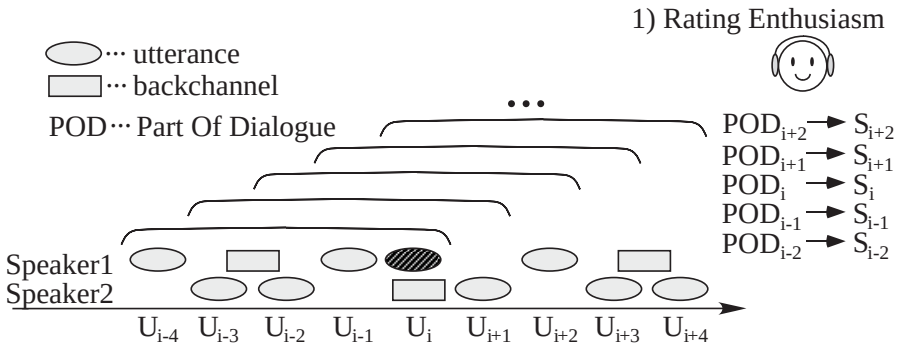


Figure 3.2: Rating the Enthusiasm

3.2.3.2 Our method of annotating Enthusiasm

In this section, we propose a method for evaluating the degree of enthusiasm. Using our method, the degree of enthusiasm can be evaluated continuously with little influence on the context while also recognizing fluctuation in the enthusiasm. We describe the process for evaluating the degree of enthusiasm.

Step 1. Rating Enthusiasm

We estimate a score for the enthusiasm corresponding to the part of dialog (POD), which is a series of utterances. To decide the number of utterances of a POD, we conducted an exploratory experiment on a set of sample PODs. Four subjects listened to 10 samples. Each sample contained 6 patterns of POD (1 utterance, 2 utterances, 3 utterances, 5 utterances, 7 utterances, and 9 utterances). After listening each POD, the subjects completed a questionnarire which asked easiness and accuracy to evaluate the degree of enthusiasm. An analysis of the results revealed that it was difficult to rate enthusiasm when the POD is shorter than 3 utterances, because subjects were sometimes not be able to understand the semantic content of a given POD a part from the context. On the other hand, it was also difficult to annotate Enthusiasm when the POD is longer than 7 utterances because the degree of enthusiasm changed within a single POD. Thus, we decided a series of 5 utterances constitute a POD in enthusiasm rating.

In Figure 3.2, U_i denotes an utterance, while S_i denotes the score for the enthusiasm of POD_i. A score ranges from 10 to 90. For example, a score of 68 is less than ”Moderate Enthusiasm”.

90 ... Extremely Enthusiastic

70 ... Moderate Enthusiasm

50 ... Neutral

30 ... Low Enthusiasm

10 ... No Enthusiasm

When rating the score, a rater must obey the following four rules.

1. Listen to each POD more than three times.
2. Perform estimation based on the entire POD and not just part of the POD.
3. Listen to PODs given the same score during rating, and then modify the rating if there is any difference from the rater's standard.
4. Estimate as participants, not as side-participants.

Furthermore, the score is estimated based on the viewpoint of each speaker independently. Therefore, two scores are labeled for each POD.

Step 2. Calculate the score of enthusiasm for each utterance

The score of enthusiasm for an utterance U_i is given by the average of the scores of the PODs that contain utterance U_i.

$$V(U_i) = \frac{1}{5} \sum_{k=i-2}^{i+2} S_k \tag{3.1}$$

Step 3. Calculate the degree of enthusiasm for an utterance

Different raters may have different absolute criteria of enthusiasm. It is effective to make a standard POD per score, but we normalize the score for enthusiasm that is given by one rater. We deal with all the degrees of enthusiasm as a normalized score, which we call **Enthusiasm**. Then, Enthusiasm for U_i is given as follows:

$$I(U_i) = \frac{V(U_i) - \overline{V(U)}}{\sigma} \tag{3.2}$$

where

$$\overline{V(U)} = \frac{1}{n}\sum_{i=1}^{n} V(U_i)$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\{V(U_i) - \overline{V(U)}\}^2}$$

n denotes the number of utterances in the dialog.

Step 4. Calculate Enthusiasm for APs

Enthusiasm for AP_i is given by the average of Enthusiasms of the utterances where are AP_i. In formula 2.3, U_i and U_j denote the ith AP and the jth AP, respectively.

$$I(U_i, U_j) = \frac{1}{2}\{I(U_i) + I(U_j)\} \tag{3.3}$$

3.2.4 Results of the annotation

3.2.4.1 Reliability of DAs and RRs

We examined the inter-annotator reliability for two annotators for DAs, RRs and APs, using four dialogs spoken by people in their twenties and sixties. We refer to these annotators as A1 and A2. A1 is one of the authors of this thesis. Before the start of the investigation, one annotator segmented a dialog into utterances. The number of segmented utterances was 697. A1 and A2 annotated them as described in steps 2 to 4 of Section 3.2.2.

DAs annotation

The Kappa statistic is usually used to measure the inter-annotator agreement. However, we can not apply the measurement in this case since the Kappa statistics cannot be applied to multiple tag annotations. We then apply formula 2.4 to examine the reliability.

$$\text{agreement} = \frac{(\text{Agreed DAs}) \times 2}{\text{Total of DAs annotated by A1 and A2}} \times 100 \tag{3.4}$$

The result of agreement was 1542 DAs (65.5%) from a total of 2355 DAs. The major reasons for the disagreement were as follows.

Table 3.3: Agreement of RRs

agree		disagree
2 annotators	0 annotators	1 annotator
233	236	67
469(87.5%)		67(12.5%)

- Disagreement of subjective/objective ... 124(15.3%)
- Disagreement of request/confirm ... 112(13.8%)
- Disagreement of partial/whole ... 72(8.9%)

Building APs

We examined the agreement of building APs between utterances. The result of agreement was 536 APs (85.2%) from the total of the 629 APs that were built by annotators. The chance agreement is very low because every utterance can construct APs. Thus, we can conclude that the building of APs is reliable.

RRs annotation

Table 3.3 shows the number of APs annotated with or without RRs. In Table 3.3, ”2 annotators” indicates the case in which both annotators annotated RRs on APs, ”0 annotators” indicates the case in which neither annotator annotated RRs on APs, and ”1 annotator” is the case where one of the annotators annotated RRs on APs. Table 3.3 shows that the agreement for the annotating of RRs is high.

We also examined the agreement of RRs annotation, using 233 APs for which both annotators annotated RRs. We applied formula 2.5 to this examination.

$$\text{agreement} = \frac{(\text{Agreed RRs}) \times 2}{\text{Total of RRs annotated by A1 and A2}} \times 100 \qquad (3.5)$$

As a result, we found agreement for 576 RRs (59.6%) out of a total of 967 RRs. We could not find any consistent explanation for the disagreement.

3.2.4.2 Evaluation of Enthusiasm

Influence of Context

Table 3.4: Correlation between random rating and sequential rating

	correlation coefficient	
	speaker1	speaker2
twenties, female	0.833	0.881
twenties, male	0.971	0.950
sixties, female	0.972	0.973
sixties, male	0.971	0.958

We examined the influence of context on Enthusiasm, using four dialogs by persons in their twenties and sixties. One rater noted Enthusiasm under two conditions.

1) Listening to PODs randomly

2) Listening to PODs sequentially as dialog

Table 3.4 shows the correlation between the random and sequential ratings. A correlation coefficient was calculated for the Enthusiasm of each of the two participants. The "speaker1" shows the correlation of the Enthusiasm rated as speaker1, and "speaker2" shows the correlation of the Enthusiasm rated as speaker2. This was found to be approximately 0.9 in both cases. These results show that Enthusiasm can be estimated stably and that the context has little influence.

Reliability of Enthusiasm

We examined the inter-rater reliability of Enthusiasm as determined by two independent annotators and described in section 3.2.4, using 10 dialogs as described in section 3.2.1. We term these raters R3 and R4.

Figures 3.3 shows the transition of the Enthusiasm of a dialog spoken by a female subject in her thirties, respectively. Table 3.5 shows the correlation coefficient and root mean square (RMS) of the inter-rater.

These indicate that our method of rating enables us to observe the fluctuation of Enthusiasm. Also, the tendency of Enthusiasm can be rated reliably.

3.2.5 Relationship between DAs/RRs and Enthusiasm

We investigated the relationship between DAs/RRs and Enthusiasm, using four dialogs by people in their twenties and sixties. The DAs/RRs corpus was annotated by

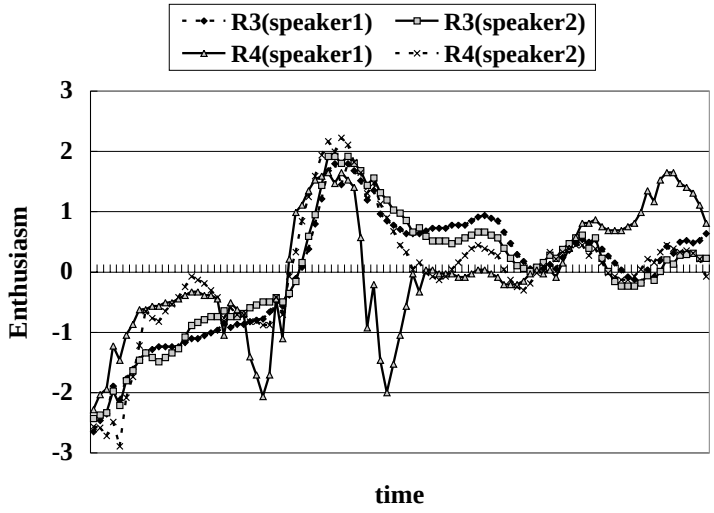


Figure 3.3: Enthusiasm of dialog of speaker1 and speaker2 (thirties, female)

A1 because A1 is one of the authors of this thesis and has a better knowledge of the DAs and RRs tagging schema than A2. The Enthusiasm corpus was annotated by R3 because we found that R4 rated Enthusiasm based on non-subjective reasons: after the examination of the rating, R4 said that speaker1 spoke enthusiastically but that it seemed unnatural because speaker1 had to manage the recording of the dialog, which appears in the results as speaker1's Enthusiasm as annotated by R4 as a notable difference (see Figure 3.3).

Figures 3.4 and 3.5 show the ratio of the frequency of DAs and RRs in each of the levels of Enthusiasm over a range of 0.5. For example, the far left bar in the *signal understanding* indicates about 0.19, which means that 19% of the DAs were *signal understanding* in the score of Enthusiasm from -2.5 to -2.0. If DAs and RRs were evenly annotated for any level of Enthusiasm, the graph will be completely even. However, the graph shows the right side as being higher if the DAs and RRs increase as Enthusiasm increases. Conversely, the graph shows the left side as being higher if the DAs and RRs fall as Enthusiasm increases. The number in Figures 3.4 and 3.5 shows the average Enthusiasm for each DA and RR. If the average is positive, it means that the frequency of the DAs and RRs is high in that part in which Enthusiasm is positive. In contrast, if the average is negative, it means that the frequency of the DAs and RRs is high in that part in which Enthusiasm is negative.

We determined the following two points about the tendency of the DAs frequency.

Table 3.5: Inter-rater agreement of Enthusiasm

	correlation coefficient		RMS	
	speaker1	speaker2	speaker1	speaker2
twenties, female	0.46	0.52	1.04	0.98
twenties, male	0.69	0.66	0.78	0.83
thirties, female	0.60	0.92	0.90	0.39
thirties, male	0.82	0.69	0.60	0.79
forties, female	0.35	0.66	1.14	0.83
forties, male	0.79	0.90	0.64	0.45
fifties, female	0.61	0.83	0.89	0.58
fifties, male	0.64	0.67	0.85	0.81
sixties, female	0.81	0.81	0.62	0.62
sixties, male	0.39	0.21	1.10	1.26

1) Tendency of subjective and objective DAs

The ratio of the frequency of those DAs related to *subjective elements* tends to increase as Enthusiasm increases (see *1 in Figure 3.4). In contrast, the ratio of the frequency of those DAs pertaining to *objective matters* tends to decrease as Enthusiasm increases (see *2 in Figure 3.4). We can thus conclude that those exchanges related to subjective elements increases in the enthusiastic dialog, but those related to objective elements decrease.

2) Tendency of affective DAs

The ratio of the frequency of *show humor* and *show interest*, which are related to the affective contents, tends to increase as Enthusiasm increases (see *3 in Figure 3.4). However, *express admiration*, which is also related to affective contents, tends to decrease (see *4 in Figure 3.4). We then analyzed several instances of *admiration*. As a result, we found that the prosodic characteristic of admiration utterance will cause this tendency.

Furthermore, we noted the following two points about the tendency of the RRs frequency.

1) Tendency of additional utterances

The ratio of the frequency of *addition*, which completes the other participant's utterance, tends to increase as Enthusiasm increases (see *5 in Figure 3.5). Figure 3.6 shows a dialog example for *addition*. There are *addition* relations be-

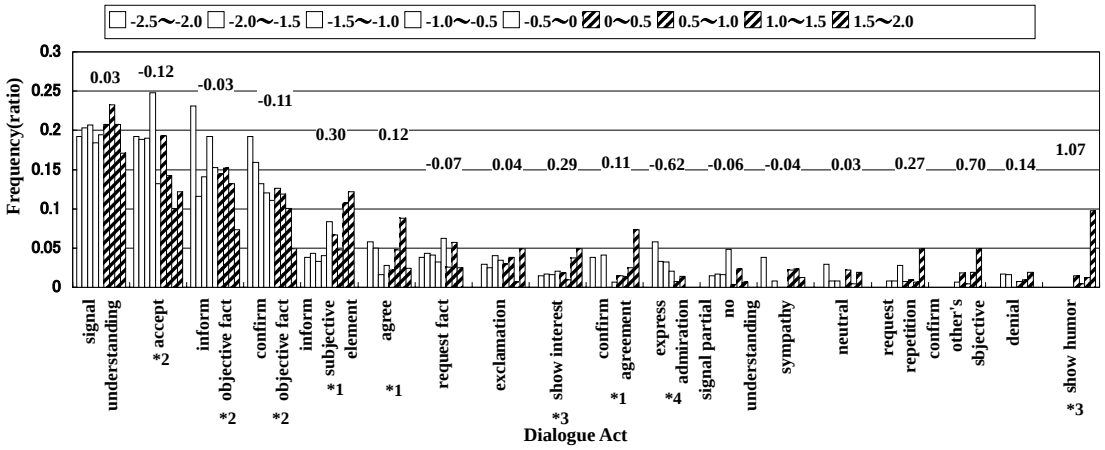


Figure 3.4: Frequency of DAs per Enthusiasm

tween lines 1 and 2. This shows that the participant makes an utterance cooperatively by completing the other's utterances in enthusiastic dialogs. Such cooperative utterance is a significant component of an enthusiastic dialog.

2) Tendency of positive evaluation

The ratio of the frequency of *positive evaluation* tends to increase at lower Enthusiasm and higher Enthusiasm (see *6 in Figure 3.5). The speaker tries to create Enthusiasm by an utterance of *positive evaluation* in a dialog with low Enthusiasm, and the speaker summarizes with a *positive evaluation* in a dialog with high Enthusiasm. Figure 3.7 shows an example of *positive evaluation* in enthusiastic dialog. In this case, speaker1 expresses *positive evaluation* on line 10 about the element on line 9. The utterance on line 10 also has the function of expressing an overall *positive evaluation* of the previous dialog. Consequently, the utterance brought the dialog to a conclusion and moved to the next topic on line 11.

3.3 The analysis of distinctive utterances in non-task-oriented dialogs

Many task-oriented dialog systems have been developed in previous work. On the other hand, very few studies have been presented for non-task-oriented dialogs. There

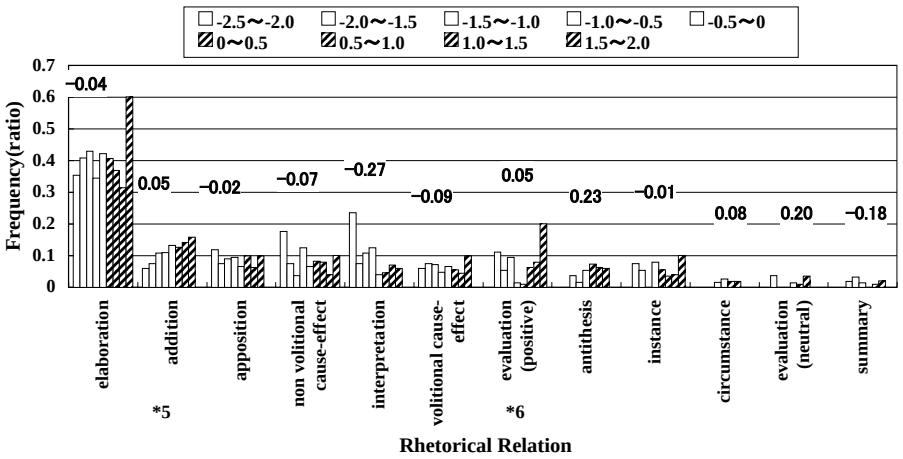


Figure 3.5: Frequency of RRs per Enthusiasm

Dialogue between speaker1 and speaker2 (twenties,female)
Context: Mother of speaker2 does not cook dinner when the father is out.

1 speaker1:but if he's there
2 speaker2:then she cooks a really delicious dinner
3 speaker1:wow

Figure 3.6: Example of *addition*

might be some differences between task-oriented and non-task-oriented dialogs, but it is not clear what they are. In this section, we investigate the difference between task-oriented dialog and non-task-oriented dialog.

3.3.1 Corpus collection

We first created two types of human-to-human dialog corpora: a task-oriented dialog corpus and non-task-oriented dialog corpus. This section explains the design of our corpora.

a) Task-oriented dialog corpus

Various task-oriented corpora have been created in previous work [1, 2, 59]. There are also various tasks in these corpora: map task, information seeking,

Dialogue between speaker1 and speaker2 (twenties,female)
Context: About a hamster and its exercise instrument.

1 speaker2:two hamsters run together in their exercise wheel.
2 speaker2:they run up and down and side by side
3 speaker1:but surely they can't they run together if they aren't getting along very well?
4 speaker2:exactly
5 speaker2:one gets carried along if it stops when the other continues to run.
6 speaker1:is it?
7 speaker1:does it lean forward?
8 speaker2:yes
9 speaker2:sometimes it falls out
10 speaker1:that's so cute
11 speaker1:when I go to a pet shop.....

Figure 3.7: Example of *positive evaluation*

etc. It is not easy, however, to analysis task-oriented dialog for every domain. Therefore, we restrict our scope to information seeking dialogs which are still applicable to a broad range of applications such as QA systems, flight reservation systems, and so on.

We created a corpus containing ten information seeking dialogs that were spoken by an operator (thirties, female) and one of ten subjects (twenties to sixties, equal numbers of males and females). We prepared several information seeking tasks, one of which, for example, was "Find a French restaurant to go with your friends near Tokyo station". We then, for each session, asked the subject to carry out one of those tasks by conversing with the operator. In each session, the subject talked with the operator about his or her information needs. The subject then retrieved restaurant information satisfying their needs from a database. Our restaurant database consists of the following components: restaurant name, cuisine, location, price range, guidance. The information needs are basically accepted by a slot filling dialog. The conversation finished when the operator was able to narrow down the search to a single restaurant.

b) Non-task-oriented dialog corpus

We created a non-task-oriented dialog corpus in section 3.2.1. We use four dialogs where by an operator (thirties, female) and one of four subjects (twenties

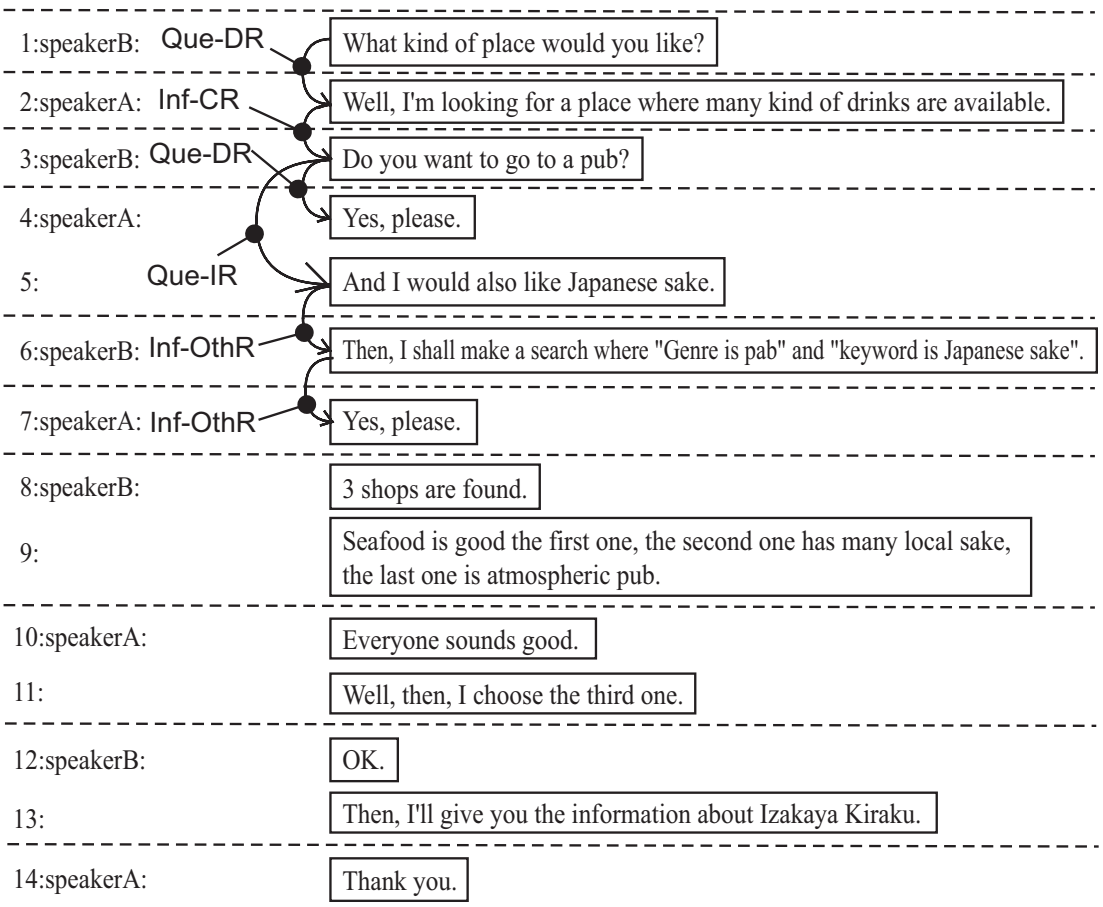


Figure 3.8: An example from a task-oriented dialog corpus

and sixties, equal numbers of males and females) were participated.

Figure 3.8 and Figure 3.9 show an excerpt of our corpus. A dotted line in the figures indicates the boundary of a *turn*, while a box indicates an *utterance*. For convenience, in this section, we refer to the subject by **speakerA**, and the operator by **speakerB**.

Both types of our corpora are mixed-initiative. It is known that the mixed-initiative dialog is more complex than the dialog in which only one of the speakers has an initiative[12]. We analyzed a part of our corpus where one of the speakers has an initiative so as to analyze the differences between task-oriented dialogs and non-task-oriented dialogs more accurately. For the task-oriented setting, we analyzed the dialog segments where speakerA informs speakerB of his/her information needs. For example, in Figure 3.8, we analyze utterances from the first line to the seventh line. For the

non-task-oriented dialogs, on the other hand, we analyzed the dialog segments where speakerA has the initiative. For example, in Figure 3.9, all of utterances are the target of analysis because speakerA informs speakerB of her family and speakerA takes the overall initiative.

3.3.2 Definition of exchange tags

As we mentioned at the beginning of this chapter, we investigate what type of utterances appears remarkably in non-task-oriented dialogs. For this end, we make a comparison between task-oriented dialog and non-task-oriented dialog. In task-oriented dialog, dialog become briefly because efficient dialog strategies have been required to achieve a given task. On the other hand, in non-task-oriented dialog, speakers often try to take a lead to continue the exchange because their aim is conversation itself. In this section, We investigate what type of utterances contribute to continuing exchanges, we analyze dialogs by focusing on exchanging utterances.

According to the findings in discourse analysis, a dialog is unfolded by exchanging utterances. Exchange of utterances is called an *exchange* in discourse analysis. We analyze the exchanges in our task-oriented and non-task-oriented dialog corpora to investigate the types of utterances are different between these dialogs.

In general, it is known that an exchange consists of three phases: **initiation**, **response**, and **follow-up** [67, 66]. To analyze exchanges more accurately, we define exchange types as follows.

1) Initiation

Initiation is sub-classified into three types: **Question (⟨Que⟩)**, **Semi Question (⟨Semi-Q⟩)**, and **Inform (⟨Inf⟩)**. These types are distinguished by whether the utterance requires any response. ⟨**Que**⟩ is an exchange where the speaker asks the hearer for something and requests the hearer's response strongly. ⟨**Semi-Q**⟩ is an exchange where the speaker does not requests the hearer for his/her response but the hearer actually responses to the speaker's implicit question. ⟨**Inf**⟩ covers all the other kind of initiation exchanges.

For example, in Figure 3.9, utterance2 is ⟨**Que**⟩ because speakerB asks speakerA and requests speakerA for some response strongly. While utterance16 is ⟨**Semi-Q**⟩ because speakerA responses to speakerB's utterance even though speakerB does not request any answer strongly. The utterance on the first line indicates

⟨**Inf**⟩ because the utterance does not neither explicit question nor implicit question.

2) Response

Let us first consider responses to explicit/implicit questions (⟨Que⟩ and ⟨Semi-Q⟩). Previous studies discuss the taxonomy of the response to a question in the context of question answering [7, 10, 69]. Benamara et al. taxonomies cooperative responses in the travel dialog [7]. Yamada et al. analyze human-to-human dialogs for an aircraft time guide and classify cooperative responses into 12 types. In our corpus analysis, however, we distinguish only two types, **Direct Response (DR)** and **Indirect Response (IR)**, because most of the fine-grained response types defined in the literature are not really relevant in non-task-oriented dialogs. **DR** indicates that the target utterance answers to a question directly, while **IR** indicates the utterance answers to a question indirectly. DR and IR are distinguished by a surface expression. For example, in Figure 3.9, utterance3 is of the **DR** class because speakerB answers to the speakerB's question directly. In contrast, the utterance4 is of **IR** because speakerA gives additional information in response to speakerB's question.

Next, consider responses to informing utterances (⟨Inf⟩). Rieser et al. taxonomies the system's clarification requests (CRs) to the user's initiating utterances in task-oriented dialogs [43]. For example, in their annotation schema, *lex* is annotated to a lexical question (e.g. *What's a double torx?*), *np-ref* is annotated to a question of reference resolution (e.g. *Which square?*). In the example below, speakerB asks speakerA because there is vagueness in the noun phrase of speakerA's utterance. That is why *np-ref* is annotated in this case.

speakerA: *I would like to book a flight on Monday.*

speakerB: *Which Monday?* ⟨*np-ref*⟩

The annotation schema proposed by Rieser et al. is useful for detailed analysis of CRs. However, only **Clarification requests (CRs)** is adapted in our annotation schema because their categories tend to be too fine-grained clarification types. Other response utterance to an inform utterance is annotated as **Other Response (OthR)**.

3) Follow-up

Table 3.6: Definition of exchange tags

	exchange tags	definition
Initiation	Question (⟨Que⟩)	The utterance where the speaker asks the hearer for something and requests the hearer's response strongly.
	Semi Question (⟨Semi-Q⟩)	The utterance where the speaker does not requests the hearer for his/her response but the hearer actually responses to the speaker's implicit question.
	Inform (⟨Inf⟩)	All the other kind of initiation utterance.
Response	Direct response (DR)	The utterance where the speaker answers to the another speaker's question directly.
	Indirect response (IR)	The utterance where the speaker answers to the another speaker's question indirectly.
	Clarification requests (CRs)	The utterance where the speaker clarifies the another speaker's informing utterance.
	Other Response (OthR)	All the other kind of Response utterance.
Follow-up	Follow-up (F)	The utterance shows receiving information and comment to another speaker's utterance.

Follow-up (F) utterance shows receiving information and comment to another speaker's utterance. We do not subdivide follow-up, the definition in previous work is used.

The overall exchange tag set, we adopt is summarized in Table 3.6.

3.3.3 Annotation to exchanges

Figure 3.8 and Figure 3.9 show examples of exchange tag annotation. The following describes the process of annotation using these dialogs.

Step1. Identifying adjacency pairs

Adjacency pairs are annotated by listening to the target dialog. Our definition of adjacency pairs follows in MRDA [14]. MRDA tends to identify more utterance pairs as an adjacency pair than traditional conversation analysis. For example, utterance4 in Figure 3.9 (*Dinner is not prepared at all*) and utterance6 (*I thought that dinner becomes simple. But dinner is not prepared, is it?*) would not considered as an adjacency pair in conversation analysis. However, in MRDA, they are identified as an adjacency pair because utterance can be interpreted as an inducement for speakerB to utterance6. We adopt this broad definition in MRDA

because we believe that it is important to capture such indirect responses in analyzing non-task-oriented dialogs.

Step2. Annotation of exchange tags

Once adjacency pairs identified, our annotation process moves on to the annotation of exchange tags. The process goes to step (2-1) if a given utterance is an initiating utterance, step (2-2) if it is a response, step (2-3) if it is a follow-up. If a given utterance functions as both initiation and response, we do steps (2-1) and (2-2).

(2-1) According to the definition shown in Table 3.6, annotate ⟨**Que**⟩ or ⟨**Semi-Q**⟩ or ⟨**Inf**⟩. These types are distinguished by whether the target utterance requires any response.

(2-2) The process goes to step (2-2-1) if a given utterance is a question, step (2-2-2) if it is not question but a preceding utterance is a question, step (2-2-3) in other cases.

(2-2-1) Annotate **CRs**.

(2-2-2) According to the definition shown in Table 3.6, annotate **DR** or **IR**. These types are distinguished by whether the target utterance answers to a question directly or not.

(2-2-3) Annotate **OthR**.

(2-3) Annotate **F**.

3.3.4 Reliability of exchange tags

In this section, we report the inter-annotator agreement in our exchange tagging. Two annotators (A1 and A2) independently annotated two dialogs (189 utterances in total) sampled from our non-task-oriented dialog corpus (see 3.3.1). MMAX2 annotation tool was used for the annotation [38][3]. The annotators were requested to listen to the speech while annotating the data. The results are the following:

[3]http://www.eml-research.de/english/research/nlp/download/mmax.php

Table 3.7: Result of reliability of adjacency pairs

speaker	number of annotated utterances	Total	Agreement	Agreement(%)
dialog1	88	67	65	97.0
dialog2	101	69	66	95.7

Table 3.8: The result of reliability of exchange tags

	Number of exchange	κ
dialog1	65	0.83
dialog2	66	0.84

(1) Step 1: Reliability of adjacency pairs

We compute the inter-annotator agreement ratios by equation (3.6).

$$\text{agreement} = \frac{\text{The number of agreed Exchange tags}}{\text{Total of Exchange tags annotated by A1 and A2}} \times 100 \qquad (3.6)$$

The results are shown in Table 3.7. The *Total* in the Table shows the total number of the exchanges annotated and *Agreement* shows the number of the agreed exchange tags. The results indicate that the annotation of adjacency pairs is sufficiently reliable.

(2) Step 2: Reliability of exchange tags

Next, we investigated the reliability of exchange tags. We used only the utterances for which the two annotators agreed in adjacency pair identification. The two annotators independently annotated exchange tags. The Cohen's Kappa is used to measure the inter-annotator agreement [11].

The results are shown in Table 3.8. The κ value is above 0.8 for both dialog, which indicates that the annotated exchange tags are sufficiently reliable.

3.3.5 Analysis of exchanges in non-task-oriented dialog

The result of the inter-annotator agreement shows that the annotated exchange tags are sufficiently reliable. One of the annotators then annotated a task-oriented and non-task-oriented dialog corpora (see 3.3.1). Table 3.9 shows the size of our exchange-tagged corpora. In this section, we investigate the difference between task-oriented

Table 3.9: Size of our exchange-tagged corpora

	utterance	exchange
Task-oriented dialog	194	115
Non-task-oriented dialog	686	501

and non-task-oriented dialogs. We then investigate what are the discriminating characteristics that differentiate them.

3.3.5.1 Tendency of exchange tags

Figure 3.10 shows the ratio of frequency of exchange tags in task-oriented and non-task-oriented dialogs. The vertical axis in Figure 3.10 shows the ratio of frequency of exchange tags. The number above each bar shows the frequency of exchange tags. For example, concerning the exchange class ⟨*Que*⟩ in a non-task-oriented dialog, the frequency is 152 and that accounts for 18.3% of the total of exchange tags in a non-task-oriented dialog.

Let us first consider the results of initiating utterances. ⟨Que⟩ appeared more frequently in a non-task-oriented dialog. On the other hand, ⟨Semi-Q⟩ appeared more frequently in a task-oriented dialog. Figure 3.11 shows the example of ⟨Semi-Q⟩ in a task-oriented dialog. In this example, speakerA interpreted speakerB's utterance as *what kind of place would you like?*. SpeakerA then answered *I would like to go to a pub*. As this example shows, an efficient dialog is accomplished by reading a speaker's real intention in a task-oriented dialog. In a task-oriented dialog, hearer would be able to guess speaker's real intention of his/her implicit question because they share a given task. On the other hand, it is difficult to foresee speaker's real intention in a non-task-oriented dialog because speakers are sharing little information mutually in a non-task-oriented dialog. The frequency of ⟨Que⟩, an explicit question, is therefore high in a non-task-oriented dialog. Meanwhile, inform utterances appeared more frequently in both dialogs. It is reported that statement utterances[4] appear a lot in the ICSI meeting corpus and the Switchboard corpus[22, 48]. Concerning an inform utterance, we reached the same result with previous works.

Next, consider the results of response utterances. IR and CRs appeared more frequently in a non-task-oriented dialog. A detailed analysis concerning IR and CRs is described in next paragraph.

[4]Statement utterance is corresponding to inform utterance in our corpora.

On the other hand, there is no difference in the frequency of follow-up utterance. The result indicates that a follow-up utterance is indispensable in both a task-oriented dialog and a non-task-oriented dialog.

3.3.5.2 Tendency of frequency of exchange tag

Figure 3.12 shows the ratio of frequency of exchanges in task-oriented and non-task-oriented dialogs. The *Initiation* in Figure 3.12 indicates the utterances annotated by initiation. The *Response* indicates the utterances annotated by response. The *Initiation/Response* indicates the utterances annotated by initiation and response. The *Follow-up* indicates the utterances annotated by follow-up.

According to Figure 3.12, we found that initiation/response utterances appeared more frequently in a non-task-oriented dialog. An efficient dialog is required in a task-oriented dialog, while the enthusiastic dialog is required in a non-task-oriented dialog. Figure 3.12 suggested that speakers often try to take a lead to continue the exchange smoothly in a non-task-oriented dialog.

Figure 3.13 shows the frequency of exchange tags of response and initiation/response utterances. The numbers in Figure 3.13 shows the frequency, the vertical line shows the ratio of the frequency. Figure 3.13 indicates that IR and CRs often work as an initiation/response utterance. According to the definition of initiation/response, an utterance continues after the initiation/response utterance. On the other hand, as described in paragraph 3.3.3, exchange tags are annotated using just surface information of an utterance. Therefore, it is not necessarily that any utterances continue after IR and CRs utterance. However, Figure 3.13 shows that 83.7% (82 cases in 98 cases) of IR and 91.7% (132 cases in 144 cases) of CRs became initiation/response utterances. The result suggests that IR and CRs contribute to make chains of utterances in non-task-oriented dialog. As we described above, continuing exchanges works on a non-task-oriented dialog because speaker's aim is a conversation itself. So, we can conclude that IR and CRs is important for a non-task-oriented dialog.

In addition, we analyzed the examples of IR and CRs. We then found that IR works as an initiation/response utterance because additional information of it leads next utterance. While CRs works as an initiation/response utterance because an explicit question leads next utterance. We will explain using examples. Let us consider the exchanges from utterance2 to utterance5 in Figure 3.9. In this case, if speakerA answered only *yes* to the speakerB's utterance2(*Your mother says so?*), the exchange might be end.

However, in fact, speaker shows the additional information in utterance4(*Dinner is not prepared at all*). SpeakerB then told *I thought dinner becomes simple. But dinner is not prepared, is it?* using the additional information shown in utterance4. In addition, it is also considered that the utterance8 leads the utterance9. Next, we will show the examples of CRs. The explicit questions on utterance6 and utterance9 lead next utterances. We can conclude that an initiation/response utterance is important for a non-task-oriented dialog because it lead next utterance naturally and make chains of utterances in non-task-oriented dialog.

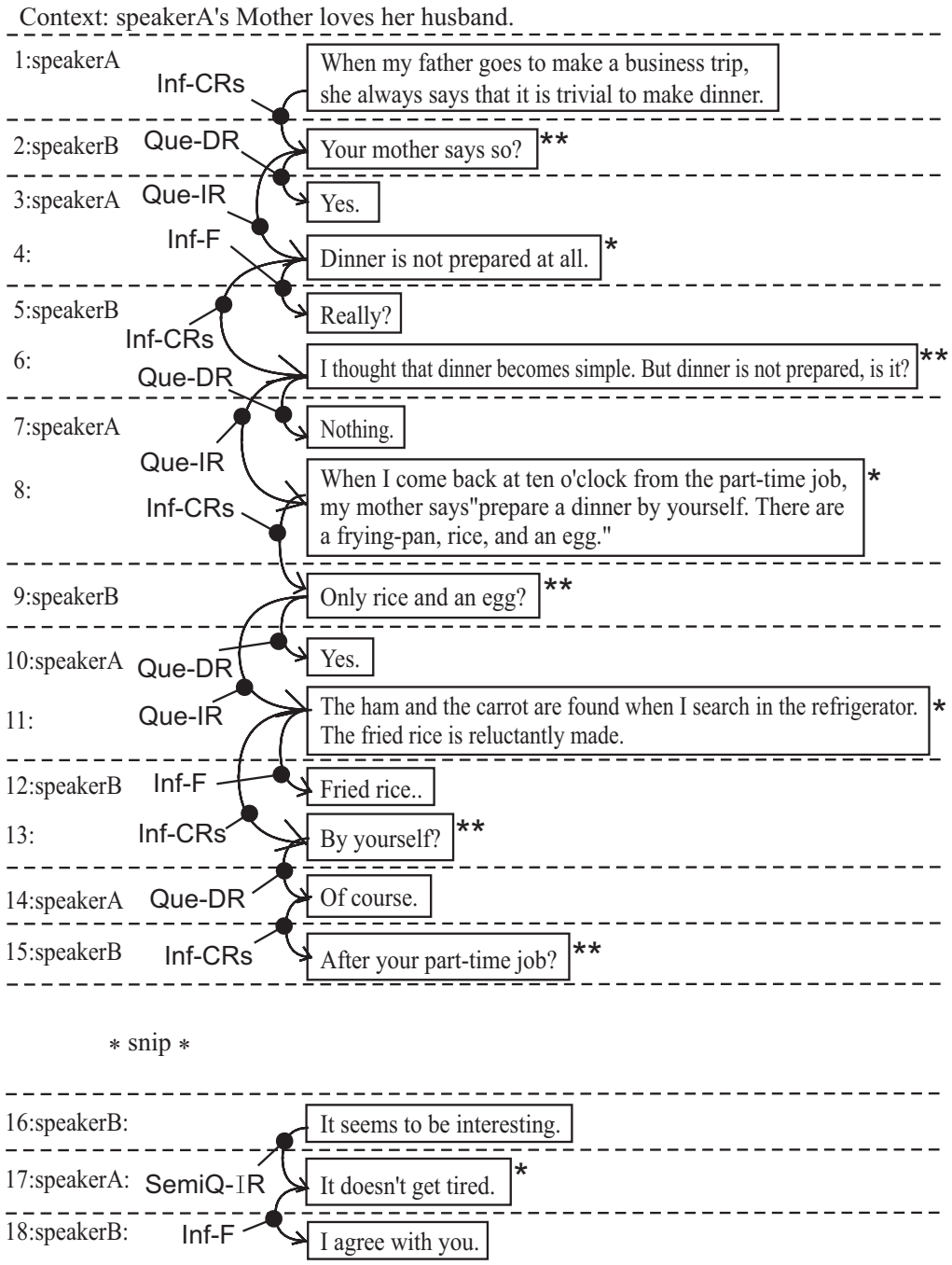
3.4 Summary

In this section, we analyzed non-task-oriented dialog from a conversation analysis viewpoint.

First, we analyzed the relationship between utterances and Enthusiasm in a non-task-oriented human-to-human dialog. We first created a non-task-oriented dialog corpus annotated with two types of tags: DAs/RRs and Enthusiasm. The DA and RR tagging schema was adapted from the definition given in a previous work for our corpus, and an Enthusiasm tagging schema is proposed. Our method of rating Enthusiasm enables the observation of the fluctuation of Enthusiasm, which enables the detailed analysis of utterances and Enthusiasm. The result of the analysis shows the frequency of objective and subjective utterances related to the level of Enthusiasm. We also found that affective and cooperative utterances are significant in an enthusiastic dialog.

Next, we created two types of human-to-human dialog corpora: task-oriented and non-task-oriented dialogs. We investigate what are the discriminating characteristics that differentiate them. We found that initiation/response utterance appeared more frequently in non-task-oriented dialog. This is because speakers often try to take a lead to continue the exchange smoothly in a non-task-oriented dialog. In addition, we also found that most indirect response and clarification requests work as initiation/response utterance in a non-task-oriented dialog.

In next chapter, we will develop a non-task-oriented dialog system using these finding.



※ The symbol * shows indirect answer, while the symbol ** shows clarification

Figure 3.9: An example from a non-task-oriented dialog corpus

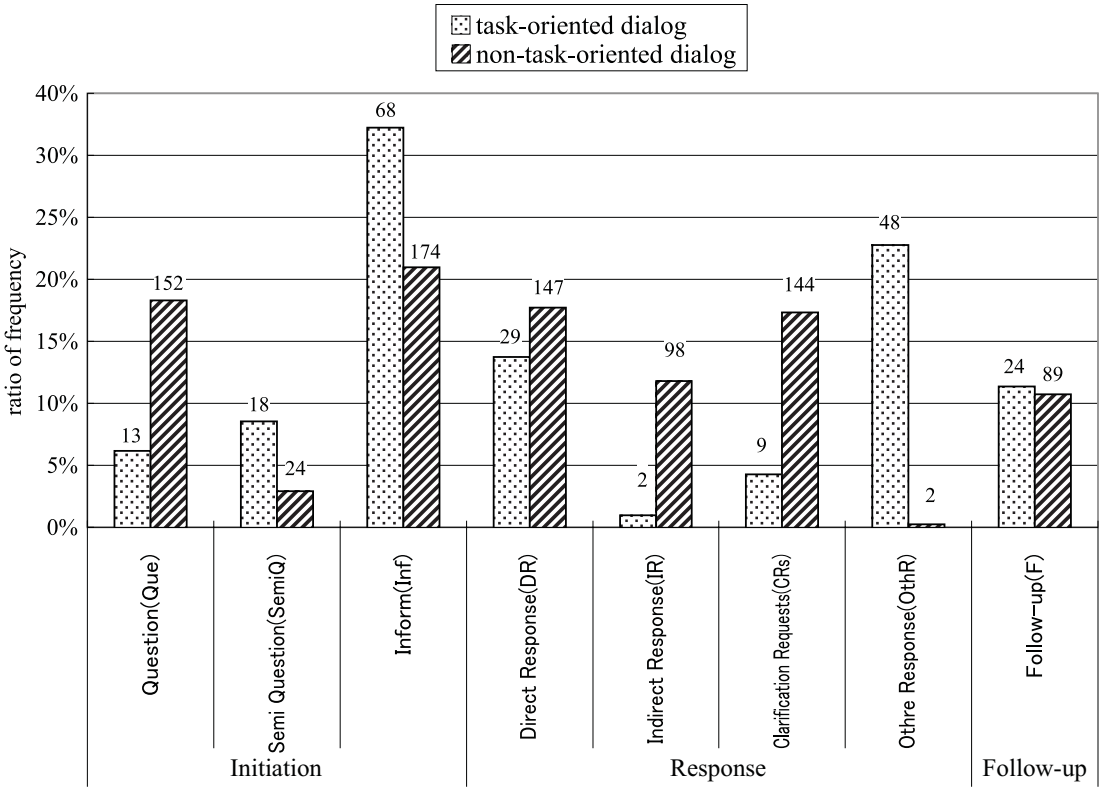


Figure 3.10: Ratio of the frequency of exchange tags in task-oriented and non-task-oriented dialogs

Dialog		Interpretation of speakerA
1:speakerB	You have not decided the genre. ***	What kind of place would you like?
2:speakerA	I would like to go to a pub.	

※ the utterance annotated symbol *** shows semi question.

Figure 3.11: example of semi question in task-oriented dialog

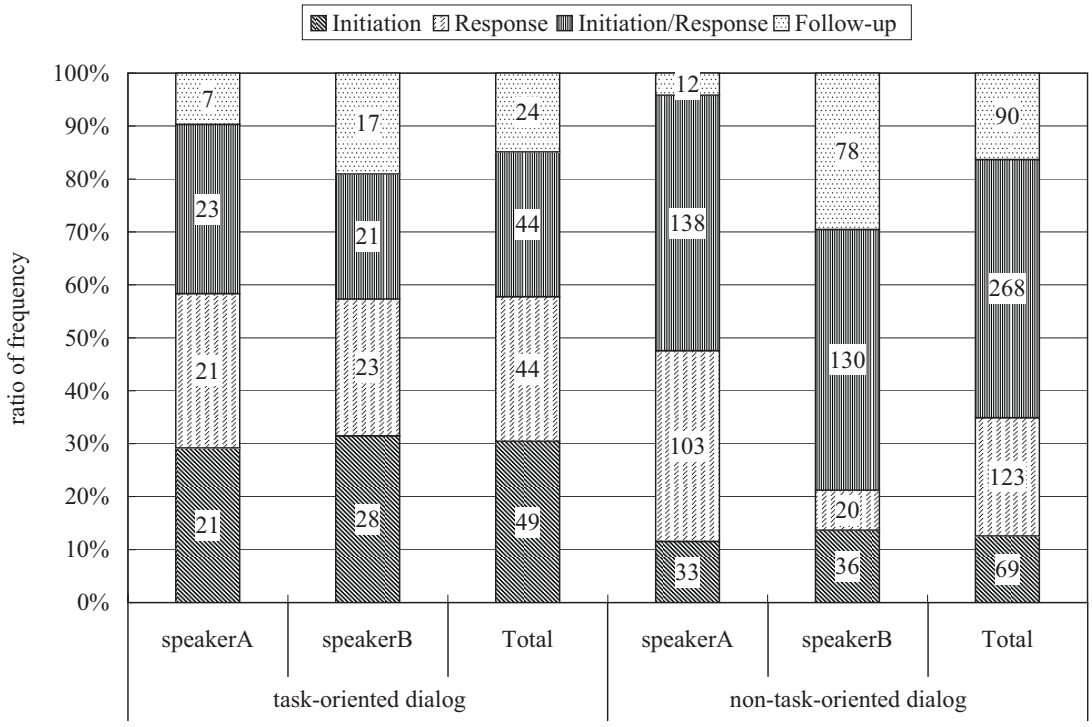


Figure 3.12: Ratio of frequency of exchanges in task-oriented and non-task-oriented dialogs

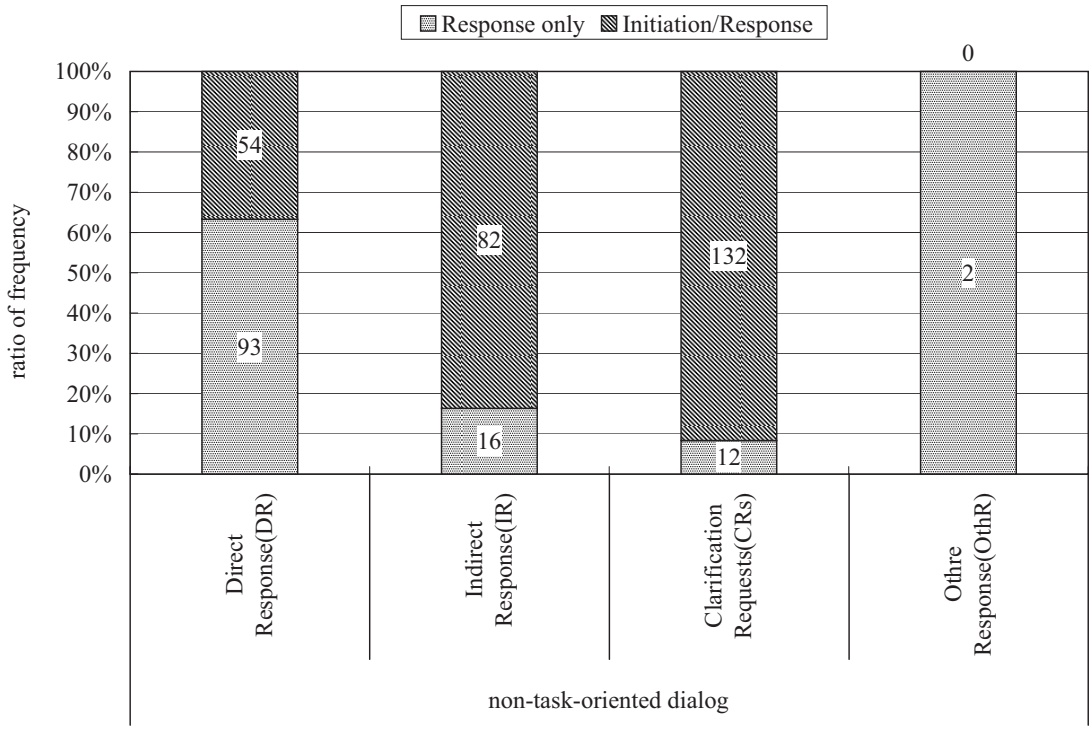


Figure 3.13: Ratio of response and initiation/response utterance

Chapter 4

Emotion classification using massive examples extracted from the web

4.1 Introduction

In the preceding chapter, the following two issues are addressed: what makes a non-task-oriented human-to-human enthusiastic dialog; what are the differences between task-oriented dialogs and non-task-oriented dialog. As a result, we found that one of the significant utterances is an affective utterance in a non-task-oriented dialog.

Based on the finding, in this chapter, we focus on generating of affective utterances. More concretely, this chapter describes the method of generating the response that shows a sympathy for the user utterance. For example, given an utterence *I traveled far to get to the shop, but it was closed* from the user, if the system could infer the user's emotion behind it, it would know that it would be appropriate to say *That's too bad* or *That's really disappointing*. It can be easily imagined that such affective behaviors of a dialog system would be beneficial not only for communication robots but also for a wide variety of dialog purposes including even task-oriented dialogs.

To be capable of generating sympathetic responses, a dialog system needs a computational model that can infer the user's emotion behind his/her utterence. There have been a range of studies for building a model for classifying a user's emotions based on acoustic-prosodic features and facial expressions [40, etc.]. Such methods are, however, severely limited in that they tend to work well only when the user expresses his/her emotions by "exaggerated" prosodic or facial expressions. Furthermore, what is required in generating sympathetic responses is the identification of the user's emotion in a finer grain-size. For example, in contrast to the above example of *disappoint-*

ing, one may expect the response to *My pet parrot died yesterday* should be *That's really sad*, wheras the response to *I may have forgotten to lock my house* should be *You're worried about that*.

In this chapter, we address the above issue of emotion classification in the context of human-computer dialog, and demonstrate that massive examples of emotion-provoking events can be extracted from the Web with a reasonable accuracy and those examples can be used to build a semantic content-based model for fine-grained emotion classification.

4.2 Related work

Recently, several studies have reported about dialog systems that are capable of classifying emotions in a human-computer dialog [5, 4, 31, 44]. ITSPOKE is a tutoring dialog system, that can recognize the user's emotion using acoustic-prosodic features and lexical features. However, the emotion classes are limited to *Uncertain* and *Non-Uncertain* because the purpose of ITSPOKE is to recognize the user's problem or discomfort in a tutoring dialog. Our goal, on the other hand, is to classify the user's emotions into more fine-grained emotion classes.

In a more general research context, while quite a few studies have been presented about opinion mining and sentiment analysis [33], research into fine-grained emotion classification has emerged only recently. There are two approaches commonly used in emotion classification: a rule-based approach and a statistical approach. Masum et al. [36] and Chaumartin [9] propose a rule-based approach to emotion classification. Chaumartin has developed a linguistic rule-based system, which classifies the emotions engendered by news headlines using the WordNet, SentiWordNet, and WordNet-Affect lexical resources. The system detects the sentiment polarity for each word in a news headline based on linguistic resources, and then attempts emotion classification by using rules based on its knowledge of sentence structures. The recall of this system is low, however, because of the limited coverage of the lexical resources. Regarding the statistical approach, Kozareva et al. (2007) apply the theory of Hatzivassiloglou et al. [20] and Turney [54] to emotion classification and propose a method based on the co-occurrence distribution over content words and six emotion words (e.g. joy, fear). For example, *birthday* appears more often with *joy*, while *war* appears more often with *fear*. However, the accuracy achieved by their method is not practical in applications assumed in this chapter. As we demonstrate in Section 4.4.2, our method significantly

outperforms Kozareva's method.

4.3 Emotion classification

4.3.1 The basic idea

We consider the task of emotion classification as a classification problem where a given input sentence (a user's utterance) is to be classified either into such 10 emotion classes as given later in Table 4.1 or as ⟨neutral⟩ if no emotion is involved in the input. Since it is a classification problem, the task should be approached straightforwardly in a variety of machine learning-based methods if a sufficient number of labelled examples were available. Our basic idea is to learn what emotion is typically provoked in what situation, from massive examples that can be collected from the Web. The development of this approach and its subsequent implementation has forced us to consider the following two issues.

First, we have to consider the quantity and accuracy of emotion-provoking examples to be collected. The process we use to collect emotion-provoking examples is illustrated in the upper half of Figure 4.1. For example, from the sentence *I was disappointed because the shop was closed and I'd I traveled a long way to get there*, pulled from the Web, we learn that the clause *the shop was closed and I'd traveled a long way to get there* is an example of an event that provokes ⟨disappointment⟩. In this chapter, we refer to such an example as an *emotion-provoking event* and a collection of event-provoking events as an *emotion-provoking event corpus (an EP corpus)*. Details are described in Section 4.3.2.

Second, assuming that an EP corpus can be obtained, the next issue is how to use it for our emotion classification task. We propose a method whereby an input utterance (sentence) is classified in two steps, sentiment polarity classification followed by fine-grained emotion classification as shown in the lower half of Figure 4.1. Details are given in Sections 4.3.3 and 4.3.4.

4.3.2 Building an EP corpus

We used ten emotions *happiness, pleasantness, relief, fear, sadness, disappointment, unpleasantness, loneliness, anxiety, anger* in our emotion classification experiment. First, we built a hand-crafted lexicon of emotion words classified into the ten emotions.

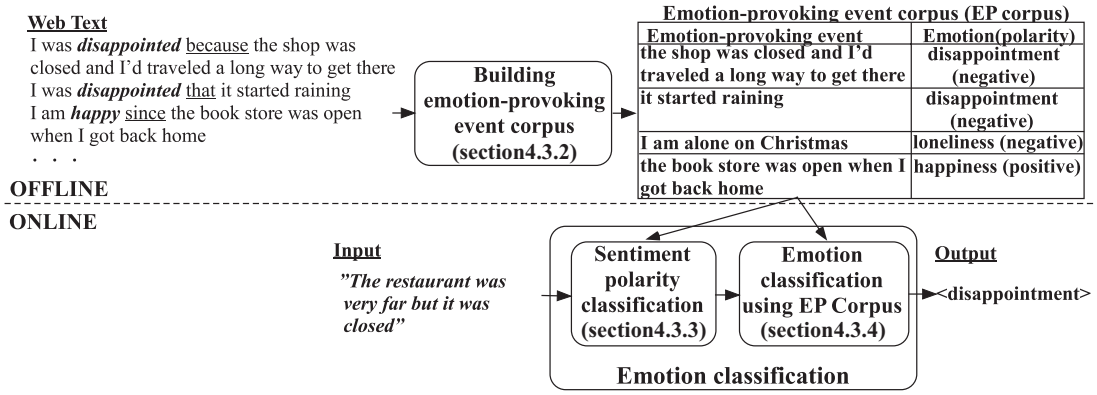


Figure 4.1: Overview of our approach to emotion classification

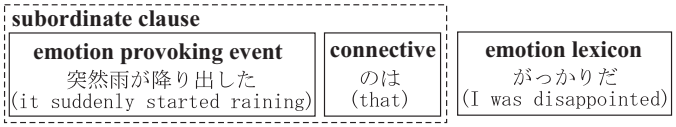


Figure 4.2: An example of a lexico-syntactic pattern

From the Japanese Evaluation Expression Dictionary created by Kobayashi et al. [24], we identified 349 emotion words based on the definition of emotion words proposed by Teramura [52]. The distribution is shown in Table 4.1 with major examples.

We then went on to find sentences in the Web corpus that possibly contain emotion-provoking events. A subordinate clause was extracted as an emotion-provoking event instance if (a) it was subordinated to a matrix clause headed by an emotion word and (b) the relation between the subordinate and matrix clauses is marked by one of the following eight connectives: ので, から, ため, て, のは, のが, ことは, ことが. An example is given in Figure 4.2. In the sentence "突然雨が降り出した のは がっかりだ *(I was disappointed that it suddenly started raining)*", the subordinate clause "突然雨が降り出した *(it suddenly started raining)*" modifies "がっかりだ *(I was disappointed)*" with the connective "のは *(that)*". In this case, therefore, the event mention "突然雨が降り出した *(it suddenly started raining)*" is learned as an event instance that provokes⟨disappointment⟩.

Applying the emotion lexicon and the lexical patterns to the Japanese Web corpus [23], which contains 500 million sentences, we were able to collect about 1.3 million events as causes of emotion. The distribution is shown in Table 4.2.

Table 4.1: Distribution of the emotion expressions and examples

Sentiment Polarity	10 Emotion Classes	Emotion lexicon (349 Japanese emotion words)	
		Total	Examples
Positive	happiness	90	嬉しい (happy), 狂喜 (joyful), 喜ぶ (glad), 歓ぶ (glad)
	pleasantness	7	楽しい (pleasant), 楽しむ (enjoy), 楽しめる (can enjoy)
	relief	5	安心 (relief), ほっと (relief)
Negative	fear	22	恐い (fear), 怖い (fear), 恐ろしい (frightening)
	sadness	21	悲しい (sad), 哀しい (sad), 悲しむ (feel sad)
	disappointment	15	がっかり (lose heart), がっくり (drop *one's* head)
	unpleasantness	109	嫌 (disgust), 嫌がる (dislike), 嫌い (dislike)
	loneliness	15	寂しい (lonely), 淋しい (lonely), わびしい (lonely)
	anxiety	17	不安 (anxiety), 心配 (anxiety), 気がかり (worry)
	anger	48	腹立たしい (angry), 腹立つ (get angry), 立腹 (angry)

Table 4.2: Number of emotion-provoking events

10 Emotions	EP event	10 Emotions	EP event
happiness	387,275	disappoint-ment	106,284
pleasantness	209,682	unpleasantness	396,002
relief	46,228	loneliness	26,493
fear	49,516	anxiety	45,018
sadness	31,369	anger	8,478

Tables 4.3 and 4.4 show the results of our evaluation for the resultant EP corpus. One annotator, who was not the developer of the EP corpus, evaluated 2000 randomly chosen events. The "Polarity" column in Table 4.3 shows the results of evaluating whether the sentiment polarity of each event is correctly labelled, whereas the"Emotion" column shows the correctness at the level of the 10 emotion classes. The correctness of each example was evaluated as exemplified in Table 4.4. *Correct* indicates a correct example, *Contex-dep.* indicates a context-dependent example, and *Error* is an error example. For example, in the case of *There are a lot of enemies* in Table 4.4, the "Polarity" is *Correct* because it represents a negative emotion. However, its emotion class⟨unpleasantness⟩ is judged *Context-dep*.

As Table 4.3 shows, the Sentiment Polarity is correct in 57.0% of cases and partially correct (Correct + Context-dep.) in 90.9% of cases. On the other hand, the Emotion is correct in only 49.4% of cases and partially correct in 73.9% of cases. These figures

Table 4.3: Correctness of samples from the EP corpus

	Polarity	Emotion
Correct	1140 (57.0%)	988 (49.4%)
Context-dep.	678 (33.9%)	489 (24.5%)
Error	182 (9.1%)	523 (26.2%)

Table 4.4: Examples from in the EP corpus

EP-Corpus			Result of evaluation	
Emotion-provoking Event	Emotion word	10 Emotions (P/N)	Polarity	Emotion
花持ちが悪い (A flower died quickly)	残念だ (disappointed)	⟨disappointment(N)⟩	Correct	Correct
敵が多い (There are a lot of enemies)	飽きる (lose interest)	⟨unpleasantness(N)⟩	Correct	Context-dep.
ちんげん菜が多い (There is a lot of Chinese cabbage)	嬉しい (happy)	⟨happiness(P)⟩	Context-dep.	Context-dep.
ジュースが飲みたい (I would like to drink orange juice)	大変だ (terrible)	⟨unpleasantness(N)⟩	Error	Error

may not seem very impressive. As far as its impact on the emotion classification accuracy is concerned, however, the use of our EP corpus, which requires no supervision, makes remarkable improvements upon Kozareva et al. (2007)'s unsupervised method as we show later.

4.3.3 Sentiment polarity classification

Given the large collection of emotion-labelled examples, it may seem straightforward to develop a trainable model for emotion classification. Before moving on to emotion classification, however, it should be noted that a user's input utterance may not involve any emotion. For example, if the user gives an utterance *I have a lunch at the school cafeteria every day*, it is not appropriate for the system to make any sympathetic response. In such a case, the user's input should be classified as ⟨neutral⟩.

The classification between *emotion-involved* and *neutral* is not necessarily a simple

Table 4.5: Distribution of the Sentiment polarity of sentences randomly sampled from the Web

Sentiment Polarity	Number	Ratio
positive	650	65.0%
negative	153	15.3%
neutral	117	11.7%
Context-dep.	80	8.0%

problem, however, because we have not found yet any practical method for collecting training examples of the class⟨neutral⟩. We cannot rely on the analogy to the pattern-based method we have adopted to collect emotion-provoking events — there seems no reliable lexico-syntactic pattern for extracting neutral examples. Alternatively, if the majority of the sentences on the Web were neutral, one would simply use a set of randomly sampled sentences as labelled data for ⟨neutral⟩. This strategy, however, does not work because neutral sentences are not the majority in real Web texts. As an attempt, we collected 1000 sentences randomly from the Web and investigated their distribution of sentiment polarity. The results, shown in Table 4.5, revealed that the ratio of neutral events was unexpectedly low. These results indicate the difficulty of collecting neutral events from Web documents.

Taking this problem into account, we adopt a two-step approach, where we first classify a given input into three sentiment polarity classes, either positive, negative or neutral, and then classify only those judged positive or negative into our 10 fine-grained emotion classes. In the first step, i.e. sentiment polarity classification, we use only the positive and negative examples stored in the EP corpus and assume sentence to be neutral if the output of the classification model is near the decision boundary. There are additional advantages in this approach. First, it is generally known that performing fine-grained classification after coarse classification often provides good results particularly when the number of the classes is large. Second, in the context of dialog, a misunderstanding the user's emotion at the sentiment polarity level would be a disaster. Imagine that the system says *You must be happy* when the user in fact feels sad. As we show in Subsection 4.2, such fatal errors can be reduced by taking the two-step approach.

Various methods have already been proposed for sentiment polarity classification, ranging from the use of co-occurrence with typical positive and negative words Turney [54] to bag of words Pang et al. [39] and dependency structure Kudo et al. [28].

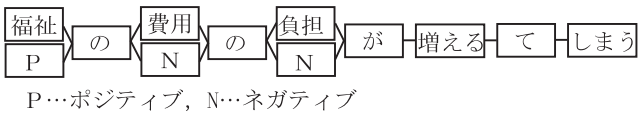


Figure 4.3: An example of a word-polarity lattice

Our sentiment polarity classification model is trained with SVMs [56], and the features are {1-gram, 2-gram, 3-gram} of words and the sentiment polarity of the words themselves. Figure 4.3 illustrates how the sentence "福祉の費用の負担が増えてしまう *(The cost of welfare increases)*" is encoded to a feature vector. Here we assume the sentiment polarity of the "福祉 *(welfare)*" is positive, while the "費用 *(cost)*" and "負担 *(cost)*" are negative. These polarity values are represented in parallel with the corresponding words, as shown in Figure 4.3. By expanding {1-gram, 2-gram, 3-gram} in this lattice representation, the following list of features are extracted: 福祉 *(welfare), Positive,* 福祉 *(welfare)-*の *(of), Positive-*の *(of),* 福祉 *(welfare)-*の *(of)-*費用 *(cost), etc.*.

4.3.4 Emotion classification

For fine-grained emotion classification, we propose a k-nearest-neighbor approach (kNN) using the EP corpus.

Given an input utterance, the kNN model retrieves k-most similar labelled examples from the EP corpus. Given the input *The restaurant was very far but it was closed* as Figure 4.1, for example, the kNN model finds similar labelled examples, say, labelled example {the shop was closed and I'd traveled far to get there} in the EP corpus. For the similarity measure, we use cosine similarity between bag-of-words vectors; $sim(I, EP) = \frac{I \cdot EP}{|I||EP|}$ for input sentence I and an emotion-provoking event EPin the EP corpus. The score of each class is given by the sum of its similarity scores.

An example is presented in Figure 4.4. The emotion of the most similar event is ⟨disappointment⟩, that of the second-most similar event is ⟨unpleasantness⟩ tied with ⟨loneliness⟩. After calculating the sum for each emotion, the system outputs ⟨loneliness⟩ as the emotion for the input I because the score for ⟨loneliness⟩ is the highest.

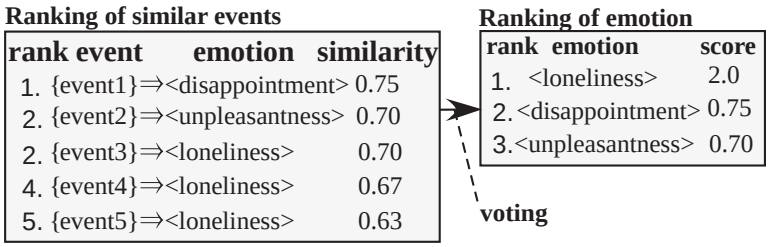


Figure 4.4: Emotion Classification by kNN (k=5)

4.4 Experiments

4.4.1 Sentiment polarity classification

We conducted experiments on sentiment polarity classification using the following two test sets:

TestSet1: The first test set was a set of utterances which 6 subject speakers produced interacting with our prototype dialog system. This data include 31 positive utterances, 34 negative utterances, and 25 neutral utterances.

TestSet2: For the second test set, we used the 1140 samples that were judged *Correct* with respect to sentiment polarity in Table 4.3. 491 samples (43.1%) were positive and 649 (56.9%) were negative. We then added 501 neutral sentences newly sampled from the Web. These samples are disjoint from the EP corpus used for training classifiers.

For each test set, we tested our sentiment polarity classifier in both the two-class (positive/negative) setting, where only positive or negative test samples were used, and the three-class (positive/negative/neutral) setting. The performance was evaluated in F-measure.

Table 4.6 shows the results for the two-class setting, whereas Table 4.7 shows the results for the three-class. "Word" denotes the model trained with only word n-gram features, whereas "Word+Polarity" denotes the model trained with n-gram features extracted from a word-polarity lattice (see Figure 4.3). The polarity value of each word is defined in Takamura's sentiment polarity dictionary, which includes 2349 positive words and 5866 negative words [50][1] .

[1]We only use the words whose confidence values are over 0.7 in the dictionary.

Table 4.6: F-values of sentiment polarity classification (positive/negative)

corpus size	Word				Word+Polarity				Word+Polarity+Dependency			
	TestSet1		TestSet2		TestSet1		TestSet2		TestSet1		TestSet2	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
1/1000	0.559	0.536	0.487	0.589	0.564	0.530	0.535*	0.572	0.565	0.523	0.537*	0.575
1/100	0.643	0.520	0.553	0.556	0.672	0.569	0.616*	0.573	0.674	0.582	0.617*	0.571
1/10	0.721	0.765	0.660	0.776	0.727	0.766	0.722*	0.790	0.722	0.764	0.712*	0.786
1.3M	0.839	0.853	0.794	0.842	0.839	0.853	0.808	0.849	0.820	0.841	0.804	0.850

Table 4.7: F-values of sentiment polarity classification (positive/negative/neutral)

corpus size	Word				Word+Polarity				Word+Polarity+Dependency			
	TestSet1		TestSet2		TestSet1		TestSet2		TestSet1		TestSet2	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
1/1000	0.267	0.249	0.400	0.433	0.234	0.234	0.245†	0.461*	0.320	0.295	0.372†	0.575*
1/100	0.484	0.326	0.531	0.458	0.459	0.319	0.447†	0.483**	0.586	0.402	0.602*	0.593*
1/10	0.526	0.562	0.632	0.718	0.534	0.572	0.591†	0.725	0.577	0.623	0.690*	0.803*
1.3M	0.743	0.758	0.610	0.742	0.769	0.769	0.734*	0.805**	0.769	0.769	0.756*	0.807**

The polarity value of each word is defined in Takamura's sentiment polarity dictionary, which includes 2349 positive words and 5866 negative words [50].

The results shown in Table 4.6. The * symbol shows a statistically significant ($p<0.01$), the ** symbol shows a statistically significant ($p<0.05$).

This is an important finding, given the degree of the correctness of our EP corpus. As we have shown in Table 4.3, only 57% of samples in our EP corpus are "exactly" correct in terms of sentiment polarity. The figures in Table 4.6 indicate that context-dependent samples are also useful for training a classifier. Table 4.6 also indicates that no significant difference is found between the "Word" and other models, where the corpus size is 1.3M. From these results, we speculate that, as far as the two-class sentiment polarity problem is concerned, word n-gram features might be sufficient if a very large set of labelled data are available.

On the other hand, Table 4.7 indicates that the three-class problem is much harder than the two-class problem. Specifically, positive sentences tend to be classified as neutral. This method has to be improved in future models.

Table 4.8: Examples of TestSet1 (2p, best)

	Annotator A	Annotator B
クリスマスにプレゼントをもらった (I got a Christmas present)	⟨happiness⟩	⟨happiness⟩
友達の家に遊びに行く (I'm going to go to my friend's house)	⟨pleasantness⟩	⟨pleasantness⟩
花見に行ったら突然雨が降り出した (It rained suddenly when I went to see the cherry blossoms)	⟨sadness⟩	⟨sadness⟩
渋滞でほとんど動かない (My car can't move because of the traffic jam)	⟨unpleasantness⟩	⟨anger⟩

4.4.2 Emotion classification

For fine-grained emotion classification, we used the following three test sets:

TestSet1 (2p, best): Two annotators were asked to annotate each positive or negative sentence in TestSet1 with one of the 10 emotion classess. The annotators chose only one emotion class even if they thought several emotions would fit a sentence. Some examples are shown in Table 4.8. The inter-annotator agreement is κ=0.76 in the kappa statistic [11]. For sentences annotated with two different labels (i.e. in the cases where the two annotators disagreed with), both labels were considered correct in the experiments — a model's answer was considered correct if it was identical with either of the two labels.

TestSet1 (1p, acceptable): One of the above two annotators was asked to annotate each positive or negative sentence in TestSet1 with all the emotions involved in it. The number of emotions for a positive sentence was 1.48 on average, and 2.47 for negative sentences. Table 4.9 lists some examples. In the experiments, a model's answer was considered correct if it was identical with one of the labelled classes.

TestSet2: For TestSet2, we used the results of our judgments on the correctness for estimating the quality of the EP corpus described in Section 4.3.2.

In the experiments, the following two models were compared:

Baseline: The baseline model simulates the method proposed by Kozareva [27]. Given an input sentence, their model first estimates the pointwise mutual information

Table 4.9: Examples of TestSet1 (1p, acceptable)

	Annotator A
クリスマスにプレゼントをもらった (I got a Christmas present)	⟨happiness⟩
友達の家に遊びに行く (I'm going to go to my friend's house)	⟨pleasantness⟩, ⟨happiness⟩
花見に行ったら突然雨が降り出した (It rained suddenly when I went to see the cherry blossoms)	⟨anger⟩, ⟨sad⟩, ⟨unpleasantness⟩, ⟨disappointment⟩
渋滞でほとんど動かない (My car can't move because of the traffic jam)	⟨unpleasantness⟩, ⟨anger⟩

(PMI) between each content word cw_j included in the sentence and emotion expression $e \in \{$*anger, disgust, fear, joy, sudness, surprise*$\}$
by $PMI(e, cw) = \log \frac{hits(e,cw)}{hits(e)hits(cw)}$, where $hits(x)$ is a hit count of word(s) x on a Web search engine. The model then calculates the score of each emotion class E_i by summing the PMI scores between each content word cw_j in the input and emotion expression e_i corresponding to that emotion class: $score(E_i) = \sum_j PMI(e_i, cw_j)$. Finally, the model chooses the best scored emotion class as an output. For our experiments, we selected the 349 Japanese emotion words showed in Table 4.1.

For hit counts, we used the Yahoo! search engine.

k-NN: We tested the 1-NN, 3-NN and 10-NN models. In each model, we examined a single-step emotion classification and two-step emotion classification. In the former method, the kNN model retrieves k-most similar examples from the all of the EP corpus. In the latter method, when the sentiment polarity of the input utterance has obtained by the sentiment polarity classifier, the kNN model retrieves similar examples from only the examples whose sentiment polarity are the same as the input utterance in the EP corpus.

The results are shown in Figure 4.5. "Emotion Classification" denotes the single-step models, whereas "Sentiment Polarity + Emotion Classification" denotes the two-step models.

An important observation from Figure 4.5 is that our models remarkably outperformed the baseline. Apparently, an important motivation behind Kozareva et al.

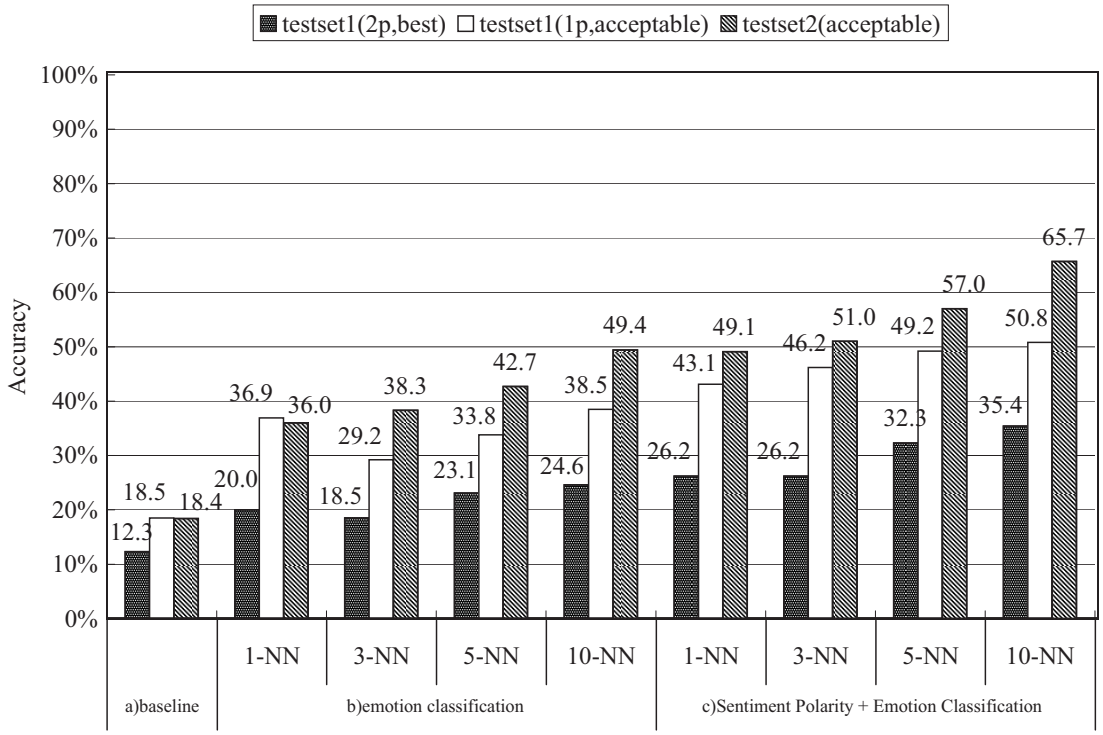


Figure 4.5: Results of emotion classification

(2007)'s method is that it does not require any manual supervion. However, our models, which rely on emotion-provoking event instances, are also totally unsupervised — no supervision is required to collect emotion-provoking event instances. Given this commonality between the two methods, the superiority of our method in accuracy can be considered as a crucial advantage.

Regarding the issue of single-step vs. two-step, Figure 4.5 indicates that the two-step models tended to outperform the single-step models for all the test set. A paired t-test for TestSet2, however, did not reach significance [2]. So we next examined this issue in further detail.

As argued in Section 4.3.3, in the context of human-computer dialog, a misunderstanding of the user's emotion at the level of sentiment polarity would lead to a serious problem, which we call a *fatal error*. On the other hand, misclassifying a case of⟨happiness⟩ as, for example,⟨pleasantness⟩ may well be tolerable. Table 4.10 shows the ratio of fatal errors for each model. For TestSet2, the single-step 10-NN model

[2]The data size of TestSet1 was not sufficient for statistical significance test

Table 4.10: Fatal error rate in emotion classification experiments

	Baseline	Emotion Classification			Sentiment Polarity + Emotion Classification
		1-NN	3-NN	10-NN	
TestSet1	49.2%	29.2%	26.2%	24.6%	15.4%
TestSet2	41.5%	37.6%	32.8%	30.0%	17.0%

made fatal errors in 30% of cases, while the two-step 10-NN model in only 17%. This improvement is statistically significant ($p<0.01$).

4.5 Summary

In this chapter, we have addressed the issue of emotion classification assuming its potential applications to be human-computer dialog system including active-listening dialog. We first automatically collected a huge collection, as many as 1.3M, of emotion-provoking event instances from the Web. We then decomposed the emotion classification task into two sub-steps: sentiment polarity classification and emotion classification. In sentiment polarity classification, we used the EP-corpus as training data. The results of the polarity classification experiment showed that word n-gram features alone are more or less sufficient to classify positive and negative sentences when a very large amount of training data is available. In the emotion classification experiments, on the other hand, we adopted the k-nearest-neighbor (kNN) method. The results of the experiments showed that our method significantly outperformed the baseline method. The results also showed that our two-step emotion classification was effective for fine-grained emotion classification. Specifically, fatal errors were significantly reduced with sentiment polarity classification before fine-grained emotion classification.

For future work, we first need to examine other machine learning methods to see their advantages and disadvantages in our task. We also need an extensive improvement in identifying neutral sentences. Finally, we are planning to apply our model to the active-listening dialog system that our group has been developing and investigate its effects on the user's behavior.

Chapter 5

Conclusion

This thesis took up the problem of the non-task-oriented dialog by two steps: first step is conversation analysis and the next step is its engineering implementation.

In the conversation analysis step, the following two issues were investigated: a) What makes a non-task oriented human-to-human conversation to be an enthusiastic one; b) What is the difference between task-oriented and non-task-oriented dialogs. The first issue was explained by investigating what type of utterances contributes to enthusiasm in a non-task-oriented human-to-human dialog. For this end, we first created a non-task-oriented human-to-human dialog corpus. We then analyzed the relationship between utterances and enthusiasm by studying the instances by studying the instances in the corpus. As a result, we found that “affective utterance” and “cooperative utterance” were related to enthusiasm. On the other hand, concerning the second issue, we investigated what type of utterances appear saliently in non-task-oriented dialog. We first created two types of human-to-human dialog corpora: a task-oriented dialog corpus and a non-task-oriented dialog corpus. We investigated what are the discriminating characteristics that differentiate them. We found that initiation/response utterance appeared more frequently in non-task-oriented dialog. This is because speakers often try to take a lead to continue the exchange smoothly in a non-task-oriented dialog. In addition, we also found that most indirect response and clarification requests work as initiation/response utterance in a non-task-oriented dialog.

Next, in the implementation step, we proposed a method to generate affective utterances that express sympathetic emotion to the partner. We first automatically collected a huge collection of emotion-provoking event instances from the World Wide Web. We classified the emotion-provoking events in terms of the emotion types and their polarity. So, the task in decomposed into the following two sub tasks: sentiment po-

larity (positive and negative) classification and emotion (e.g. happiness, sadness, fear) classification. The results of the experiments showed that our method significantly outperformed the baseline method.

For future work, we are planning to apply our work to the conversational robot that our group has been developing and investigate its effects on the user's behavior.

Bibliography

[1] J. F. Allen, L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light, N. G. Martin, B. W. Miller, M. Poesio, and D. R. Traum. The TRAINS Project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI (JETAI)*, 1994.

[2] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent. Towards Conversational Human-Computer Interaction. *AI Magazine*.

[3] J. Allen and M. Core. Draft of DAMSL: Dialog Act Markup in Several Layers. *ftp://ftp.cs.rochester.edu/pub/packages/dialogue-annotation/manual.ps.gz*, 1997.

[4] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. *Spoken Language Processing*, pp. 2037–2040, 2002.

[5] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. How to find trouble in communication. *Speech Communication*, 40(1-2):117–143, 2004.

[6] M. M. Baum, N. Bergstrom, N. F. Langston, and L. Thoma. Physiological effects of human/companion animal bonding. *Nursing research*, 33(3), 1984.

[7] F. Benamara, V. Moriceau, and P. Saint-Dizier. COOPML: Towards Annotating Cooperative Discourse. *In Proceedings of the ACL Workshop on Discourse Annotation*, pp. 9–16, 2004.

[8] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. *In Proceedints of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99)*, 1999.

[9] F.-R. Chaumartin. A knowledge-based system for headline sentiment tagging. *In Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007.

[10] B. A. Cheikes and B. L. Webber. Elements of a Computational Model of Cooperative Response Generation. *In Proceedings of the ACL Workshop on Speech and Natural Language*, pp. 216–221, 1989.

[11] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, pp. 37–46, 1960.

[12] R. Cohen, C. Allaby, C. Cumbaa, M. Fitzgerald, K. Ho, B. Hui, C. Latulipe, F. Lu, N. Moussa, D. Pooley, A. Qian, and S. Siddiqi. What is Initiative? *User Modeling and User-Adapted Interaction*, 8(3-4):171–214, 1998.

[13] K. Colby. Artificial Paranoia. *Artificial Intelligence*, 2, 1971.

[14] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. Meeting Recorder Project: Dialog Act Labeling Guide. *ICSI Technical Report*, (TR-04-002), 2004.

[15] F. Doshi and N. Roy. The Permutable POMDP: Fast Solutions to POMDPs for Preference Elicitation. *Proceedings of the Seventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, 2008.

[16] R. Epstein, G. Roberts, and G. Beber. The Turing Hub as a Standard for Turing Test Interfaces. *Parsing the Turing Test*.

[17] M. Eskenazi, A. W. Black, A. Raux, and B. Langner. Let's Go Lab: a platform for evaluation of spoken dialog systems with real world users. *In Proceedings of Interspeech*, 2008.

[18] E. Friedmann, A. H. Katcher, J. J. Lynch, and S. A. Thomas. Animal companions and one-year survival of patients after discharge from a coronary care unit. *Public Health Reports*, 95(4), 1980.

[19] D. Graff and S. Bird. Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies. *LREC2000*, 2000.

[20] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997.

[21] R. He, S. Prentice, and N. Roy. Planning in Information Space for a Quadrotor Helicopter in a GPS-denied Environments. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2008)*, 2008.

[22] D. Jurafsky, L. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. *www.dcs.shef.ac.uk/nlp/amities/files/bib/ics-tr-97-02.pdf*, 1997.

[23] D. Kawahara and S. Kurohashi. Case Frame Compilation from the Web using High-Performance Computing. *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006.

[24] N. Kobayashi, R. Iida, K. Inui, and Y. Matsumoto. Collecting Evaluative Expressions for Opinion Extraction. *Lecture Notes in Artificial Intelligence*, 3248, 2005.

[25] K. Komatani, T. Kawahara, R. Ito, and H. Okuno. Efficient Dialogue Strategy to Find Users' Intended Items from Information Query Results. *In Proceedings of the COLING*, 2002.

[26] K. Komatani, S. Ueno, T. Kawahara, and H. Okuno. User Modeling in Spoken Dialogue Systems for Flexible Fuidance Generation. *In Proceedings of the EUROSPEECH*, 2003.

[27] Z. Kozareva, B. Navarro, S. Vazquez, and A. Nibtoyo. UA-ZBSA: A Headline Emotion Classification through Web Information. *In Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007.

[28] T. Kudo and Y. Matsumoto. A Boosting Algorithm for Classification of Semi-Structured Text. *In Proceedings of the EMNLP*, 2004.

[29] D. Lago, M. Delaney, M. Miller, and C. Grill. Companion animals, attitudes toward pets, and health outcomes among the elderly: A long-term follow-up. *Anthrozoos*, 3(1), 1989.

[30] E. Levin, S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. D. Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker. The AT&T DARPA communicator mixed-initiative spoken dialog system. *ICSLP*, 2:122–125, 2000.

[31] D. J. Litman and K. Forbes-Riley. Predicting Student Emotions in Computer-Human Tutoring Dialogues. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.

[32] D. Litman, S. Singh, M. Kearns, and M. Walker. NJFun: A Reinforcement Learning Spoken Dialogue System. *In Proceedings of the ANLP/NAACL*, 2000.

[33] B. Liu. Web Data Mining. *Springer*, pp. 411–440, 2006.

[34] L. S. Lopes, A. J. S. Teixeira, M. Quindere, and M. Rodrigues. From Robust Spoken Language Understanding to Knowledge Acquisition and Management. *EUROSPEECH*, pp. 3469–3472, 2005.

[35] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281, 1988.

[36] S. M. A. Masum, H. Prendinger, and M. Ishizuka. Emotion Sensitive News Agent: An Approach Towards User Centric Emotion Sensing from the News. *In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2007.

[37] Y. Matsusaka, T. Tojo, and T. Kobayashi. Conversation robot participating in group conversation. *IEICE Trans. on Information and System*, pp. 26–36, 2003.

[38] C. Muller and M. Strube. Multi-Level Annotation in MMAX. *In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, 2003.

[39] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 76–86, 2002.

[40] M. Pantic and L. J. M. Rothkrantz. Facial Action Recognition for Facial Expression Analysis From Static Face Images. *IEEE Transactions on SMC-B*, 34(3):1449–1461, 2004.

[41] A. Raux, D. Bohus, B. Langner, A. W. Black, and M. Eskenazi. Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience. *In Proceedings of Interspeech*, 2006.

[42] A. Raux, B. Langner, A. W. Black, and M. Eskenazi. LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives. *In Proceedings of Eurospeech*, 2003.

[43] V. Rieser and J. Moore. Implications for Generating Clarification Requests in Task-Oriented Dialogues. *In Proceedings of the ACL*, pp. 239–246, 2005.

[44] M. Rotaru, D. J. Litman, and K. Forbes-Riley. Interactions between Speech Recognition Problems and User Emotions. *Proceedings 9th European Conference on Speech Communication and Technology*, 2005.

[45] A. I. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, and A. Oh. Creating Natural Dialogs in the Carnegie Mellon Communicator System. *EUROSPEECH*, pp. 1531–1534, 1999.

[46] J. R. Searle. Speech Acts. *Cambridge University Press)*, 1969.

[47] T. Shibata, T. Mitsui, K. Wada, and K. Tanie. Subjective Evaluation of Seal Robot: Paro - Tabulation and Analysis of Questionnaire Results -. *Journal of Robotics and Mechatronics*, 14(1):13–19, 2002.

[48] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *HLT-NAACL SIGDIAL Workshop*, 2004.

[49] A. Stent and J. Allen. Annotating Argumentation Acts in Spoken Dialog. *Technical Report 740*, 2000.

[50] H. Takamura, T. Inui, and M. Okumura. Extracting Semantic Orientations of Words using Spin Model. *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)*, pp. 133–140, 2005.

[51] I. Tashev, M. L. Seltzer, Y.-C. Ju, D. Yu, and A. Acero. Commute UX: Telephone Dialog System for Location-based Services. *In Proceedings of the 8nd SIGdial Workshop on Discourse and Dialogue*, 2007.

[52] H. Teramura. Japanese Syntax and Meaning. *Kurosio Publishers (in Japanese)*, 1982.

[53] S.-C. TSENG. Toward a Large Spontaneous Mandarin Dialogue Corpus. *In Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, 2001.

[54] P. Turney. Thumbs up? thumbs down? semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, 2002.

[55] D. Vanderveken. Meaning and speech acts. *Cambridge University Press)*, 1990.

[56] V. N. Vapnik. The Nature of Statistical Learning Theory. *Springer*, 1995.

[57] K. Wada, T. Shibata, T. Saito, and K. Tanie. Effects of Robot Assisted Activity for Elderly People and Nurses at a Day Service Center. *In Proceedings of the IEEE*, 92(11):1780–1788, 2004.

[58] M. A. Walker, J. C. Fromer, and S. Narayanan. Learning Optimal Dialogue Strategies: A Case Study of a Spoken Dialogue Agent for Email. *In Proceedings of COLING/ACL*, 1998.

[59] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. I. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection. *EUROSPEECH*, pp. 1371–1374, 2001.

[60] M. Walker, L. Hirschman, and J. Aberdeen. Evaluation for DARPA Communicator Spoken Dialogue Systems. *Language Resources and Evaluation Conference(LREC)*, 2000.

[61] R. Wallace. A.l.i.c.e. articial intelligence foundation. *http://www.alicebot.org*, 2005.

[62] J. Weizenbaum. ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine. *Communications of the ACM*, 9(1), 1966.

[63] B. Wrede and E. Shriberg. Spotting ”Hot Spots” in Meetings: Human Judgements and Prosodic Cues. *Eurospeech-03*, pp. 2805–2808, 2003.

[64] B. Wrede and E. Shriberg. The Relationship between Dialogue Acts and Hot Spots in Meetings. *IEEE ASRU Workshop*, 2003.

[65] A. Yokoyama. The possibility of the psychiatric treatment with a robot as an intervention - From the viewpoint of animal therapy. *International Conference Soft Computing and Intelligent Systems*, 2002.

[66] 荒木, 伊藤, 熊谷, 石崎. 発話単位タグ標準化案の作成. 人工知能学会誌, 14(2), 1999.

[67] 石崎, 伝. 談話と対話. 東京大学出版会, 2001.

[68] 金森, 鈴木, 田中. 症例報告 ペット型ロボットによる高齢者の Quality of Life 維持・向上の試み. 日本老年医学会雑誌, 39(2):214–218, 2002.

[69] 山田, 溝口, 原田. 質問応答システムにおけるユーザ発話モデルと協調的応答の生成. 情報処理学会論文誌, 35(11):2265–2275, 1994.

[70] 神田, 石黒, 小野, 今井, 前田, 中津. 研究用プラットフォームとしての日常活動型ロボット“ Robovie ”の開発. 電子情報通信学会論文誌, 85(4):380–389, 2002.

List of Publications

Journal Papers

1. Ryoko Tokuhisa, Ryuta Terashima : “The Relationship between Utterances and “Involvement” in Conversational Dialogue”, *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 21, No. 2, pp. 133-142, 2006. (in Japanese).
2. Ryoko Tokuhisa, Ryuta Terashima : “An Analysis of ‘Distinctive’ Utterances in Non-task-oriented Conversational Dialogue”, *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 22, No. 4, pp. 425-435, 2007. (in Japanese).
3. Ryoko Tokuhisa, Kentaro Inui, Yuji Matsumoto : “Emotion Classification using Massive Examples Extracted from the Web”, *IPSJ Journal*, Vol. 50, No. 4, pp. 1365-1374, 2009. (in Japanese).

International Conference/Workshop Papers

1. Ryoko Tokuhisa, Ryuta Terashima: “Relationship between Utterance and ”Enthusiasm” in Non-Task-Orieinted Conversational Dialogue”, *In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pp. 161-168, 2006.
2. Ryoko Tokuhisa, Kentaro Inui, Yuji Matsumoto: “Emotion Classification Using Massive Examples Extracted from the Web”, *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 881-888, 2008.

Other Publications

1. Ryoko Tokuhisa, Ryuta Terashima : “The Relationship between Utterances and “Involvement” in Conversational Dialogue”, *Transactions of the Japanese Society for Artificial Intelligence*, 2005. (in Japanese).

2. Ryoko Tokuhisa, Ryuta Terashima: “An Analysis of Distinctive Utterances in Non-task-oriented Conversational Dialogue”, *In Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing*, 2006. (in Japanese).

3. Ryoko Tokuhisa: “A Survey of a Non-task-oriented Conversational Dialogue ∼ Toward A Humanlike Dialogue System ∼”, *Transactions of the Japanese Society for Artificial Intelligence*, 2006. (in Japanese).

4. Ryoko Tokuhisa, Kazuya Shitaoka, Ryuta Terashima: “An Analysis of the Rhetorical Relation in Conversational Dialogue”, *In Proceedings of the 13th Annual Meeting of the Association for Natural Language Processing*, 2007. (in Japanese).

5. Ryoko Tokuhisa, Kentaro Inui: “Emotion Classification Using Massive Examples Extracted from the Web”, *In Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, 2008. (in Japanese).