Syracuse University

SURFACE

School of Information Studies - Dissertations

School of Information Studies (iSchool)

5-2013

An Investigation of Digital Reference Interviews: A Dialogue Act Approach

Keisuke Inoue
Syracuse University

Follow this and additional works at: https://surface.syr.edu/it_etd

Part of the Library and Information Science Commons

Recommended Citation
Inoue, Keisuke, "An Investigation of Digital Reference Interviews: A Dialogue Act Approach" (2013). *School of Information Studies - Dissertations*. 81.
https://surface.syr.edu/it_etd/81

This Dissertation is brought to you for free and open access by the School of Information Studies (iSchool) at SURFACE. It has been accepted for inclusion in School of Information Studies - Dissertations by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.

Abstract

The rapid increase of computer-mediated communications (CMCs) in various forms such as micro-blogging (e.g. Twitter), online chatting (e.g. digital reference) and community-based question-answering services (e.g. Yahoo! Answers) characterizes a recent trend in web technologies, often referred to as the *social web*. This trend highlights the importance of supporting linguistic interactions in people's online information-seeking activities in daily life – something that the web search engines still lack because of the complexity of this human behavior. The presented research consists of an investigation of the information-seeking behavior of digital reference services through analysis of discourse semantics, called dialogue acts, and experimentation of automatic identification of dialogue acts using machine-learning techniques. The data was an online chat reference transaction archive, provided by the Online Computing Library Center (OCLC). Findings of the discourse analysis include supporting evidence of some of the existing theories of the information-seeking behavior. They also suggest a new way of analyzing the progress of information-seeking interactions using dialogue act analysis. The machine learning experimentation produced promising results and demonstrated the possibility of practical applications of the DA analysis for further research across disciplines.

An Investigation of Digital Reference Interviews: A Dialogue Act Approach

by

Keisuke Inoue

B.A. Law, Waseda University, 1996

M.S. Computer Science, Syracuse University, 2002

M.A. Linguistics, Syracuse University, 2005

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information Science and Technology
in the Graduate School of Syracuse University

May 2013

Copyright © 2013 Keisuke Inoue

All Rights Reserved

Acknowledgements

I would like to thank my advisor, Liz Liddy, for always supporting her worst advisee through all the troubles, whether it is a last minute conference paper submission, an unexpectedly long bike ride, or an absurd April Fool's joke. Without her patience and generosity, I would have not finished my work.

I would also like to thank my committee members, Bob Oddy, Nancy McCracken, David Lankes, Bei Yu, and Howard Turtle, for their feedback, ideas, and support. Having discussions with them was the most luxurious part of my PhD student life.

Thanks to Dr. Lynn Silipigni Connaway and Dr. Marie L. Radford for generously sharing the data. And thanks to Jeremy Browning at OCLC Research for providing the data.

Thanks to Bridget Crary for always taking care of the PhD students whenever we are in trouble (and we are in trouble all the time).

Thanks to Eileen Allen for the editorial help.

Thanks to my wife, Sarah Inoue, for all sorts of help and for her patience.

Thanks to Beta Phi Mu and Dr. Eugene Garfield for assistance in the form of a Eugene Garfield Doctoral Dissertation Fellowship.

Thanks to ALISE and OCLC for assistance in the form of the ALISE/OCLC Library & Information Science Research Grant.

And lastly, thanks to my mother, Tomoko Inoue, for all the encouragement (or complaints) and support for all the long years.

This dissertation is dedicated to my father, Masakatsu Inoue.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context

While fully automated web search engines have become the primary tool for information searching by the majority of web users, human-powered information services, such as community-based question-answering services (e.g. Yahoo! Answer, Quora, WikiAnswers), online chat services with information experts (e.g. help desk and digital reference services), or question-answering services using short messaging service (SMS) on mobile phones (e.g. ChaCha, kgb answers) have also been increasing in popularity. By facilitating computer-mediated communication (CMC), these systems provide opportunities for information-seeking communication using natural language, enabling interactions that are more complex, personalized, and richer in content than simple exchanges of keywords and hyperlinks using algorithmic mechanisms as is the case of the traditional “one-shot” web search engines. Such systems often employ specially-tailored algorithms that aggregate or process the input and feedback from masses of people, enabling their users to utilize the information in turn. This also provides social opportunities, facilitating interactions and further encouraging information sharing among the users. The facilitation of CMC on the web is not only in the context of information-seeking or searching. More and more people are utilizing and contributing to the information resources or digital media archives that are created by masses of people (e.g. Wikipedia, Youtube, Flickr). These trends in web-based information systems, called *social web systems* in this document, raise questions regarding information behaviors in this new digital environment. At first glance, the trends seem to suggest a new type of information behavior. As the term *prosumer* (Toffler, 1984) suggests, the distinction between information producers and information consumers has become vague. Creation, accumulation, and dissemination of information artifacts has become possible with tremendous speed and volume, and in various forms, media, and modes. But how exactly do people seek information in this new environment? And how can we improve the systems for even better information experiences? With the rapidly increasing volume of information creation by the human society – IBM (2011) reported that 2.5 quintillion bytes of data are created every day – the answers to these questions are becoming critical.

In order to address the above questions, this study was designed with two major goals: 1) to enhance understanding of the information-seeking behavior in the current context, and 2) to apply the understanding to experiments for new information technologies. The research consisted of an investigation of the information-seeking behavior of digital reference services through analysis of discourse semantics, called *dialogue acts* (defined in Section 2.3.2.2), and experimentation of automatic identification of dialgoue acts using machine learning techniques.

1.2 Research Questions

The following are research questions of this study:

RQ 1: What is the discourse of question negotiation in digital reference?

1.1) What are the components of the question negotiation process in digital reference and how are they distributed in the process?

1.2) What are the structural characteristics of the process?

RQ 2: Can the discourse-level semantics of question negotiation be automatically detected? If so, how?

2.1) Which machine learning algorithms are suited for detecting the discourse-level semantics of question negotiation in digital reference?

2.2) What types of linguistic evidence are useful for automatic recognition of the discourse-level semantics of question negotiation in digital reference?

1.3 Methods

This study was designed with two major stages – 1) the discourse analysis stage, and 2) the machine learning stage – to answer **RQ 1** and **RQ 2**, respectively. **RQ 1.1** was answered through the development of a new annotation scheme and by analyzing the distribution of the annotations. In order to answer **RQ 1.2**, relations among the components were analyzed in terms of structural characteristics, such as transitional dependency and their relative positions. **RQ 2.1** and **RQ 2.2** were answered in the machine-learning stage, which consisted of semi-factorial experiments of different algorithms and features, following a common machine learning experiment procedure. Chapter 3 describes the methods in more detail. This section provides a brief overview.

1.3.1 Discourse Analysis

The discourse analysis in this study utilized a dialogue act annotation scheme, which had been developed through an exploratory study by the researcher (Inoue, 2009). The coding scheme focused on identifying the following four aspects of the interactions: 1) exchanging information, 2) assigning tasks, 3) maintaining and managing the dialogue, and 4) maintaining the social relationship. Building a coding scheme with these four aspects was motivated by the Dynamic Interpretation Theory (Bunt, 1994), which hypothesized that dialogues were always carried out by participants performing the following two kinds of tasks: 1) tasks to achieve the goal that motivated the dialogue and 2) tasks to maintain the dialogue itself in order to achieve goals that were associated to the context of the dialogue. The notion of these two kinds of tasks fit in with observations from previous studies about reference research which showed how the communicative task was carried out during chat reference (e.g. Radford (1993; 2006a; 2006b)), and from studies about information-seeking behaviors which examine how information exchanges occur during an information-seeking process (e.g. Belkin et al. (1983), Spink et al. (1995), Saracevic et al. (1997), and Wu (2005)). The coding scheme for the proposed study was developed based on DIT++, proposed by Bunt (2007), and was extended to accommodate the goals of the study.

The analysis of the annotated data was done by examining the data from overall distributions of annotations, distributions over the progress of interviews, transitions, and linguistic forms. The observations were then compared with the "classic" theories and models of information-seeking behaviors such as the ASK Hypothesis (Belkin et al., 1982) and the Berrypicking model (Bates, 1989), as well as more recent developments such as micro-level information seeking processes (Wu, 2005) and exploratory search (Marchionini, 2006).

1.3.2 Machine-learning Experiments

While various algorithms have been proposed for automatically detecting dialogue acts, these algorithms can be categorized into two approaches: sequential labeling and text classification. The sequential labeling approach is used widely in speech and dialogue research, where many early automatic dialogue act annotation experiments were conducted. Text classification algorithms are a more recent approach applied to dialogue act annotations. Hidden Markov Models (HMM) is the most successful algorithm in the earlier approach (Kita et al., 1996; Reithinger et al., 1996; Stolcke et al., 2000), while Support Vector Machines (SVM) is the most successful algorithm in the later approach (Cohen et al., 2004; Carvalho and Cohen, 2006; Hu et al., 2009). In this study, the performance of HM-SVM, a learning algorithm that combines HMM and SVM, was examined. The experiment also examined the effects of different linguistic attributes (features) to the performance of the machine learning task.

Based on a literature review and analysis of the data, the following features were used in the experiments: 1) word vector, 2) message sequence number, 3) speaker, 4) text segment length, 5) message position, and 6) word bigram vector. Using the standard SVM as the baseline, the experiment used two-way semi-factorial design, examining isolated effects and interactive effects of the two algorithms and the six features listed above.

1.3.3 Data

The data, provided by the Online Computer Library Center (OCLC), is a log of digital reference service dialogues, collected between July 2004 and December 2006.[1] Out of 800 interview sessions in the original data, 211 interviews were selected based on the types of questions asked in the reference interviews. This selection was to ensure that the information problems presented in the interviews required question negotiations. Each interview used for the analysis consists of an average of 26 messages sent between a librarian and a user.

1.4 Significance of the Study

Over 40 years ago, Taylor (1968) investigated challenges of information-seeking interactions as a science of human behavior in his seminal work "Question-negotiation and Information Seeking in Libraries". He used the term "question negotiation" to describe how librarians needed to "negotiate" between the ill-defined information need of the user and formal requirements of input to information systems. He suggested investigating mechanisms that enable efficient information-seeking communication should be part of the primary agenda of the information science community. Since then, the need to incorporate an interactive process into information retrieval systems has been repeatedly claimed by various researchers (Belkin et al., 1982; Bates, 1989; Belkin, 1993; Ingwersen, 1996; Spink and Saracevic, 1998; Marchionini, 2006). An abundance of models of information-seeking interactions have been

[1] The data was originally prepared for an on-going research project by Radford and Connaway (2005) and became available to the researcher by courtesy of Dr. Radford, Dr. Connaway, and the OCLC.

proposed and some experimental systems with interactive interfaces (Oddy, 1977; Croft and Thompson, 1987) have been developed. Prolonged efforts have been made to experiment with the implementation and evaluation of interactive processes for information retrieval. A most notable effort to test such systems was made by the Text Retrieval Conference by conducting the series of investigations with the Interactive track (1997, 1998, 1999, and 2000) and the HARD track (2003, 2004, and 2005). Yet, the integration of interactive process into the daily-life IR systems has not been realized.

While there may be various reasons why these efforts haven't borne fruit, one possible explanation is the discordance between the advances in information technologies and development of theories of information behavior. Wu and Liu (2003) cites the following two reasons:

1. Difference in the levels of contexts

   Many studies of information-seeking behavior in the information science field look at the social context, rather at system interactions and thus lack of understanding of changes in the nature of information-seeking interactions along the advancement of information technologies.

2. Lack of specification of technological details

   Wu and Liu (2003) states "the descriptive models of information behaviors may not be detailed enough for interactive searching in IR systems because IR interaction involves both the individual task level (e.g., task complexity) and the IR operational task level (e.g., query formulation)."

This study was designed to pay attention to the two concerns above. A combination of linguistic analysis and machine learning experimentation was chosen for the investigation in order to contribute to both the conceptual understanding of information-seeking interactions, and, using this understanding, to the development of information technologies.

The first stage of this study, a discourse analysis of digital reference transactions, contributes to the body of science with the following aspects: 1) a new discourse analysis scheme that describes components of the information-seeking processes in digital reference,

2) refinement of existing information-seeking behavior models, by confirming or refuting the models with empirical results, and 3) enhanced understanding of the phenomena, by providing a new way to observe and analyze information-seeking interactions. Some researchers have suggested that the development of web information technologies have changed people's attitude towards information seeking. For example, according to Radford and Connaway (2007b), people who grew up in the digital communication environment "want easy access to full-text documents and become impatient with complex searching" (page 5). The analysis revealed how librarians are responding to such users' needs, with important implications for the design of information services, such as the traditional library reference service or the emerging social web systems discussed above, as well as training of the information professionals at such services.

The outcomes of the discourse analysis are also useful for designing new IR systems that incorporate social web systems in at least the following two ways: 1) incorporating an interactive interface in the IR process – by understanding how the information requirement is specified in the process, e.g. how information problems are specified initially and how clarifying questions should take place (See Section 4.1.5 for the analysis.), the systems will be able to interact with the user in an efficient manner; and 2) utilizing social web media as information resources – by understanding how the information is provided, e.g. which utterance in a conversation or which part of an utterance is relevant to the information problem and which is not, the systems will be able to better utilize the social web media as an information resource.

The second part of the study, a machine learning experiment, provides a proof of concept for the dialogue model, by confirming that there is linguistic evidence that represents the discourse semantics (dialogue acts) that the linguistic analysis in this study attempted to capture, and that the semantics can be learned by following certain procedures (algorithms). The experiments, which consisted of semi-factorial combinations of different algorithms and features, produced promising results, indicating a good potential for practical applications of the dialogue act analysis for further research and development across disciplines. Specifically, the outcome of the experiments contributes to further research by providing the following: 1) a new aspect for evaluating digital reference services, 2) new data attributes for information extraction / retrieval algorithms (document models), and 3) a prototypical dialogue model for constructing fully-automated dialogue systems.

1.5 Summary

In this study, the researcher analyzed the information-seeking interactions between librarians and library service users through online reference services. By taking an interdisciplinary approach, the analysis provides a holistic view of the phenomenon and contributes to the existing theories and models of information-seeking behavior. The research also included machine learning experiments to find the optimal combination of algorithms and attributes for learning linguistic attributes, dialogue acts, of interactions. Thus the intellectual merit of the study includes the theoretical development of information-seeking behavior as well as the technological development of information systems, both of which lead to more efficient and satisfying information experiences for users of reference services and information systems in general.

Chapter 2

Related Studies

2.1 Introduction

This chapter discusses previous studies that informed the design of this research, e.g. the analytical method, experimentation, data collection, and theoretical justification. These studies are from the fields of information-seeking behavior, information retrieval, dialogue analysis, reference, digital reference, text classification, discourse modeling, speech acts, and dialogue acts, which spread cross at least three major academic disciplines: library and information science, linguistics, and computer science. Figure 2.1 illustrates how each field is located among the three disciplines, based on the literature review presented in this chapter. Relationships among the disciplines that influence the field of library and information science are discussed elsewhere, e.g. Ingwersen (1992). The diverse collection of work presented in this chapter focused on three areas of studies, each of which is more or less a subfield of each of the three disciplines:

- Studies of information-seeking interactions (Section 2.2),
- Studies of discourse semantics (Section 2.3), and
- Studies of machine learning dialogue act annotation (Section 2.4).

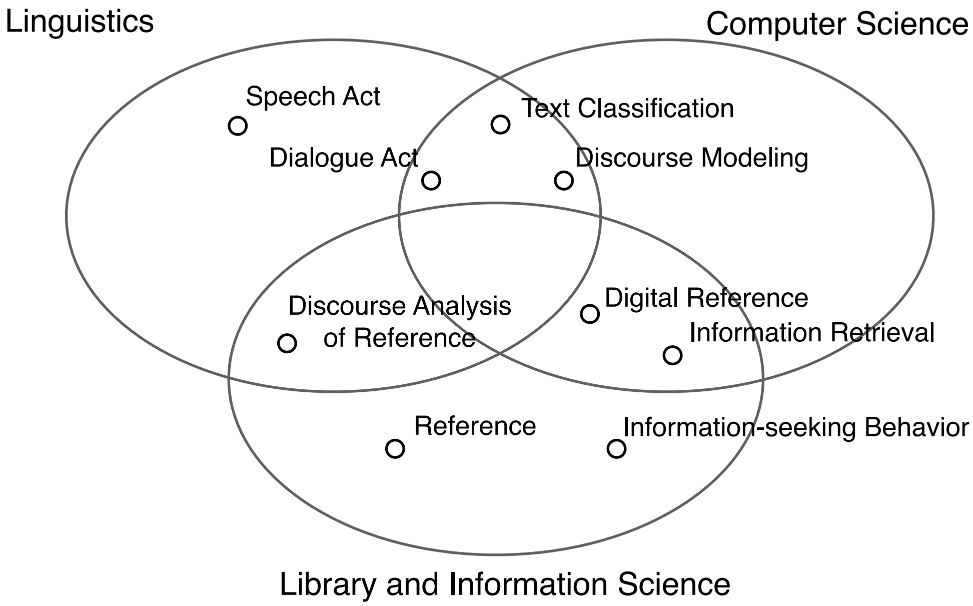


Figure 2.1: Related fields of studies across the three disciplines

Previous studies of information-seeking interactions informed the formation of the research questions. While information-seeking behavior traditionally is a topic in library and information science, previous studies employed methods and theories from different fields, e.g. cognitive science, psychology, linguistics, and communication. Section 2.2 reviews how the field of information-seeking behavior science has evolved in the past several decades, in terms of conceptual and methodological development.

The studies of discourse semantics, a subfield of linguistics, informed the conceptual framework and the analytical methods of the discourse analysis stage of the study. Linguistics offers a range of methods and theories for analyzing texts that may be utilized for research and development of information technologies. Section 2.3 reviews such theories and methods designed to explain or analyze the meaning of texts in dialogue.

The studies of machine learning dialogue act annotation informed the the experimental design used in this study. Machine learning is a vast field, which started as an outgrowth of computer science and mathematics (statistics), and has grown rapidly in the past decades (Mitchell, 2006). While machine learning is applied to a wide variety of human behavior or natural phenomena, Section 2.4 focuses on applications of the approach for recognizing discourse semantics of dialogue and for determining the right methods for the experiments.

2.2 Studies of Information-seeking Interactions

Taylor (1968) investigated the challenges of information-seeking interactions as a science of human behavior in his seminal work "Question-negotiation and Information Seeking in Libraries". He coined the term "question negotiation" to describe how librarians needed to "negotiate" between the ill-defined information need of the user and formal requirements needed as input to information systems. Information scientists have investigated various ways to improve information-seeking experiences since then, expanding the disciplinary boundaries by incorporating linguistic analysis to understand users' needs or the information in documents, developing evaluation measures for information systems, experimenting with different computational processes (algorithms) to predict the information objects that users are looking for, exploring the digital environment, and so on. In the following, the studies of information-seeking interactions are described in two subsections. Section 2.2.1 reviews the conceptual development of information-seeking interactions in terms of the evolution of *models* of information-seeking behavior. Section 2.2.2 reviews studies that examined information-seeking behavior in reference interactions. The studies in Section 2.2.2 report concrete observations about the interactions, rather than abstract models that the studies in Section 2.2.1 propose.

The studies from both sections collectively led to the formation of this study's research questions, by explicating assumptions that have been made and proposing hypotheses that needed to be tested.

2.2.1 Models of Information-seeking Behavior

Information-seeking interaction is one of the most well-studied aspects of information-seeking behavior. Various researchers (Belkin et al., 1982; Bates, 1989; Belkin, 1993; Ingwersen, 1996; Spink and Saracevic, 1998; Marchionini, 2006) repeatedly claimed the need to incorporate interactive processes in information retrieval systems in order to improve users' information experience. Theoretical developments along such arguments are best represented by the evolution of conceptual *models* of information-seeking behavior proposed by those researchers. A *model* of information-seeking behavior provides constructs and relationships among them in order to explain and/or predict human information-seeking behavior with different levels or aspects: some models describe the human behaviors that are related to information need, seeking, and use as a whole (Wilson, 1999), while others describe a particular aspect of the information seeking process e.g. the anomalous states of knowledge (Belkin et al., 1982), berrypicking, (Bates, 1989), and the information search process (Kuhlthau, 1991). Regardless of the aspects or levels, models tend to have the following properties (Saracevic, 1996).[1]

- *Abstract:* The model describes the researcher's perception of an aspect of the information-seeking process in an abstract manner.
- *Operationalizable:* The model is operationalizable as a working system theoretically and practically.
- *Verifiable:* The model and its operationalized system can be tested and evaluated.

Furthermore, an ideal model may have the following properties:

- *Comprehensive:* The model explains all of the aspects of the information-seeking process.
- *Simple:* The model is as simple as possible without losing any necessary details.

Models described in this section were developed as an antithesis to the traditional conceptualization of IR processes, which was based on the Cranfield experiments (Cleverdon, 1967). In this tradition, information-seeking processes proceed as follows (also shown in Figure 2.2):

1. An IR system is developed with a collection of information artifacts and batch processes to create the representation of information contained in the artifacts (index) prior to access by users.

[1] Similar descriptions can be found in Wilson (1999) or Case (2002).

2. A user formulates and provides a query to the system, based on his/her information need.

3. The system matches the query and the representations of information artifacts based on some algorithm and returns a list of artifacts to the user as a response to the query.

4. The performance of a system is measured based on the *relevance* between the query and the information artifacts that the system returned.

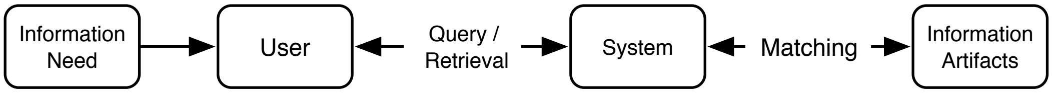


Figure 2.2: Traditional IR process (from Bates (1989))

There are three major assumptions in this view of IR processes:

- Users have an information need that is defined and fixed at the beginning of an information-seeking process;
- Users are capable of representing their information need with a single input (query) to the system; and
- The *relevance* of an information artifact is defined based on the query.

In the following paragraphs, models of information-seeking behavior that focused on the nature of information-seeking interactions are described. All the models below reject one or more of these assumptions and suggest more involvement from users and systems for a better information experience.

The ASK Hypothesis

The focus of the ASK hypothesis (Belkin et al., 1982) is on the very early stage of human information seeking when people recognize an information need and initiate an information-seeking process. According to this hypothesis, people are not able to specify their information need precisely at this stage of information seeking, called *anomalous states of knowledge* (ASK). Thus, Belkin et al. claim that information retrieval systems need to employ a strategy that elicits users' information needs interactively. Some early studies in information-seeking behavior (Brooks and Belkin, 1983; Daniels et al., 1985; Brooks et al., 1986; Belkin et al., 1987a) confirmed this hypothesis by analyzing the conversation structure of various reference interviews (See Section 2.2.2.1 for more about the studies). Based on this hypothesis, Oddy (1977) implemented THOMAS, one of the earliest IR systems with an interactive component.

Berrypicking Model

Like the ASK hypothesis, the *berrypicking* Model (Bates, 1989) also questions the classic system-based information-seeking behavior models but with a different approach. In contrast to ASK, the focus of this model is on the evolving nature of the information requirement and information seeking strategies throughout a process. Specifically, Bates characterizes information-seeking behavior as follows:

- Users' information needs are not static, but evolve throughout IR processes.
- A single IR process may be satisfied not by a single piece of information but by a series of information pieces, which the user obtains bit by bit.
- Users may employ a wide variety of IR strategies, other than a "single-shot" query.
- Users may utilize a wide variety of information resources.

Figure 2.3 illustrates the idea. At each stage of the search, the user's conception of their query changes based on their thoughts and perceptions of retrieved information from the search. The query is not satisfied by the last piece of information retrieved, but by the series of information pieces retrieved throughout the process. Although some of Bates' proposals and implications for system design have become reality, it is still a challenge for today's IR research to satisfy the notion of an evolving search.

Kuhlthau's Information Search Process Model

Kuhlthau (1991) defines the information search process (ISP) as "the user's constructive activity of finding meaning from information in order to extend his or her state of knowledge

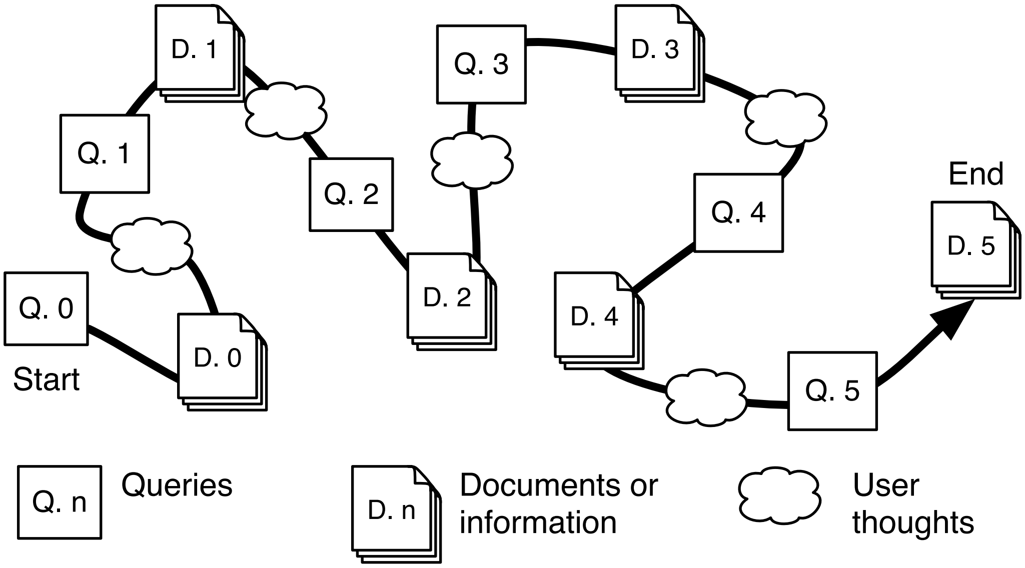


Figure 2.3: Berrypicking Model (from Bates (1989))

ISP Stages	**Feelings**	**Thoughts**	**Actions**
Initiation	Uncertainty	General / Vague	Seeking background information
Selection	Optimism		
Exploration	Confusion / Frustration / Doubt		Seeking relevant information
Formulation	Clarity	Narrowed / Clearer	
Collection	Sense of direction / Confidence	Increased interest	Seeking relevant or focused information
Presentation	Relief / Satisfaction or Disappointment	Clearer or Focused	

Table 2.1: ISP Stages (from Kuhlthau (1991))

on a particular problem or topic" (p. 361). By conducting a series of empirical studies based on the personal constructs theory (Kelly, 1963), she discerned six stages of the ISP, each of which is characterized by three aspects: affective aspects (*feelings*), cognitive aspects (*thoughts*), and physical aspects (*actions*). The unique contribution of this study is that it sheds light on the affective aspect of the ISP, which had not been studied previously, and establishes a relationship between *uncertainty* or *anxiety* and the progression in the ISP, based on empirical research (Table 2.1 summarizes the common characteristics of the three aspects in each stage of ISP).

		Mode of Definition	
		Well-Defined	**Ill-Defined**
Intrinsic Stability	**Stable**	Rich, stable cognitive state System-based IR	Weak, stable cognitive state The ASK hypothesis
	Variable	Rich, variable cognitive state Berrypicking	Weak, variable cognitive state Berrypicking, The ASK hypothesis

Table 2.2: Matrix of information needs and other models

Cognitive IR Theory

The cognitive IR theory of Ingwersen (1996) attempts to amalgamate the models of the previous IR research into a single theory that analyzes information-seeking behavior with a cognitive view point. For example, in this theory, the forms of information need found in the ASK hypotheses, the berrypicking model, and the system-based IR model can be placed in a single matrix in which each row represents a different stability of cognitive structure of the information seeker (*intrinsic stability*) and each column represents a different level of definition of the information need in the cognitive space of the information searcher (Table 2.2). His theory not only incorporates results from various earlier studies, but also reveals a number of complex interactions involved in information-seeking processes, by providing a model of IR interactions. Saracevic (1996), however, points out that the model is so complex that no studies have yet been conducted to test his model empirically nor to apply his model in an implementation.

Wilson's Information Behavior Models

Wilson developed two models of information behavior from 1981 to 1996. In his first model, attention was paid to covering various aspects of the information seeking process (Wilson, 1999). The model suggested that information-seeking behavior is a consequence of the rise of an information need, which, in turn, is a consequence of information use, which in turn, is a consequence of the previous information-seeking behavior. The model is similar to what Meadow (2007) proposed, in that it demonstrates the recursive nature of human information behavior. The problem with this model is, as Wilson himself points out, that it merely provides a conceptual map of information behavior and does not generate any testable hypotheses.

The second model Wilson developed is based on the first model, but incorporates theories from various studies to explain the choices made through the information-seeking process: stress/coping theory from psychology to explain why some of the information needs require prompt information seeking, risk/reward theory from consumer research to explain why some information sources are preferred, and social learning theory from psychology to explain why some users pursue their goal successfully while others do not. An important aspect of this model is that by using theories from disciplines outside information science, the model includes psychological or social variables that affect information seeking behavior. Thus by incorporating this model, system researchers may design a system based on predictions that these constructs and psychological / sociological theories provide. The problem of this model is that, although the model attempts to explain both the internal, psychological aspects and the external, social aspects of information behavior, it still lacks a good explanation of how socio-emotional or interpersonal context affects information needs. For example, the model does not explain how social behavior, such as communication, is related to how information-seeking processes are carried out.

Saracevic's Stratified Interaction Model

Saracevic's stratified interaction model is an application of stratified linguistics to the human-computer interaction in IR. The concept is parallel to the levels of linguistic analysis applied to NLP, such as the one proposed by Liddy (1998) (described in Section 2.3). The basic premise of the model is that IR interactions take place at multiple levels, shown in Figure 2.4, and thus the system needs to facilitate different interfaces and processes for each level in order to accommodate successful IR sessions.

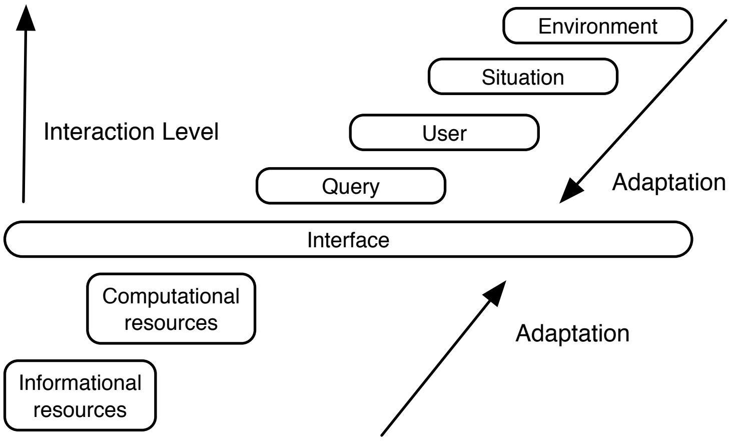


Figure 2.4: Saracevic's model of IR interaction

Exploratory Search

Marchionini (2006) claims that in order to meet the expectations of generations who grew up with web technologies, IR systems must allow users to "lookup", "learn", and "investigate" fluidly. He argues that as people spend more time online, they increase their expectations about their experience with information technologies and content, and seek richer information experiences, such as data mining and in-depth analysis of the information presented. Similar observations were also made in a focus group study to investigate the use of digital reference services by teenagers (Connaway and Radford, 2007). IR systems, therefore, need to provide a user interface that allows the user to seamlessly transfer from looking up information to learning and investigating the information. Exploratory search is a new type of information seeking, which requires search systems to clarify users' information needs, learn from information in collections, and investigate complex information problems (White and Roth, 2009).

While lookup searches tend to require simple, single answers from IR systems, exploratory searches require more involvement from users and systems, maintaining a balance between browsing and focused searching.[2] In iterative searches, the user's information need is as-

[2] White and Roth (2009) adopt the definition of browsing by Kwasnik (1992), which is movement in an information space with the following activities: 1) persistently orienting to the environment, 2) marking of potentially relevant items for potential second considerations, 3) identification or recognition of potentially

sumed to be fixed, and every iteration of the search is essentially narrowing down the exact information requirement. On the other hand, in exploratory search, searchers explore more of the information space, and switch activities between browsing and focused searching (depicted in Figure 2.5).

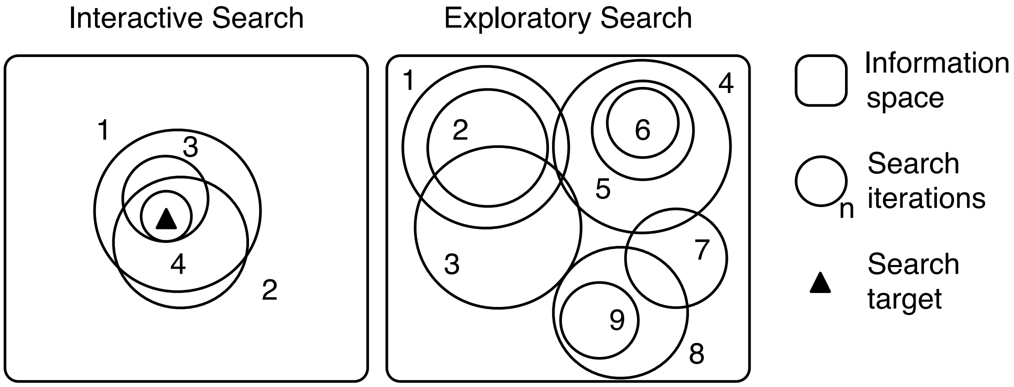


Figure 2.5: Iterative Search vs Exploratory Search (from White and Roth (2009))

2.2.2 Studies of Reference Interactions

2.2.2.1 Studies for IR Applications

The abstract models described above are often not detailed enough to be actually incorporated into system development (Wu and Liu, 2003). Thus, various researchers conducted systematic investigations concerning the nature and process of reference interactions with the goal of providing practical guidance for designing information retrieval systems or implementing reference services.

In the 1980s, a group of researchers at the City University of London (Belkin, Brooks, Daniels, and Oddy) conducted a series of studies that analyzed face-to-face reference interviews in order to develop design implications for an interactive component of IR systems. Based on their analysis, Belkin et al. (1983) proposes a specification of information system

relevant or definitely not relevant items, 4) resolution of anomalies, 5) comparison between items that serve to orient, identify, or solidify purposes and aims, and 6) transitions from one item to another (White and Roth, 2009, p.18).

interface, which they named the information provision mechanism (IPM). IPM specifies all the functions that are necessary for an interface of information systems. Based on the ASK hypothesis, these functions are designed to explicate users' anomalous states of knowledge and translate them into formal information requirements through interactive processes. A series of studies (Brooks, 1986; Daniels et al., 1985; Belkin et al., 1987b) further specified descriptions of the functions.

The studies of Belkin et al. also suggest that discourse analysis contributes to studies of information-seeking behavior by revealing the structure of information-seeking dialogues. The studies utilized the theory of focus shifting, proposed by Grosz (1978), in order to identify the concepts or themes that are the focus of attention in dialogue. The interview data were transcribed using rules based on the previous discourse analysis studies by Jefferson (1972) and Crouch and Lucia (1980).

White (1998) characterizes the question behavior in reference interviews on two dimensions: types of questions and categories of information. The findings include: a) about 50% of the questions asked by the librarians are verifications; b) about 22% of the questions asked by the librarians are judgmental questions or requests; c) users often ask questions as requests; and d) subjects and service dominates the content of dialogues. According to White, these findings suggest that the minimalist approach to interface incorporated into some Web search engines (in the 1990s) may not be effective, and that IR systems should not rely on the users to provide all the relevant search terms.

Wu and Liu (2003) analyzed how intermediaries elicit information from users within three dimensions: linguistic forms, utterance purposes, and communicative functions. Their statistical analysis revealed three distinct types of elicitation styles among the intermediaries:

- *Situationally oriented*, where intermediaries change forms, purposes, and functions, based on users and users' information needs,

- *Functionally oriented*, where intermediaries have a tendency to use specific communicative functions regardless of the users or the users' information needs, and

- *Stereotyped*, where intermediaries tend to use similar questioning strategies (for all three aspects) even with different users and user needs.

In this study (and in her dissertation study in 1993), Wu uses a term "elicitation", rather than "question" as in the previous studies, to explicitly refer to the micro-level information-seeking behavior "by which one seeks information to fill the gap in one's internal state of knowledge" (p. 1117), and to distinguish from a user's search statement or indication of information needs.

Wu (2005) further investigated the micro-level information-seeking (MLIS) by analyzing transcripts of online interviews between librarians and patrons. The distinction, which Wu makes, between the "question" by a user in a reference dialogue and "elicitation" by the user or the librarian is important in analyzing information-seeking dialogues. In this study, a user's "question" is treated as an "information provision" of his/her information need, while "elicitation" by either of the participants is treated as an "information request" for some aspect of the information need (Section 3.3.2 discusses details of the coding scheme).

2.2.2.2 Studies for Improving Reference Service

While the studies described above investigated reference interactions for designing and developing IR systems, other researchers investigated reference transactions for the improvement of reference services.[3]

Dervin and Dewdney (1986), based on the theoretical framework called "Sense-Making" (Dervin, 1983), propose a reference interview strategy, *neutral questioning*. Neutral questioning is designed to allow a librarian to better understand a user's viewpoint by asking open-ended, less-structured questions. However, studies show mixed results. Crouch and Lucia (1981) conducted discourse analysis of pre-search interviews and found some corre-

[3] The evaluation of reference services, although an important part of library studies, is considered outside the scope of this literature review. This subsection discusses studies that examined information-seeking behavior specifically. The researcher aims to elaborate the outcomes of this study to contribute to the field in the future.

lation between user satisfaction and open questions.[4] Allen (1988), however, investigated user-intermediary interactions and found no statistically significant difference between open and closed questions asked during mediated online searching.

Other studies combine the investigation of question strategies and information transfer. Auster and Lawton (1984) investigated the relationship among three constructs: 1) the interview techniques used by the librarian, 2) the amount of new information gained by the user as the result of the search, and 3) the user's ultimate satisfaction with the reference service. Their findings include: a) asking of open and closed questions has a modest effect on the amount that reference users learned; b) overall satisfaction is higher when open questions are asked; and c) those who learn more about their topics are more satisfied than those who learn less.

Ingwersen (1982) also investigated interview strategies at a reference service and reports that the use of open questions is scarce and the use of closed questions depends on the intermediary's knowledge of the subject area. The researcher suggests that provision of conceptual clarification and elicitation of the underlying problem or task are crucial to understanding the knowledge state of the user.

Radford (1993) investigated and compared perceptions by library users and librarians regarding their interpersonal communication in academic library reference interactions based on communication theory. The findings indicate that interpersonal communication in both its relationship-defining (relational) and information-transfer (content) dimensions is important to librarians' and users' perception of the reference interaction. Users attach great significance to the librarians' attitude and personal qualities. On the other hand, librarians emphasize information transfer to the user, and perceive relationship qualities to be of lesser importance. The study also produced a categorical scheme of interpersonal communication with three major themes: 1) *goals*: participants' desired outcomes for the interactions, 2) *facilitators*: qualities that have a positive impact on the perceptions of the interactions, and

[4] Crouch and Lucia's study was conducted here in the School of Information Studies, Syracuse University. The researchers published the details of the coding scheme, which was often consulted during the coding scheme development for this study.

3) *barriers*: characteristics that have a negative impact.

Radford's work contrasts with the studies described earlier, by focusing on the socio-emotional aspects of communication rather than the information transfer. The study set the interactional theory (Watzlawick et al., 1967) as its theoretical foundation, demonstrating the applicability of the theory to describing and explaining communicative behavior seen in reference interactions. Watzlawick's theory states five basic axioms of communication, one of which is the dual nature of interpersonal communication.[5] In this study, this dual nature of communication is included as part of the definition of dialogue acts (Section 2.3 discusses how dialogue act theory fits in with this notion).

Dewdney and Ross (1994) also claim that the librarian's behavior is the major factor in user satisfaction regarding the quality of reference services. Ross (1999) suggests the importance of "active engagement", "affective dimension", "trustworthiness" and the "social context" in non goal-oriented reference transactions (p. 783).

Since the emergence of internet technologies, many studies of information-seeking interactions started focusing on how to integrate online searching tools into reference transactions. A large number of studies (e.g. Kuhlthau et al. (1992), Wu (1993), Spink et al. (1995), and Saracevic et al. (1997)) examined reference interviews that require an online search, investigating how and why intermediaries or users make elicitation. In the field of reference, Janes (1999; 2002), Zumalt and Pasicznyuk (1999), and Ross and Nilsen (2000) examined how to utilize web technologies (e.g. search engines) as tools for reference services effectively. These studies led to the further integration of technologies and human expertise for information seeking – a premise of digital reference.

Studies of Digital Reference Interactions

One of the central goals of digital reference, as a field of research, is the incorporation of human expertise into information systems (Lankes, 2009)[6] Digital reference services started

[5] The idea of of duality of communication was first described from the perspective of psychiatry by Ruesch and Bateson (1951). Watzlawick's axioms are based on Ruesch and Bateson's work.

[6] In this document, the definition of digital reference does not subscribe to any specific media of delivery.

as an extension of traditional reference services using emails to extend the hours of reference services (Howard and Jankowski, 1986; Weise and Borgendale, 1986) or to experiment with the new technology for enhancing the services (Bristow, 1992). Soon after emails were introduced, online forms started being used for submitting questions for digital reference (Lagace, 1999; Janes et al., 1999) and by 2000, most academic libraries offered reference services through e-mail or web forms and some libraries started using instant messaging or chat as an additional medium of digital reference (Foley, 2002).

Initially, articles about digital reference were mostly limited to descriptive studies that surveyed the current states or report cases of digital reference implementations (Howard and Jankowski, 1986; Weise and Borgendale, 1986; Bristow, 1992; Lagace, 1999; Janes et al., 1999; Foley, 2002), or guidelines or suggestions for developing digital reference services (Lipow, 2002; Hirko and Ross, 2004). However, as digital reference services have become a viable alternative to the traditional face-to-face library reference transactions, studies that examine digital reference processes have increased rapidly. Under the term digital reference (or virtual reference), studies have been conducted with a wide variety of focuses and domains of inquiry: surveying users' expectations and satisfaction with digital reference (Ruppel and Fagan, 2002; Nilsen, 2004), evaluation of efficiency or effectiveness of digital reference services (Carter and Janes, 2000; Kaske and Arnold, 2002; White et al., 2003; Arnold and Kaske, 2005), setting the quality standards of the service (Kasowitz et al., 2000; McClure et al., 2002; Arnold and Kaske, 2005), analyzing the types of questions asked at digital reference (Carter and Janes, 2000; Diamond and Pease, 2001), etc.

Construction of a knowledge base that allows librarians and/or users to access outcomes of previous reference interviews is an important research area for the efficiency and effectiveness of the digital reference services. There have been efforts in research and practice in different fields such as science, education and libraries for creating archives of online information-seeking interactions. AskA services (Lankes and Kasowitz, 1998) are one type of such services that facilitate and archive information-seeking conversations between users and subject experts. However, archiving information-seeking conversations for future use, or automating digital references processes involves various issues.

Pomerantz (2003) claims that in order to develop such systems, one must identify the types of questions that are received by digital reference services. Pomerantz surveyed three bodies of literature: question answering systems, linguistics, and reference, and compiled five taxonomies of questions: 1) wh- words, 2) subjects of questions, 3) the functions of expected answers to questions, 4) the forms of expected answers to questions, and 5) types of sources from which answers may be drawn. The complexity of his taxonomies represent one of the challenges of developing systems to automate reference services. The taxonomies may be interpreted as facets of questions that need to be clarified for a system to respond to. In this view, these dimensions overlap with the coding schemes for analyzing reference interviews used by Crouch and Lucia (1981) or Belkin et al. (1983).

Nicholson and Lankes (2007) suggest developing a unified schema to archive digital reference transactions from different services in order to create a fielded, searchable knowledge base. Nicholson and Lankes argue that their framework is useful for developing bibliomining or data mining tools as library services, evaluation of digital reference services and service management, and modeling the collection of digital reference transactions as a complex adaptive system.

Lankes et al. (2006) seek an implementation of the library as a facilitator of technology-enhanced information interactions, which they call *Participatory Network*. This initiative is motivated by conversation theory by Pask (1975, 1976), which posits that knowledge is created through conversations. Conversation theory is based on cybernetics, the study of control processes in machineries or biological systems. The fundamental idea of the theory is that knowledge is created through conversations where instances of languages about a subject matter are exchanged and the understanding is enhanced. The theory is an attempt to formalize this process, in order to support the process with technologies. While Pask did not achieve the goal, his study provides two useful observations: 1) Pask identifies two levels of language used in conversations: L_1, in which conversation participants discuss the subject matter, and L_2 in which the process of learning about the subject matter is discussed; and 2)

Pask categorizes conversation participants based on their learning strategies: *serialists*, who progress in a sequential fashion, and with structure, and *holists*, who look for surrounding matters and higher order relations. According to Pask, an ideal learning process involves balancing both learning strategies. His observations coincide with models of information-seeking behavior that explain the different aspects of the information-seeking process, as well as the linguistic and communication theories that explain the dual nature of communication.

Some researchers have analyzed the informational or socio-emotional aspects of digital reference interviews, applying the methodologies for analyzing face-to-face reference transactions. Wu (2005) investigated micro-level information-seeking (MLIS), which she developed from her earlier work (Wu, 1993; Wu and Liu, 2003), by analyzing transcripts of online interviews between librarians and patrons (users). The findings include: a) patrons' and intermediaries' elicitation behavior differs in terms of frequency and time frame, supporting the assumption that intermediaries' elicitation is strategically pre-planned, while the patrons' behavior is situational; b) patrons' perplexities are situational; and c) patrons' elicitation behavior is related to their contextual variables.

Radford (2006b) analyzed interpersonal communication in digital reference services as part of a series of studies using chat reference data from OCLC (Radford and Connaway, 2005). The results show that interpersonal skills are important to successful face-to-face reference sessions and that they are present, modified from face-to-face sessions, in digital reference. Connaway and Radford (2011) also conducted a series of studies investigating the use of digital reference services, employing multiple methods (focus group, interviews, survey and content analysis). Among many findings, the study indicates that accuracy, a positive attitude by the librarian, and good communication are critical for the success of the reference service, and that query clarification is the key for accuracy and effectiveness.

2.2.3 Summary

Throughout the history of information science, various researchers repeatedly claimed the need to incorporate interactive processes into information retrieval systems in order to improve users' information experience. They did so by describing the interactions between intermediaries and information system users with abstract models of information-seeking behavior. This section reviewed the models proposed by Taylor (1968), Belkin et al. (1982), Bates (1989), Belkin (1993), Ingwersen (1996), Saracevic et al. (1997), Wilson (1999), and Marchionini (2006). Although these models are helpful in understanding the phenomenon, they are less often applied to the implementation of information retrieval systems in the real world because of a lack of details for input to system design and development. The studies described in Section 2.2.2 reported concrete observations for more practical implications. Daniels et al. (1985) and Brooks (1986) investigated the execution of specific functions in information-seeking interactions during reference interviews using discourse analysis. Wu (1993), Wu and Liu (2003), and Wu (2005) analyzed the elicitation behavior in different contexts. Radford (1993, 2006b) focused on the socio-emotional aspect of reference interactions based on communication theory.

The models and observations described in this section set expectations to the phenomena (assumptions) and hypotheses that need to be tested in this research. In addition, these studies helped the researcher in choosing the conceptual and methodological framework for the research, by demonstrating the applicability of discourse analysis and introducing the dual nature of the communication to the study of information-seeking behavior.

2.3 Studies of Discourse Semantics

2.3.1 Linguistic Analysis

The study of linguistics offers a range of methods and theories for analyzing linguistic behavior. The methods and theories are often adopted by other fields of study or disciplines that deal with linguistic phenomena in some way: art, law, sociology, anthropology, computer science, information science, etc. According to Liddy (1998), there are seven levels of linguistic analysis that may be utilized for information retrieval research and implementation. They are, in order of increasing complexity, as follows:

- *Phonological*: interpretation of speech sounds
- *Morphological*: analysis of components of words such as prefixes, suffixes, and roots
- *Lexical*: analysis of words such as word meaning and part of speech
- *Syntactic*: analysis of the grammatical structures of sentences
- *Semantic*: determining the meaning of a sentence, including disambiguation of word meanings in context
- *Discourse*: interpreting structure and meaning conveyed by texts of more than a sentence
- *Pragmatic*: understanding the use of language in situations and with world knowledge

Many IR systems, either research or commercial systems, utilize linguistic analysis at the morphological or lexical level by automating the process. Stemming is an example of an application of morphological level analysis that is used by many IR systems. Stemming deletes derivational morphemes, such as "ation" of the word "organization", or inflectional morphemes, such as the third-person singular "s", in order to associate relevant terms in different forms. Using a thesaurus or lexicon for indexing is an example of lexical level analysis, widely used for IR. By looking up a term in a thesaurus, an IR system can automatically add query terms (*query expansion*) to increase the recall, or identify the meanings of ambiguous words in a query or a document (*word sense disambiguation*).

NLP techniques have been proven to be effective in a wide range of information technologies such as information extraction, information management and knowledge organization (el Hadi, 2004), and commercial IR systems have been adopting NLP techniques gradually over time (e.g. adoption of stemming by Google, as documented by Brin and Page (1998) and Uyar (2009)).

The recent proliferation of computer-mediated communication and social media suggests more web contents are created in an interactive context or as part of conversation. As Lease (2007) argues, higher levels of linguistic analysis still have the potential to improve IR systems, especially in restricted domains. Discourse analysis, in particular, has been used to reveal the structure of a certain type of information object in order to improve a variety of text processing systems that deal with a particular type of text, such as document summarization, information extraction, and text retrieval systems. A pioneering work in this thread of research was by Liddy (1991), who analyzed the discourse-level structure of information abstracts of empirical work. Liddy's study is based on an array of discourse linguistic studies that discovered predictable structures of various kinds of texts (e.g. Propp (1958) for folk-tales, Cohen (1987) for arguments, and van Dijk (1980) for newspaper articles), and cognitive psychology studies that suggested that human cognition processes information by organizing the information in a structure (Miller, 1956; Rumelhart, 1977, 1980). The results of Liddy's study indicate the presence of a detectable structure in empirical abstracts, which would be useful in a variety of text-based information processing systems. Application of discourse level linguistic analysis to automatic information extraction has also been suggested by Kando (1997), and Teufel (1999), and IR applications have been suggested by Oddy et al. (1992) and Kando (1995) among others.

The proposed research applies discourse analysis to digital reference transactions with two goals in mind: 1) revealing the structure of such interactions to inform the design of interactive processes for IR systems; and 2) understanding how information requests and information provision are expressed in dialogues to utilize the interactions as information resources.

2.3.2 Discourse Analysis

Discourse analysis, among other sub-disciplines of linguistics, is unique in that it does not refer to a particular set of methods or theories. Rather, the term "discourse analysis" often means simply a study of language beyond the sentence level. Tannen (2007) argues:

> "[T]he term "discourse analysis" does not refer to a particular method of analysis. It does not entail a single theory or coherent set of theories. Moreover, the term does not describe a theoretical perspective or methodological framework at all. It simply describes the object of study: language beyond the sentence". (p. 6)

Johnstone (2002) argues most discourse analysis studies can be grouped into one of two kinds of research, based on the epistemological assumptions: descriptive research and critical research. Descriptive discourse analysis studies are based on the belief that the world can be described or measured accurately and the role of scientific research is to produce such descriptions or measurements in order to create solutions for problems in the world. Thus, the goal of descriptive research is to describe phenomena surrounding linguistic behavior: the structure of language (descriptive linguistics), the mechanics of social interactions (Conversation Analysis), the interactions between texts and social contexts (linguistic anthropology), etc.

The idea that the structure of language reflects meaning dates back to work by Prague School linguists in the 1930s, but it was Zellig S. Harris, who defined the term as a formal method for analyzing "connected speech (or writing)" in terms of linguistic structure (Harris, 1952, p.1). Intensive work on discourse structure was done during the following time periods by linguists such as Labov and Waletzky (1967), Pike (1967), Grimes (1975) and Halliday and Hasan (1976), all of whom attempted to provide a way to describe human linguistic behavior in a universal manner. Pike (1967) generalized the notions of *phonemes* and *phonetics* in phonology and applied them to the higher levels of linguistic analysis. His theory, *Tagmemics*, is an attempt to provide a unified theory that explains the structure of language and human behavior at all levels.[7] Grimes (1975) analyzed discourse structure in various languages and different contexts to investigate the parameters of discourse structure in languages across

[7] Pike coined the terms *etic* and *emic*, which are widely used by anthropology today: emic refers to a behavior that is meaningful to the observer in the context or culture, while etic refers to a behavior that is insignificant in the given context or culture, but can be observed in different contexts or culture (Harris, 1976).

the world. Halliday and Hasan (1976) provided a detailed account of the discourse structure of English. They claim that a text is not simply a sequence of sentences but consists of a grammatical and semantic unit, called *cohesion*. Texts are *cohesive* when sentences hold *cohesive relations*, such as *reference*, *substitution*, *ellipsis*, and *conjunction*.

The presuppositions of descriptive research, which were based on the positivistic attitude to world knowledge, have been called into question in the past several decades under the influence of linguistic relativism and critical social theories. As a result, discourse analysis has increasingly come to be used in the service of critical goals in social sciences (Johnstone, 2002). On the other hand, application of discourse level analysis to language technologies has been rapidly increasing, as research at lower levels of analysis (e.g. morphological or syntactic) have matured and new aspects of linguistic behavior started attracting attention (e.g. coreference, question answering, sentiment).

In the most general term, discourse analysis can be characterized with the following two properties: 1) The unit of analysis is larger than a sentence; and 2) Subjects of analysis are actual linguistic utterances in use, rather than artificially constructed exemplars for research. The proposed study follows the tradition of descriptive linguistics and seeks the relationship between structure, meaning (or function), and use of language in information-seeking interactions. Thus, in the rest of the document, the term *discourse analysis* refers to the descriptive discourse analysis studies.

2.3.2.1 Speech Act Theory

Analysis of the relationship between structure, meaning and use of language in information-seeking interactions, requires a theoretical framework that orients the research with the right focus and that provides a theoretical lens to the investigation. Among the most prominent work that focused on describing the relationship between the structure and function of language is speech act theory proposed by Austin (1975). Speech act theory was developed as an antithesis to the mainstream approach of the philosophy of language, which then focused on the relationship between formal semantics as the representation of meaning and syntactic construction as structure. Such approaches, most notably developed by philosophers of language such as Bertrand Russell and Friedrich Ludwig Gottlob Frege, excluded any utterance other than ones that declare a statement that can be represented in formal logic (e.g. “Socrates is a human.” as opposed to greetings, e.g. “Hello.” or expressions of gratitude, e.g. “Thank you.”) from analysis. Austin conducted a long series of systematic investigations of those utterances that had been left out and proposed the theory, which claims that speech is a kind of action with different types of effects (speech acts). In Austin’s definition, speech acts are analyzed at three levels: *locutionary acts*, which is the actual utterance and its literal (or denoted) meaning – this is the level that the traditional philosophy of language focused on, *illocutionary acts*, which is the intended (or implied) meaning, and *perlocutionary acts*, which is the actual effect of the utterance, such as persuading, scaring, enlightening, inspiring, etc. While Austin’s theory was intended to provide a more general description of linguistic behavior than earlier approaches, the contribution of the theory is often attributed to the inclusion of the illocutionary effect of language into analysis. Searle (1969) elaborated the idea of illocutionary acts, which he called *indirect speech acts*, and developed a classification scheme of illocutionary acts.

2.3.2.2 Dialogue Acts

Dialogue act (DA) classification is a newer development, which classifies functions of utterances in dialogues in particular. Like speech acts, DAs treat an utterance in a dialogue as a kind of action, but often incorporate theoretical developments in analyzing dialogues in terms of the maintenance or management of a dialogue, such as understanding of turn-taking mechanisms (Sacks et al., 1974) or specification of underlying expectations for adjacency pairs (Goffman, 1981) – a pair of utterances, such as a question and an answer or a greeting and a response. Bunt (1994) defines DAs as “functional units used by the speaker to change the context” (p. 3). But there is not much agreement within research communities on the definition of a DA, and some studies adopt the notion a priori, while other studies define the notion inductively, as coding schemes are developed. Below are some of the definitions/descriptions that previous studies have provided:

- “the meaning of an utterance at the level of illocutionary force” (Stolcke et al., 2000, p. 340)
- “a concise abstraction of a speaker’s intention” (Samuel et al., 1998a, p. 1150)
- “the project of speakers’ intention in communication, the illocutionary force that an utterance accomplishes through processes of interactions, with particular reference to the notion of ’talk-in-interaction’ in conversation analysis where the structure of talk emerges through processes of interaction between participants” (Wu and Liu, 2003, p. 1120)
- “some function of an utterance in a dialogue, not reducible to its syntactic or semantic content” (Clark and Popescu-Belis, 2004, p. 2)

Most studies use the term as a variation of speech act, but DAs are often defined for a small domain of interest and specific focus, and thus classifications tend to be domain dependent and involve more details. There have been efforts, however, to create domain-independent DA classification schemes, such as DAMSL (Core and Allen, 1997), DIT++ (Bunt, 2006), and DiaMSL (Bunt et al., 2010). These coding schemes are, naturally, even larger and more complex than other DA coding schemes, which may be disadvantageous for inter-coder reliability and machine learning experiments. Researchers sometimes use a subset of the original code set in order to overcome such difficulties, e.g. Jurafsky et al. (1997) and Stolcke et al. (2000).

The DA analysis was used predominantly in the speech and dialogue communities to create a structural representation of dialogues for designing a spoken dialogue system (Jurafsky and Martin, 2008, Ch. 19). But as mentioned earlier, as higher levels of linguistic analysis (semantics, discourse and pragmatics) are increasingly applied to research on information technologies (e.g. IR and Question Answering), DA analysis, given its focus on the functional aspects and interaction of linguistic behavior, is becoming a promising approach for integrating high-level linguistic analyses into research and development of new information technologies. Among various notions of DAs proposed in the previous studies, this study adopts the definition of dialogue acts by Bunt (1994), and his theoretical framework, the Dynamic Interpretation Theory.

2.3.2.3 Dynamic Interpretation Theory

Dynamic Interpretation Theory, proposed by Bunt (1994), provides the theoretical assumptions for the dialogue act classification scheme, DIT++, in the following two ways:

1) The theory defines dialogue acts (quoted earlier) as the unification of formal semantics and speech act theories.

2) The theory hypothesizes that two types of tasks are carried out during a conversation: *underlying tasks* and *communicative tasks*.

In the following, these two arguments and their relations for the proposed study are discussed.

2.3.2.4 Defining Dialogue Acts

Bunt (1994) defines DA by unifying some of the modern formal semantic theories (e.g. Dynamic Montague Grammar by Groenendijk and Stokhof (1989)) and speech act theory. His definition of dialogue acts (quoted above) entails that a dialogue act specifies the function of an utterance, as well as the unit of the function. Utterances must be analyzed by their effects on the context while the unit of analysis is defined by the function. The theory organizes the analysis of DAs into three levels: 1) the *utterance form*, which is the observed, surface form of the utterance, 2) the *semantic content*, which is the formal predicative representation of the meaning, and 3) the *communicative function*, which is the effect of an utterance. The relationship between 1) and 2) is the focus of the formal semantics, the relationship between 1) and 3) is the focus of the speech act, and Bunt's theory is an attempt to analyze the relationship among the three aspects, thus unifying the previous two approaches. In the proposed study, analysis of the utterance forms is done by identifying text segments that correspond to dialogue acts in messages. The analysis of communicative function is done by specification of dialogue act labels following the coding scheme that has been developed for the study (The coding scheme is described in Section 3.3.2). It is assumed that the semantic content is treated in two ways in this study: 1) through the annotation of data, wherein human annotators identify communicative functions based on the data, and thus implicitly (unconsciously or consciously) relate the communicative functions to the utterance form and/or semantic content; and 2) through the feature representations for machine learning, wherein lexical (word-level) semantics are assumed to be represented by the word vector features (See Section 3.4.2 for the descriptions of the features used in the machine learning experiments.).

As a function, Bunt defines a DA as a transformation of participants' cognitive states, which he calls *contexts*. In defining dialogue acts, Bunt (1994) also identifies the following five categories of contexts that dialogue acts may have effects on:

1. *linguistic context*: properties of the surrounding linguistic material (textual or spoken)
2. *physical context*: physical circumstances where the dialogue takes place
3. *semantic context*: underlying tasks of the dialogue and their related circumstances
4. *social context*: types of interactive situations and roles of the participants in terms of social obligations and rights
5. *cognitive context*: participants' cognitive states

In effect, Bunt classifies the functions of utterances by their domains much like mathematical functions are formally categorized. While the present study does not make use of all the categorization of contexts that Bunt proposed, it does follow the approach of categorizing functions based on their contexts (See Section 3.3 for the details of the method of analysis.).

2.3.2.5 Duality of Dialogue Process

The dynamic interpretation theory hypothesizes that dialogues are always carried out by participants performing the following two kinds of tasks:

a) tasks to achieve the goal that motivated the dialogue (*underlying tasks*), and

b) tasks to maintain the dialogue itself in order to achieve goals that are associated to the context of the dialogue (*communicative tasks*).

Thus, the theory implies that a coding scheme for analyzing functions of utterances must be defined with at least two major categories: ones that are related to performing the underlying tasks, and the others that are related to performing the communicative tasks.

The underlying task of reference interviews is to satisfy the user's information need. And in order to achieve that goal, the librarians and the users are engaged in two kinds of tasks: 1) information exchanges between the user and the librarians and 2) physical tasks that are related to the information-seeking process, e.g. searching online, looking up in a catalogue, examining information objects, etc. The detailed descriptions of the functions are presented in Section 3.3.2.

2.3.3 Summary

Many IR research researchers once agreed that performance of IR systems could be improved more effectively by statistical methods rather than by linguistically motivated approaches (Lewis and Jones, 1996; Sparck Jones, 1999; Smeaton, 1999; Allan, 2000; Robertson, 2008). Most major search engines and research IR systems, however, now employ low-level NLP techniques and have been slowly employing techniques at higher levels. As Lease (2007) argues, higher levels of linguistic analysis, such as discourse analysis, have the potential to improve IR systems, by revealing the semantics beyond the sentence structure.

The proposed research applies discourse analysis to digital reference transactions with two goals in mind: 1) revealing the structure of such interactions to inform the design of interactive processes for IR systems; and 2) understanding how information request and information provision are expressed in dialogues to utilize the interactions as information resources.

This section reviewed the theories and methods for analyzing discourse semantics, in particular, dialogue acts. Dialogue acts are "functional units used by the speaker to change the context" (Bunt, 1994, p. 3), based on speech act theory (Austin, 1975; Searle, 1969). Dialogue acts classify utterances in dialogues along two dimensions: functions and contexts. Dynamic interpretation theory, proposed by Bunt (1994) explains the two aspects of information-seeking interactions, information transfer and socio-emotional aspects, by stating that dialogues are carried out by achieving two goals: communicative goals and underlying goals. Bunt's theory confirms the models and observations produced by the studies in Section 2.2, in particular, by Radford (1993, 2006b). Bunt's classification scheme (DIT++) is used as the basis for the coding scheme of the study, while his theory explains the theoretical motivation of the scheme.

2.4 Dialogue Acts and Machine Learning

This section provides an overview of the machine learning dialogue act studies. The studies discussed in this section informed the design of the machine learning experiment stage of the study – the choice of the machine learning algorithms, features, software packages, and overall design. In this section, the notion of *dialogue act* is loosely defined as the intention of the speaker. Some of the studies discussed here may not use the same term. The studies described in this section are summarized in Table A.3 in the Appendices.

Different approaches for learning dialogue acts have been researched and categorized into two types. The first approach is to induce an abstract temporal model of dialogues based on observed evidence. The most common example of this approach is the Hidden Markov Model (HMM), which constructs a probabilistic temporal model, represented by a weighted finite state automaton. The HMM has been widely used for speech recognition or hand-writing recognition and was applied to many of the early automatic dialogue act annotation studies by researchers from the speech and dialogue communities.

The second approach for automatic dialogue act annotation is to treat the problem as a machine learning text classification problem, where a system attempts to learn a function that maps utterances to dialogue act labels based on observed evidence (features). The performance of such systems mainly depends on the algorithm that the system deploys and the features available to the system. Some algorithms operate directly based on the observed evidence (*instance-based learning*), while many other algorithms operate on mathematical abstraction of the evidence. For example, Naive Bayes operates on an inference network (Bayesian Network), back-propagation operates on an artificial neural network, and the Support Vector Machine (SVM) operates on a vector space. Among the various algorithms, the SVM has been one of the most-widely used algorithms for text classification tasks in recent years.

In the following, Section 2.4.1 describes the label sequence learning in detail, including examples of previous studies and their design principles. Section 2.4.2 describes text classification tasks and SVM. Lastly, Section 2.4.4 describes recent developments in machine learning dialogue annotation studies.

2.4.1 Label Sequence Learning with HMM

2.4.1.1 Previous Studies

The Hidden Markov Model (HMM) is the most common example of the label sequence learning approach for automatic dialogue act annotation. The approach has been widely used for speech recognition or hand-writing recognition and applied to many of the early automatic dialogue act annotation studies done by researchers from the speech and dialogue communities. Applications of HMMs to dialogue act annotation are found in studies from the late 1990s (Kita et al., 1996; Reithinger et al., 1996; Reithinger and Klesen, 1997; Jurafsky et al., 1997; Stolcke et al., 2000).

Kita et al. (1996) used an Ergodic HMM and the ALERGIA algorithm to learn probabilistic automata to represent typical dialogue structures such as turn-taking and speech act sequencing. The study used an annotation scheme called Illocutionary Force Type (IFT), which represents "an abstraction of the speakers' intention in terms of the type of action the speaker intends by the utterance" (page 1). IFT consists of nine types, which are similar to the speech acts defined by Searle (1969). The researchers annotated conversations between conference staff and participants by phone and constructed a probabilistic model by applying the algorithm. Kita et al. claim that the Ergodic HMM and the probabilistic automaton derived by their algorithm successfully captured the local discourse structure of the dialogues. The study is one of the earliest studies that attempts to automatically identify discourse characteristics of human communication using a probabilistic approach.

While the goal of the study by Kita et al. was to generate a model of dialogues, Reithinger et al. (1996) developed a system to predict dialogue acts. The prediction was used by different modules of their speech-to-speech translation system called VERBMOBIL. The study employed an annotation scheme which included 42 labels and used over 300 transcribed scheduling dialogues as data. The experiment compared two different algorithms: the Markov chain method and the dynamic adaptation method, and four kinds of features: the baseline n-gram language model, speaker information, dialogue grammar, and mirroring.[8] Overall, the system achieved accuracies of 72.24% to 76.05% with different configurations, which is relatively high, even by today's standards.

As for the algorithms, the Markov chain method outperformed the dynamic adaptation method with all the combinations of features. As for the features, the results show the advantages of using speaker information and mirroring over simply using the baseline features.

While previous studies for automatic detection of dialogue acts focused on dialogues with specific tasks, and thus contained specific dialogue acts, a group of researchers worked on dialogue acts of unconstrained, spontaneous conversations (Jurafsky et al., 1997; Stolcke et al., 1998, 2000). In order to accommodate a wide variety of dialogue acts observed in the data, the researchers employed DAMSL (Dialogue Act Markup in Several Layers), a high-level dialogue act annotation scheme for general conversations, developed by Core and Allen (1997). The tag set was further extended to include 220 tags and then reduced to 42 tags (SWBD-DAMSL) by clustering. The series of work by Jurafsky, Stolcke, and others

[8] Mirroring refers to a process where the size of the data is doubled by simply exchanging the speaker information of all the utterances.

established a benchmark of automatic dialogue annotation tasks by using one of the most used data sets (SWBD) and annotation scheme (SWBD-DAMSL). By combining three types of features: prosodic (tones), word occurrences, and surrounding dialogue acts, their system achieved 65% of accuracy.

2.4.1.2 Algorithm

HMM constructs a probabilistic temporal model, represented by a weighted finite state automaton, based on observed pieces of evidence. In the case of dialogue act classification, each hidden state represents a dialogue act (or a combination of dialogue acts) of an utterance, while each piece of observed evidence represents an actual utterance (text segment). A transition from a state to another state represents a possible sequence of dialogue acts (Figure 2.6). Formally, given a sequence of observations of utterances, $E = \langle e_1, e_2, ..., e_n \rangle$,

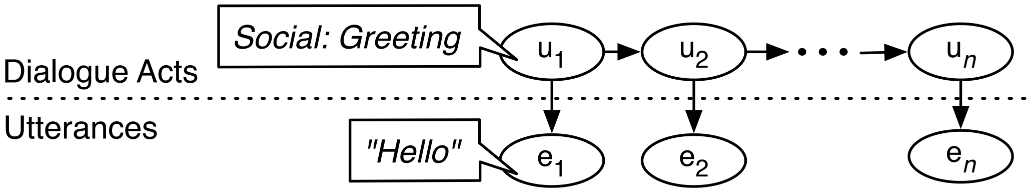


Figure 2.6: Probabilistic Temporal Model of Dialogue Acts

the task of automatic dialogue act annotation is to find a sequence of dialogue act labels, $U* = \langle u_1^*, u_2^*, ..., u_n^* \rangle$ that has the highest *posterior probability* $P(U|E)$, i.e. the probability of the sequence of dialogue acts is indeed U, given the observations E, among all the possible sequences of $U = \langle u_1, u_2, ..., u_n \rangle$. That is:

$$U^* = argmax_{U'} P(U|E)$$

By Bayes' Theorem:

$$= argmax_U \frac{P(U)P(E|U)}{P(E)}$$

$$= argmax_U P(U)P(E|U) \tag{2.1}$$

Using Bayes' Theorem, the posterior probability is transformed into a combination of a *prior probability*, in this case, a *state model*, $P(U)$, the probability of the dialogue act sequence is U, and a *likelihood*, in this case, an *observation model*, $P(E|U)$, the probability of the evidence set E is observed, given that the dialogue act sequence is U. The benefit of this transformation is that, in general, constructing the *state model* and the *observation model* is more straightforward than estimating the posterior probability. In this case:[9]

$$P(U) = \prod_{1 \leq i \leq n} P(u_i|u_0, ..., u_{i-1}) \tag{2.2}$$

$$P(E|U) = \prod_{1 \leq i \leq n} P(e_i|e_0, ..., e_{i-1}, u_0, ..., u_i) \tag{2.3}$$

These formulas are often further simplified by making a couple of assumptions on the probability distribution. The first assumption, called *Markov assumption*, assumes that each hidden state only depends on a finite history of previous states. For example, one may assume that each state depends only on the immediate successor of the state, i.e.:

$$P(u_x|u_0, ..., u_{x-1}) = P(u_x|u_{x-1}) \tag{2.4}$$

In general, a Markov assumption that considers the states with *n*-state apart is called the n*th-order Markov process*, for example, the case above is a *first-order Markov process*. By using this assumption, the state model (2.2) can be simplified to the following:

$$P(U) = \prod_{1 \leq i \leq n} P(u_i|u_{i-1}) \tag{2.5}$$

[9] The variable u_0 here is used to denote the starting of the sequence. For example, $P(u_1|u_0)$ represents the probability of the dialogue act u_1 occurs as the first state. The same applies to the variable e_0.

The second assumption assumes that each observation depends only on the corresponding hidden state, i.e.:

$$P(e_x|e_0, ..., e_{x-1}, u_0, ..., u_x) = P(e_x|u_x) \tag{2.6}$$

By accepting this assumption, the observation model (2.3) is now simplified to the following:

$$P(E|U) = \prod_{1 \leq i \leq n} P(e_i|u_i) \tag{2.7}$$

And lastly, the entire distribution (2.1) can be simplified by 2.5 and 2.7 to the following:

$$U = argmax_U \prod_{1 \leq i \leq n} P(u_i|u_{i-1})P(e_i|u_i) \tag{2.8}$$

The transformation of a posterior probability into a combination of a prior probability and a likelihood estimation is a fundamental technique to many probabilistic approaches that has been applied to human language technologies – the most relevant work to this proposed research is the probabilistic approach for information retrieval proposed by Robertson (1977). Figures 2.7 and 2.8 show the simplification of the model that was made by the two assumptions above.

The basic structure of the HMM presented above can be used for different tasks, such as the following:[10]

1. Estimating the dialogue act of the current utterance, based on the history of observations, i.e. $P(U_t|e_1, ..., e_t)$.

   This task is useful for generating the semantic representation of the current utterance, for example, for the purpose of simultaneous translation.

2. Predicting the dialogue act of s state in the future, i.e. $P(U_{t+k}|e_1, ..., e_t)$, for some $k > 0$.

[10] This list is based on the list of the basic inference tasks appearing on Russell and Norvig (2009).

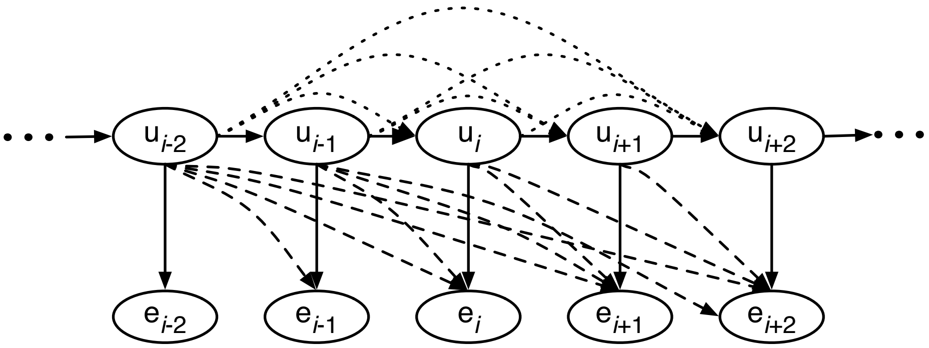


Figure 2.7: Probability dependencies of the original model

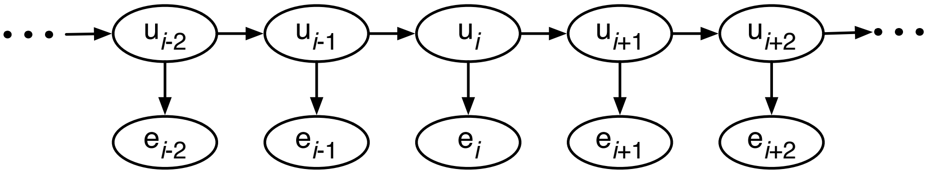


Figure 2.8: Probability dependencies of the simplified model

This function is commonly used for dialogue systems, where responses are needed to be generated based on evaluation of possible courses of actions.

3. Estimating the past state, i.e. $P(U_k|e_1, ..., e_t)$, for some $0 < k < t$.

   This task is useful for generating the semantic representation of an utterance in the past, for example, for the purpose of consecutive translation or summarization.

4. Finding out the sequence of states that is most likely to have generated the observations, i.e. $P(u_1, ..., u_t|e_1, ..., e_t)$.

   Algorithms for this task are commonly used for speech recognition systems, where the goal is to find a sequence of words that is most-likely to have generated observed sounds.

The tasks 1), 2), and 3) compute posterior probabilities for a single state, while the last task computes posterior probabilities for all the states in a dialogue. Thus, while the former tasks require a relatively simple iterative algorithm (e.g. recursive estimation or forward-backward algorithms), the last problem involves additional complexity. The HMM solves this last task using an iterative algorithm, called the Viterbi algorithm (or equivalent matrix operations). Note that the task of a simple text classifier in this context would be to classify the dialogue act of a given utterance, which is roughly equivalent to calculating $P(U_t|e_t)$, which, in turn, is comparable to the tasks 1), 2), or 3), with the zeroth-order Markov process.

2.4.2 Text Classification with SVM

2.4.2.1 Previous Studies

The goal of text classification is to classify pieces of text into a fixed number of predefined categories (*classes*). The goal of machine learning text classification systems is, then, to learn the decision process of classification from exemplars, and to automate the process. Given the rapid increase of digital communication, the text classification algorithms have been applied to dialogue act analysis of various digital documents such as emails and discussion board messages.

Cohen et al. (2004) proposed an approach to automatically classify e-mail texts based on "email speech acts", which represent "the intent of the sender" (page 309). Email speech acts are defined by taxonomies of verbs such as request, deliver, propose, or commit, and taxonomies of nouns such as opinion, data, meeting, etc. Cohen et al. compared a baseline SVM system, which used TFIDF-weighted bag-of-words, with different additional features such as: word bigrams, extracted time and date expressions, words near proper nouns or personal pronouns, and POS counts. They also compared different learning algorithms (with the baseline features) such as voted perceptron, decision tree, decision tree with boosting, and SVM. The experiment showed that SVM with additional features such as POS tag frequencies, could learn the proposed email acts reasonably well. Cohen et al.'s work is an example of an application of domain-specific dialogue acts.

Carvalho and Cohen (2006) followed up the work of Cohen et al. (2004) and improved the classification algorithms by incorporating n-gram features. Carvalho and Cohen also analyzed the most " meaningful" n-gram sequences by computing the information gain of each sequence, and concluded that n-gram features with high information gain agree with linguistic intuitions about expressing different email speech acts.

Hu et al. (2009) developed an annotation scheme for verbal interaction which can be applied to corpora that vary across many dimensions such as modality of signal (oral, textual), medium (e.g. email, voice alone, voice over electronic channel), register (such as informal conversation versus formal legal interrogation), number of participants, or immediacy (online versus offline). The researchers showed that a structured SVM classifier successfully identified the dialogue acts in the corpora. Their intention for developing the new annotation scheme was to compare different modes of communication, and thus, the scheme was designed to be more abstract than previous work.

2.4.2.2 Algorithm

The SVM was originally invented by Cortes and Vapnik (1995), and introduced to text classification tasks by Joachims (1998). The SVM algorithm operates over a vector space and finds a hyperplane that separates the instances in the vector space with the largest margin.

The algorithm is based on the mathematical fact that a vector space is always linearly separable if the dimension is high enough (namely n - 1 for n features), and the features are represented in the vector space in a certain, reasonable manner.[11] Figure 2.9 demonstrates a comparison between the separation in a two-dimensional space and a three-dimensional space, based on Russell and Norvig (2009). The two graphs show that the circular decision boundary (on the left) is transformed to a linear boundary (on the right), after mapping the two-dimensional vector to the three-dimensional vector.

Finding the optimal linear separator in a vector space is a quadratic programming problem, which finds parameters to minimize or maximize a quadratic function. The major advantage of SVMs comes from the fact that a solution of quadratic programming problems is known and can be effectively solved by using dot products of points[12] and avoiding re-

[11] This is defined in terms of the characteristics of a kernel function, by Mercer's theorem (1909), discussion of which is outside the scope of this document. Description and proof of the theorem can be found elsewhere, e.g. Shawe-Taylor and Cristianini (2000).

[12] A dot product is calculated from two equal-sequences of numbers, by multiplying corresponding entries and adding up those products.

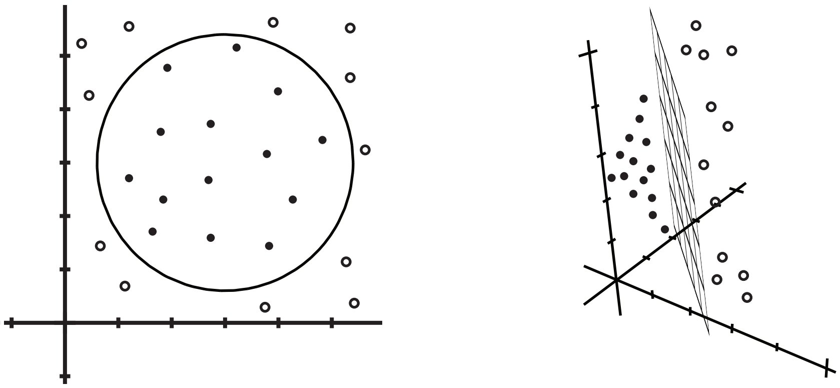


Figure 2.9: Examples of separations in different dimensions

ferring to each point in the vector space, which reduces the complexity of the computation greatly. Furthermore, by Mercer's Theorem (1909), SVMs can utilize different functions, called *kernel functions*, to calculate the dot products of points in different high dimensional spaces. Thus, while often the simplest kernel (linear kernel) provides acceptable experiment results, researchers may utilize different kernel functions for different tasks or even invent new ones to improve the performance.

It is believed that the SVM algorithm is well suited for text classification tasks, because its learning ability is independent from the number of dimensions – the algorithm finds the optimal separator based on margins measured by the number of data points, rather than the number of features. Thus SVMs can perform learning tasks that require high dimensions with a lower risk of overfitting. In addition, the SVM learning is a fully automatic process, eliminating the need for manual parameter tuning, and tends to provide better performance over other learning methods (Cardoso-Cachopo and Oliveira, 2003). For dialogue act classification, Cohen et al. (2004) and Hu et al. (2009) have used SVMs and showed promising results.

SVMs are available in various forms of software packages: LIBSVM, developed by Chang and Lin (2001), is an open source software library in C++ and Java, which provides interfaces for various tools (e.g. WEKA, RapidMiner, and R) and languages (e.g. Ruby, Python, and Perl), while Joachims (1999) has implemented a program in C, called SVM^{light} (See Ivanciuc (2007) for a list of SVM software packages).

2.4.3 Hidden Markov Support Vector Machines

While the HMM has been the predominant formalism for modeling and predicting label sequences, its limitations have been pointed out by some researchers. McCallum et al. (2000) argue that many text applications would benefit from a richer representation than the simple language models that HMMs typically employ. Additionally, using a *generative* model, such as HMMs, for *conditional* problems, such as text classification, is inappropriate and, in particular, it is not practical to rely on a joint probability over observation and label sequences, because the inference for such models is intractable (Lafferty et al., 2001). Altun et al. (2003) lists three major limitations with the approach as follows: 1) The algorithm is typically trained in a non-discriminative manner; 2) The Markov assumption is too restrictive; and 3) The algorithms rely on the feature representations and lack of the power of kernel-based methods.

Various algorithms have been invented to overcome the shortcomings of the HMMs. In order to segment questions and answers in FAQ documents, McCallum et al. (2000) proposed the Maximum Entropy Model, which represents the probability of a state given an observation and the previous state (Maximum Entropy Markov Model, MEMM).Lafferty et al. (2001) invented Conditional Random Fields (CRFs), avoiding a fundamental limitation of the MEMM, which is a bias towards states with few successor states. And lastly Altun et al. (2003) combined arguably the two most successful learning algorithms, namely HMM and SVM, and proposed Hidden Markov Support Vector Machines (HM-SVM).

2.4.4 Other Machine Learning Dialogue Act Studies

Other Algorithms

While HMM and SVM are the most successful algorithms, there certainly are studies that use different algorithms for learning dialogue acts.

Samuel et al. (1998a) uses Transformation Based Learning (TBL, Brill (1995)) for detecting dialogue acts. TBL is an effective learning algorithm for Part-of-Speech tagging, which is, to an extent, similar and related to dialogue act tagging. TBL's learning processes are easier to interpret for humans and are often considered to be similar to the thought processes of humans. Samuel et al. applied TBL to various features and achieved similar accuracies (75.12% on average over five runs) to Reithinger et al.'s system with the same experiment. One of the unique features that Samuel used is called "dialogue act cues", which is a set of substrings that appear frequently in the dialogues and that help determine the appropriate dialogue acts.

Walker and Passonneau (2001) developed a dialogue act tagging scheme for providing quantitative dialogue metrics in evaluating DARPA COMMUNICATOR spoken dialogue systems. Walker's system automatically tags utterances by pattern matching. The study shows that the dialogue acts are used to quantify the amount of effort spent to maintain the dialogue rather than actually carrying out the task that the system is supposed to do.

Kim et al. (2006) classified messages on discussion boards for an online class based on dialogue acts in order to assess participants' roles and identified discussion threads that might reveal users' confusions or have unanswered questions. In the study, messages on the discussion forums were annotated using customized 6-label speech acts ($kappa = 0.7$). The annotated data is analyzed by frequency distributions, transition probabilities (bi-gram distributions), roles that students played in the discussions, and patterns observed within threads. The dialogue acts were then used to improve automatic detection of conversation focus, by combining a graph-based algorithm with the dialogue act annotations. The researchers claim "analysis of human conversation via online discussions provides a basis for the development of future information extraction and intelligent assistance techniques for online discussions" (page 10).

Lan et al. (2008) used the Maximum Entropy (ME) model to learn dialogue acts. Using the SWBD as the corpus and SWBD-DAMSL as the classification scheme, Lan et al.'s system achieved a promising accuracy of 75.03%.

Other Types of Discourse Semantics

Given the proliferation of computer-mediated communication utilizing such data is becoming important also in the development and research of question-answering systems. Previously, successful question-answering systems had a similar “pipeline structure” (Feng et al., 2006a, p. 172), a structure that combines a natural language parser, information retrieval engine, and information extraction component. Lin (2007), however, questions if the NLP components are actually adding value to the IR engine. Some researchers worked on developing a corpus-based system that automatically generates responses for online communication such as a discussion board (Feng et al., 2006a) and a help desk service (Marom and Zukerman, 2009). Given the advancement of data storage technologies, the corpus-based approach has the potential to provide a better framework. Researchers have developed systems that automatically detect questions and answers in online documents such as email threads (Shrestha and McKeown, 2004), community QAs (Jeon et al., 2005), and online forums (Hong and Davison, 2009; Ding et al., 2009; Hong and Davison, 2009).

Shrestha and McKeown (2004) developed a machine learning system to detect question-answer pairs in e-mail threads using a rule-induction system. The system detected the question-answer pairs with relatively good results (precision: 69.8%, recall: 61.9%).

Jeon et al. (2005) propose statistical methods (language models) to find questions that are semantically similar in the community-based question answering service of Naver (Korean portal site). Their approach uses the similarity ranking of answer pairs to estimate the similarities of the corresponding question pairs. Their assumption is that two questions may be lexically different even if they are meant to have the same meaning. Their experimental results indicated they can indeed automatically find semantically similar questions by measuring similarities between answers. The top ten results from the two methods they proposed showed accuracies of 80% and 90%.

Cong et al. (2008) developed a system that detects questions and corresponding answers in online forums. The detection of questions is done by a rule-based classifier, Ripper, using labeled sequential patterns as features. A graph based propagation method was used to detect answers for questions in the same thread. Experimental results show that the techniques are very promising. The precision and recall for the question detection task were 97.8% and 97.0% respectively, at the highest, and the precision at the first position for the answer detection task was 66.5%.

Ding et al. (2009) used Conditional Random Fields, and Hong and Davison (2009) used SVM for similar tasks, and achieved similar results.

2.4.5 Summary

This section discussed previous studies that applied machine learning methods to dialogue act annotation. Developing systems that analyze the semantics of online communication, is increasingly important, and analyzing dialogue acts of messages is one such approach, based on the discourse linguistic theory, Speech Act Theory (Austin, 1975; Searle, 1969). Researchers have developed systems that detect dialogue acts of different discourse genres such as telephone conversations (Stolcke et al., 2000; Lan et al., 2008), multi-party meetings (Clark and Popescu-Belis, 2004; Galley et al., 2004), online discussion boards (Kim et al., 2006; Feng et al., 2006b), or emails (Cohen et al., 2004; Carvalho and Cohen, 2006; Hu et al., 2009). Given the nature of reference interviews, automatic identification of dialogue acts has promise to be beneficial not only for understanding of the nature of information-seeking interactions but also for development of systems that extract and/or retrieve information that is embedded in information-seeking interactions.

The previous studies show that HMM and SVM are reasonable choices as the algorithms for such experiments. The studies also show a variety of features that may be used for learning dialogue acts.

Some researchers have worked on developing corpus-based systems that automatically generate responses for online communication, such as in a discussion board (Feng et al., 2006a), and a help desk service (Marom and Zukerman, 2009). Given the advancement of data storage technologies, the corpus-based approach has potential to provide a better framework. Other researchers have developed systems that automatically detect questions and answers in online documents such as email threads (Shrestha and McKeown, 2004), community QAs (Jeon et al., 2005), and online forums (Hong and Davison, 2009; Ding et al., 2009; Hong and Davison, 2009). This approach, however, is not applicable to cases where the information need is complex and questions and answers are addressed through multiple exchanges of messages. Thus, a more general approach to investigate the discourse of online reference communications is needed.

Chapter 3

Method of Investigation

3.1 Overview

This study consisted of two stages:

1) Discourse analysis, wherein a dataset was annotated by trained annotators and analyzed by the researcher, and

2) Machine learning experimentation, wherein the annotated data was used to train and test machine learning systems to perform the same annotation task done manually by the annotators.

Discourse analysis was chosen as the first method of investigation because it allowed researchers to analyze the interactions with minimum interference between the librarians and the reference service users in the process. While the term discourse analysis encompasses various methods and theories across disciplines, the study employed dialogue act analysis, a method of analyzing communicative functions of linguistic interactions, based on the speech act theory (Austin, 1975; Searle, 1969). The dynamic interpretation theory, proposed by Bunt (1994), provided the theoretical motivation of the method.

The coding was done by three annotators, who were all MLIS students. Standard statistical measures were used to evaluate the reliability of annotation. The annotated data was then analyzed for three aspects: 1) the distribution of dialogue acts, 2) linguistic forms of text segments, and 3) transitions (sequences) of dialogue acts. The observations were compared with the "classic" theories and models of information-seeking behaviors such as the ASK Hypothesis (Belkin et al., 1982) and the Berrypicking model (Bates, 1989), as well as more recent developments such as micro-level information seeking processes (Wu, 2005) and exploratory search (Marchionini, 2006).

In addition to the evaluation and analysis, the annotated data were used to train machine-learning systems to automatically recognize the dialogue act of a given text segment. Through the experiments, the researcher investigated the suitability of each of the machinelearning algorithms for the automatic annotation of dialogue acts (DAs), as well as the usefulness of the defferent linguistic evidence for machine learning. The goal of these machine learning experiments was to provide proof of concept, confirming that there was linguistic evidence to identify the dialogue acts and that this linguistic evidence could be automatically learned by following certain procedures (algorithms).

The experiments produced promising results, indicating that dialogue act analysis has potential for practical applications in futher research. For example, it can provide: 1) a new measurement for evaluating virtual reference services, 2) new data attributes for information extraction / retrieval algorithms, and 3) a prototypical dialogue model for constructing fully-automated dialogue systems.

3.2 Data

3.2.1 Overview

The data for this study was an archive of reference interviews at QuestionPoint, the virtual cooperative reference service provided by the Online Computer Library Center (OCLC). QuestionPoint provides a 24/7 digital reference service through chat and e-mail management systems, with staffing from over 200 participating libraries worldwide.

Two data sets were provided by OCLC. The first dataset was collected from December

2005 to December 2006, and consisted of 500 interview sessions. The second dataset provided was originally collected from July 2004 to June 2005 (before the first dataset) and consisted of 293[1] interview sessions. Collectively, the data consisted of over 12,000 chat messages between librarians and users. These data were originally prepared for a research project contacted by Radford and Connaway (2005) and became available to the researcher, courtesy of Dr. Radford, Dr. Connaway, and OCLC.

Any information that might identify the participants of the reference sessions, e.g. e-mail addresses, names, and phone numbers, had been replaced by a place holder with a general descriptor, such as *[Patron name]*, *[Librarian e-mail address]*, prior to the release of the data by OCLC.

As part of primary observations, simple descriptive statistics (frequency counting) were used in order to obtain a general picture of the data. Table 3.1 shows the results and illustrates how different the nature of interactions in digital reference interviews is from other forms of information-seeking interactions, e.g. interactions between an internet user and a web search engine. One of the major difficulties for the development of any IR system is determination of the information need from limited user input. While, over time, web search engine queries have been getting longer (as shown in Tatham (2009)), it is known that only one to three terms are used most often for web search engine queries (Silverstein and Henzinger, 1999; Spink et al., 2002).[2] In contrast, the average number of terms in a single online chat reference session in this data was over 200, and the average number of terms used in a users' first message of a reference session (, which typically describes the users' information need) was approximately twenty. These numbers indicate how information interactions at the digital reference service have potential to provide richer semantics than interactions between web search engines and users.

[1] 300 were provided, but seven were ill-formed and could not be processed.

[2] According to a blog post (Ussery, 2008), the average number of Google query terms is up to four.

Number of sessions:	793
Messages in total:	12,634
Messages in each session on average:	15.9
Terms used in total:	167,515
Terms used in each session on average:	209.4
Terms used in each message on average:	13.3
Terms used in the first messages:	16,000
Terms used in the first message in each session on average:	20.2

Table 3.1: Numbers in the Chat Reference Transcripts

3.2.2 Rationale

In this section, the rationale behind choosing online chat reference transcripts as the data for the study is explained.

First, the researcher expects that online chat reference sessions are purposefully structured by the librarians to solve the information needs of the users in the same way that the face-to-face reference sessions do (Taylor, 1968). Thus the data was expected to reveal the nature of the information-seeking process in an ideal form and, by providing an optimal model of interaction, therefore useful inn the design of information systems. This is an assumption of many previous studies that examined reference interviews for design of new IR systems (Belkin et al., 1983; Daniels et al., 1985; Brooks, 1986; Belkin et al., 1987b; Wu, 1993; Spink et al., 1995; Saracevic et al., 1997; Wu and Liu, 2003; Wu, 2005). In addition, the online chat reference interviews are most often performed by two people – a librarian and a user – and thus are less complicated in terms of the structure of interactions, and less challenging for interpreting and coding the data compared to many other forms of online information seeking such as community-based question answering or online forums that involve multiple participants.

Second, in chat reference, the interactions happen through the digital environment, namely through a web browser, which makes the interchanges more comparable to other information-seeking exchanges in the digital environment, such as use of web search engines. While interactions between a search engine and its users are often limited to exchanges of queries (keywords) and search results (mostly URLs), many researchers (Silverstein and Henzinger, 1999; Lau and Horvitz, 1999; Spink et al., 2002; Beitzel et al., 2004; Teevan et al., 2008) have analyzed the interactions as an information-seeking process. Given that chat reference services take place through exchanges of texts in an online environment, observations of online chat reference transcripts invite an interesting comparison not only to the findings of the previous studies of the information-seeking behavior where face-to-face reference sessions, but also to studies that are based on the use of web search engines.

Third, online chat transcripts are near-complete reproductions of actual interactions. Texts available to the researcher for analysis are exactly the same texts as the conversation participants actually exchanged, which is not the case for transcripts of face-to-face reference sessions, where para-linguistic aspects of the interaction (e.g. tones, gestures, or eye movements) are always lost. Online chat transcripts are stored automatically and unobtrusively (with the user's permission), so it is reasonable to assume that the data collection process has little effect on the users' behavior, which is often not the case if face-to-face reference sessions are recorded.

The characteristics of online chat reference interviews described above made the data ideal for the method of the study and its purpose.

3.2.3 Format

The data were originally provided to the researcher in Microsoft Word document format. The document files were first exported to plain text format and then imported into a MySQL database, using a set of Perl scripts. The database was constructed using two database tables: 1) reference session, which consisted of metadata (date and time entered, institution referred by, institution assigned to, wait time, session time) and the content, which is a series of messages, and 2) message, which consisted of metadata (date and time the message was sent and sender) and the content, which is a text message. As Table 3.1 shows, the dataset for the study consisted of 793 interview sessions, which included 12,634 messages in total. Figure 3.1 shows the structure of the interview and message data as initially imported.

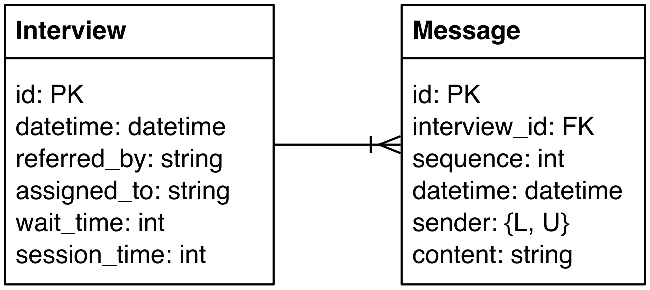


Figure 3.1: Structure of interview and message data

3.2.4 Preprocessing

Through initial observations of the data, it became clear to the researcher that a substantial portion of the data was not applicable to the analysis. Sometimes, it was due to the quality of the conversation (e.g. inappropriate questions, questions that were too simplistic, test sessions), and other times it was because of the quantity of text in the messages (i.e. too little text was exchanged in an interview). Thus, systematic preprocessing was needed in order to make the later linguistic analysis efficient and meaningful. After importing the original dataset to a MySQL database, the following three stages of preprocessing were applied:

1. **Message filtering**, wherein messages were automatically labeled with message types using simple rules (regular expressions).

2. **Interview filtering**, wherein interviews that were determined to be to be inapplicable to the analysis were set aside, based on the metadata and the labels (applied in the previous step).

3. **Classification of reference questions**, wherein the user's question or information need in each interview session was classified.

In order to avoid producing biases while excluding these data, the preprocesses were done systematically and automatically, as much as possible. As such, the original data were preserved and the processes were kept reversible or repeatable for different applications (e.g. the arrival of new data). The following paragraphs provide more descriptions of the steps

Message content	**Description**
[Page sent]	A indication that the librarian sent an URL to the user through a co-browsing feature of the system. (The URLs are saved as text messages.)
Chat session ended	Often, but not always, inserted at the end of an chat session.
Note to staff: ...	Notes that often describe e-mail or face-to-face interactions that followed-up the chat session.
Set Resolution: ...	Used occasionally (136 sessions out of the 500 interviews in the first dataset) to indicate the status of the question.

Table 3.2: Examples of messages excluded from analysis

above.

The following are the types of messages or interviews that were excluded from the analysis. Each type is futher described below.

- Messages that were generated only for the purpose of record.
- Interviews with extremely few message exchanges.
- Test or training interviews by the staff at OCLC.

3.2.4.1 Message Filtering

In the original data, there were messages that were not generated either by the librarian or by the user, but were inserted for the purpose of record. These messages were considered not applicable to the analysis since the texts did not represent messages that were exchanged during the interviews and were not seen by librarians or users. These messages were labeled and set aside from the analysis. Examples of these types of messages are shown in Table 3.2.

3.2.4.2 Message Labeling

While considering the exclusion of some of the messages as described above, other types of messages or interviews caught the eye of the researcher due to their distinctive nature. While the messages above were identified and set aside, other messages were labeled and kept for the analysis. The following are the types of messages that were labeled:

- Scripts

  Librarians often utilize prescribed scripts to generate frequently used phrases. These messages are often used to open or a close an interview, but, in some occasions, are used during a question negotiation. For example:

> *[Hello and welcome to Homework Help. I'm just reading your question...]*
>
> *[If you need further assistance, please feel free to contact us again. Thank you for using 24x7 Ask! We hope you will use our service often!]*
>
> *[Can you tell me where you have searched, so I don't duplicate your search?]*

Such messages were labeled as *Scripts*.

- Practice/Test Questions

  The data included interview sessions for testing or practice purposes by the OCLC staff or librarians. In such cases, the initial messages were often (but not always) identified with the term *[Practice]* or *[Test]*. Messages with such a tag were labeled as *Practice*.

- Greeting

  The most common pattern of prematurely terminated interviews was a simple exchange of greetings. In order to detect such simple conversations, greeting utterances, such as *Hi [Patron Name]* and *Hello?* were captured by regular expressions and labeled as *Greeting*.

- Default

  All the other messages were labeled as *Default*. This label was used to best estimate the numbers of messages that were actual exchanges by a librarian and a user in an interview.

3.2.4.3 Interview Filtering

Some interview sessions had very few message exchanges, involving no clarifying or negotiation processes, or provision of information. While there might be various reasons for such cases, e.g. technical difficulties, intentions of users, miscommunication, long waiting time, etc., the study needed to exclude these interviews from the analysis as they provided little information as to how information-seeking dialogues proceed. In order to do so, the following steps were taken:

1. The number of messages in each interview was counted and recorded.
2. The number of turn takings in each interview was counted and recorded. [3]
3. Interviews with the following conditions were excluded from the analysis: 1) the number of *default* messages were less than 2; or 2) the number of turn takings was less than 3.

The thresholds were kept very low, since false negatives were preferred over false positives – the conditions automatically excluded only interviews that were most likely to be unproductive for the study, while leaving the data near the boundary to be manually inspected and excluded at the researcher's discretion.

3.2.4.4 Classification of Reference Questions

While the main focus of the study was to investigate the structure of question negotiations, reference interviews involve all types of questions, which might not necessarily require negotiation processes (e.g. *"What is alliumphobia?"*). Thus, prior to the main analysis, the interviews in the data were coded with the types of reference questions. Classifying questions for reference services is a well-established practice among both researchers and practitioners, and a number of research articles (Arnold and Kaske, 2005; Radford and Connaway, 2007a) and reference service guidelines (Katz, 2002; Bopp and Smith, 2001) have suggested classification schemes. The classification of reference questions used in this study was based on the classification scheme used by Radford and Connaway (2007a), which was, in turn, based on Katz (2002) and Arnold and Kaske (2005). Table 3.3 shows the classification scheme. This classification was done solely by the researcher, with occasional consultations with reference librarians, when the researcher was not confident of the decision.

[3] The number of turn takings could not be calculated from the number of messages, since in online chat, one can send multiple messages before the other person responds or interrupts.

Classification	Description
Ready Reference	Typical, uncomplicated questions that require simple lookup of reference materials.
Research/Subject Search	Questions about a certain topic or subject, that require question negotiations.
Procedural/Policy	Questions regarding policies or procedures within a library.
Holding/Reader's Advisory	Questions regarding the collection of a library or asking for reader's advisory.
Location/Directional	Questions asking the location of or the direction to a library or a location within a library.
Inappropriate/Unknown	Questions regarding personal information of the librarian, pranks, etc., that are not appropriate for the reference service or questions of which the intention or the purpose is not clear.
Legal	Questions that involve legal process or status, such as filing legal forms or documents, or taking legal actions in situations.

Table 3.3: Classification of Reference Questions

3.3 Dialogue Act Analysis

3.3.1 Overview

After the preprocessing described in the previous section, the data were coded by human annotators using a coding scheme and a coding environment that had been developed prior to the proposal of the study (described in Section 3.3.2 and 3.3.4). Three annotators (MLIS students) collectively annotated 210 online reference interview sessions, which, in total, consisted of 5489 messages between a librarian and a user. Three annotators (Annotator 1, 2 and 3a) were hired initially, but Annotator 3a withdrew after the second week of training, and another annotator (Annotator 3) was hired as a replacement. During the initial two week period, the researcher explained the purpose of the research, annotation scheme, and annotation environment to the annotators, and the annotators went through several practice coding sessions. After three months, Annotator 2 left the project for graduation and employment, but since about two thirds of the data were annotated at that point, no replacement was hired. Thus, only Annotator 1 was involved from the beginning to the end of the process – Annotator 3 was absent for the first three weeks and Annotator 2 was absent for the second half of the annotation process (about 3 months). These absences did not affect the completion of annotation over the prepared data, but Annotator 2 annotated less data than the other two and Annotator 3 had less overlap with other two than Annotator 1 and 2 (described further below).

During the main annotation process (March to August), the researcher and the annotators met every week to discuss problems that annotators faced during the annotation in the past week. Suggestions were often made by the reseacher or the annotators, and based on the suggestions, the coding scheme and coding environment were modified. When it was necessary, annotators went back to the old annotated data to reflect new agreement. The annotators' work was also monitored in terms of the duration of the work, as well as the number of interviews, messages, and text segments they annotated. The annotators were assigned to a new set of reference interviews every week (See Section 4.1.2 for more details.). The annotators manually segmented each message in a reference interview in into one or more text segments and labeled each text segment with one or more dialogue acts. Each dialogue act consists of two labels: a function and a domain. Functions represent what utterances are intended to perform, e.g. providing information or conforming to social obligations. Domains represent more detailed descriptions of the dialogue acts by specifying which part of cognitive states the function operates upon, e.g. what kind of information is exchanged, which aspect of social relationship is dealt with by the utterance, etc.

Annotators 1 and 3 went through a four-week period for additional brush-up, from October to November 2011. In that period, each annotator revisited their own work to make sure there were no errors in the annotation and to add details that had been overlooked.

3.3.2 Annotation Scheme

3.3.2.1 Theoretical Framework

The annotation scheme was developed based on the Dynamic Interpretation Theory referenced in Section 2.3.2.3 (For the complete list of annotation labels, see Tables A.1 and 5.2 in the Appendices; for examples of coding, see Table A.5.). The theory was suited to the study because it enabled the classification of reference interview messages with a conceptual unit (dialogue act), and their organization based on certain aspects of their dialogue acts (functions and domains). In addition, the notion of underlying tasks and communicative tasks aligned in with observations from previous reference studies such as Radford (1993, 2006b), which showed how the communicative task is carried out during chat reference, and with studies in information-seeking behavior, such as the studies by Belkin et al. (1983) and Wu (2005), which analyzed the information exchanges during information seeking. Figure 3.2 illustrates the relationship between the duality of conversation that the Dynamic Interpretation Theory hypothesizes and aspects of information-seeking behavior that different researchers investigated.[4]

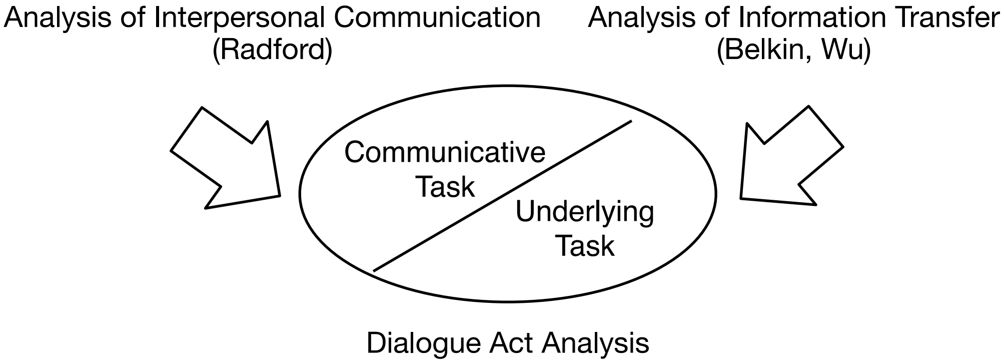


Figure 3.2: Dialogue act analysis and library and information science studies

3.3.2.2 Dimensions of Analysis: Function and Domain

In this study, dialogue acts were analyzed along two dimensions: *function* and *domain*, based on the notions of *function* and *context* of the Dynamic Interpretation Theory. According to the theory, dialogue acts are transformations of participants' cognitive states, which is called *context*. A function, in a formal definition, is a relationship (or mapping) between two sets of elements, where for each element in one set (called *domain*) as an input, an element from another set (called *codomain*) is specified as an output. Thus, a dialogue act can be

[4] Radford's studies were based on the interactional theory (Watzlawick et al., 1967), which stated the similar duality of communication as the dynamic interpretation theory.

considered a function of which domain and codomain are the cognitive states of participants (Figure 3.3 illustrates this analogy using an example of a dialogue act function and other kinds of formal functions.). With this regard, dialogue acts can be categorized, in principle, in two ways, as categories of the function, which represent what the dialogue act does to the cognitive state, and as categories of the domain, which represent which cognitive states the dialogue act operates on. The latter categorization assumes that human cognitive states can be conceptually divided into multiple spaces, which Bunt (1994) calls *local contexts*. Bunt proposed several such local contexts, such as *local semantic context* and *local social context* (, which roughly correspond to the *information domain* and the *social domain* in this study), but since Bunt's notion of contexts includes other aspects of dialogue acts in this study, the term *domain* is used instead (See Sections 2.3.2.3, 2.3.2.4 and 2.3.2.5 for more on Bunt's notion of dialogue acts.).

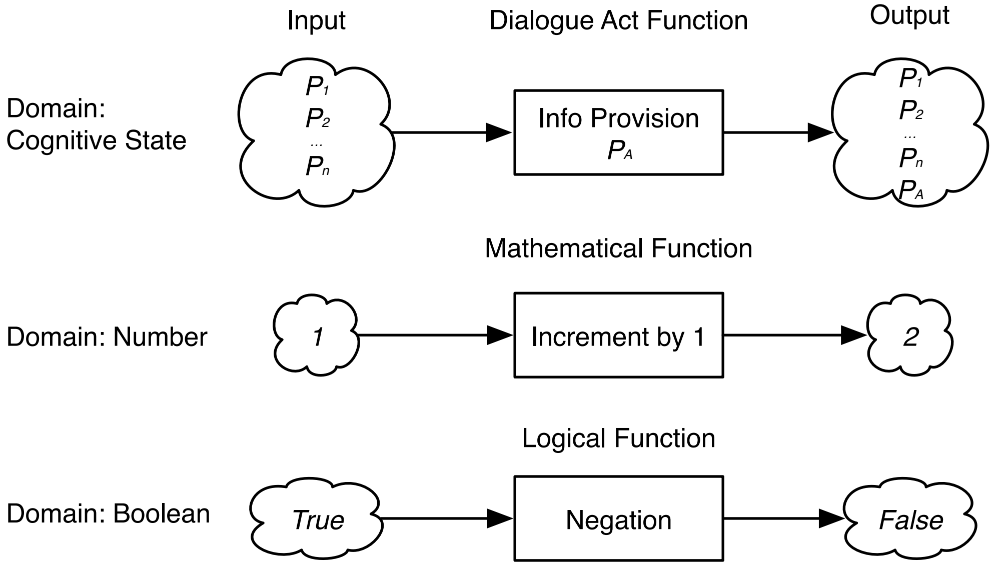


Figure 3.3: Examples of Functions and Domains
A cognitive state is represented as a set of predicates P_x.

The annotation scheme was developed with this principle in mind, separating the categories of the function and the categories of the domain as independent dimensions. However, the annotation scheme was further developed as the coding was performed by the annotators and the two dimensions appeared to be more intertwined than the original design. In the end, the domain categories were used as the detail-level categories of the function categories, except for the categories of the information domains and the categories of information transfer functions – they remained as independent dimensions. For this reason, combining function categories and domain categories (or subcategories, described in the later sections) did not increase the number of possible classes by a simple multiplication, since many of the combinations were not possible (e.g. a combination of *Info Provision* and *Social:Greeting*). For example, there were seven function categories and twenty domain categories, while combining the function categories and the domain categories yielded only 25 possible classes. Table 3.4 summarizes the relationship between the level of analysis and the number of possible classes. The following sections describe the categories of the functions and domains in detail.

Level of Analysis	**Number of Classes**
Function	5
Domain	18
Function + Domain	23
Domain Subcategory	41
Function + Domain Subcategory	70

Table 3.4: Level of Analysis and Number of Classes

3.3.2.3 Categories of Dialogue Act Functions

In reference interviews, librarians and users are engaged in two kinds of dialogue acts that motivate the communication: 1) *Information Transfer* (e.g. asking clarifying questions or describing of the information need) and 2) *Task Management* for performing information-seeking activities (e.g. looking up a catalogue or examining information objects). And in order to maintain the communication to realize these functions, librarians and users are engaged in two other kinds of dialogue acts: 3) *Social Relationship Management*, which deals with the socio-emotional aspects of the communication (e.g. greeting, thanking) and 4) *Communication Management*, which deals with the physical aspects of the communication (e.g. pausing the dialogue, checking the communication channel). These four fundamental categories of dialogue acts, called dialogue act functions in this document, were adopted from

Function	Task Level	Description
Info Provision	Underlying	To provide information
Info Request	Underlying	To request information
Task Mgmt	Underlying	To assign or commit to tasks
Social Rel Mgmt	Communicative	To manage socio-emotional aspects of communication
Comm Mgmt	Communicative	To manage physical aspects of communication

Table 3.5: Dialogue Act Functions

information transfer functions, *action discussion functions*, *social obligation management functions*, and *interaction management functions*, respectively, from DIT++ proposed by Bunt (2000). (DIT++ is described in Section 2.3.2.3.).

In order to enable in-depth analysis of information exchanged in dialogues, the information transfer function was further divided into two subcategories: *Information Provision* and *Information Request*. Table 3.5 summarizes the categories of functions.

3.3.2.4 Categories of Dialogue Act Domains

As described in Section 3.3.2.2, the coding scheme for the study incorporated the notion of context as the domain of the functions to further identify dialogue acts, following Bunt (1999). While Bunt's theory suggests five types of contexts (linguistic, social, cognitive, semantic, and physical), this study identifies four types of domains: 1) *Information*, 2) *Task*, 3) *Social Relationship*, and 4) *Communication*, each of which corresponds to one of the four types of functions described above (*Information Provision* and *Information Request* share the *Information* domain.). Previous analyses of information-seeking processes (Belkin et al., 1983; Wu, 2005) were incorporated for analyzing the *Information* domain, while reference studies that analyzed the socio-emotional aspects, such as Dewdney and Ross (1994), Radford et al. (1999) and especially Radford (2006b), were incorporated for defining the categories for the *Social Relationship* domain and the *Communication* domain.

3.3.2.5 Categories of the Information Domain

The following were identified as categories of the *Information* domain:

Domain	Description
Info Problem	Description of the user's problem or information need, e.g. the topic of information, the background of the information need, the use of the information required, etc.
Search Process	Description of the search process and related matters, e.g. search strategies, previous search experiences, the on-going search tasks.
Info Object	Description of a particular information object, e.g. physical description, location, reference information (title, author, publisher), URL, overall impression, etc.
Feedback	Feedback for information objects or other types of information.
Other	Other information such as the user's e-mail address, or library's location, etc.

Table 3.6: Categories of the Information Domain

1. *Problem*: description of the user's problem or information need,
2. *Search Process*: description of the search process and related matters, such as a search strategy,
3. *Information Object*: description of a particular information object,
4. *Feedback*: feedback for information objects or other dialogue acts, and
5. *Other information*: other information such as the user's e-mail address, and the library's location.

As shown in Table 5.2 in the Appendices, these categories were further divided into subcategories in order to enable more detailed analysis. For example, the *Info:Object* category was further divided into: *Info:Object:Description*, *Info:Object:Excerpt*, etc. All the annotation was done at the subcategory-level. The analysis, however, was done at the higher, category-level, because of the low inter-coder agreement at the subcategory level annotation (See Section 4.1.3 for the description.). The subcategories provided the detailed descriptions of the codes and made the coding tasks easier. So even though the inter-coder agreements at the subcategory-level annotation were lower than the satisfactory level, they were beneficial for the annotators to complete their tasks (The evaluation of the annotation is discussed in Section 4.1.3.).

Categories of the Task Domain

The *Task Management* domain was categorized into three: *Librarian's task*, *User's task* and *Other's task*. The description of each task is provided in Table 3.7.

Categories of the Social Relationship Management Domain

The categories of the *Social Relationship Management* domain were defined based on the categories of social obligations found in DIT++ (Bunt, 2000) and classes of relational facilitators defined by Radford (2006b). A challenge in identifying categories in this domain was that the domain encompasses various aspects of linguistic behavior such as sociological, cultural, emotional and psychological aspects, which are intertwined, and thus fixing the analytical dimension was difficult. In the end, the coding scheme included the most commonly-recognized types of utterances that are related to the social obligations in the data: *Apology*, *Downplay*, *Exclamation*, *Gratitude*, *Greeting*, *Rapport Building* and *Valediction* (See Table 3.8 for details.).

Categories of the Communication Management Domain

The categories of the *Communication Management* domain were defined based on the categories of the dialogue management functions and the interaction management functions found in DIT++ (Bunt, 2000). Among various subcategories that Bunt defines, three categories were identified in the data: *Pausing*, which corresponds to the time management function in Bunt's definition, *Channel Checking*, which corresponds to the contact management function, and *Feedback*, which corresponds to the auto-feedback function (See Table 3.9 for details.).

3.3.3 Annotation Procedures

During the main annotation process, the annotators manually segmented each message into one or more text segments and labeled each text segment with one or more dialogue acts.

Domain	Description
Librarian's Task	Description of the librarian's task. If provided by a librarian him/herself, the utterance is often a commitment of a certain task or explanation of the plan of action, e.g. "I'm going to do a search to see if I can find anything good". If provided by a user, the utterance is a suggestion or a question about the direction that the librarian could take, e.g. "but if any more information is found, may please be emailed?"
User's Task	Description of the user's task. If provided by a librarian (i.e. Info Provision + Task: User) it often is a suggestion, instruction, or directive, e.g. "Now scroll down to World Book Online and click on it..." If provided by a user, it is often a description of the user's future action, e.g. "i'll browse through them".
Other	Tasks for someone other than the librarian or the user in the conversation, e.g. "Someone from UW will respond to you by email."

Table 3.7: Categories of the Task Domain

Domain	Description
Greeting	The sender informs the presence of him/herself to the receiver, e.g. "Hello.", "Hi [User Name]."
Valediction	The sender informs that he/she is ready to close the dialogue, e.g. "Bye."
Exclamation	The sender expresses surprise, confusion, and other emotional / psychological remarks, e.g. "hmm", "wow!"
Apology	The sender informs his/her regret of his/her failure or offense to the receiver's experience, e.g. "I'm sorry."
Gratitude	The sender informs that he/she is thankful to the receiver, e.g. "Thank you."
Downplay	The sender acknowledges an apology or a gratitude, e.g. "You are welcome."
Closing Ritual	The sender implies that he/she is ready to close the dialogue, e.g. "Thank you for using Maryland AskUsNow! If you have any further questions, please contact us again."
Rapport Building	Other rapport building, such as humor, praise and encouraging remarks, e.g. ":-)", "God bless you!", "you've been great!"

Table 3.8: Categories of the Social Relationship Management Domain

Domain	Description
Channel Checking	Checking the communication channel, e.g. "Hello?" "Are you still there?", "Can you stay online?"
Pausing	Informing the receiver that communication is being paused or stalled, e.g. "Hold on...", "Just a sec."
Feedback	Informing the receiver that his/her previous message was received and understood, e.g. "Ok."

Table 3.9: Categories of the Communication Management Domain

For each message in each reference session, annotators identified dialogue acts recognized in the message and specified the text segment associated with each dialogue act. Multiple dialogue acts might be associated with the same text segment, but each text segment must not overlap with other text segments. Figure 3.4 shows a simplified view of the data structure of annotated data.

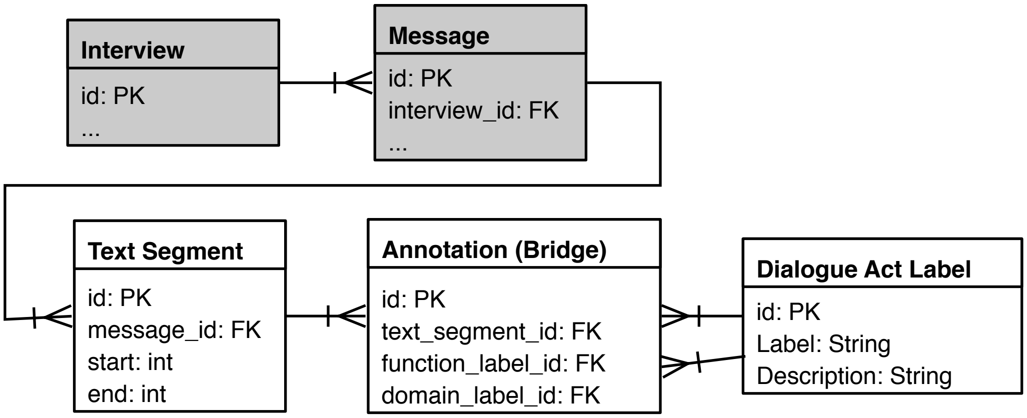


Figure 3.4: Structure of data and annotation
The gray containers represent original data while white containers represent annotations.

The following are brief descriptions of the steps that each annotator took to complete the annotation task.

1. Launch a web browser and go to http://sanka.syr.edu:8080/cae. Log in to the system using the user name and password provided by the researcher.

2. Click “Assignment” from the main menu and select a dataset.

3. Select an interview from a list of assigned interviews.

4. Once the interview data is displayed, repeat the following steps for each message:

   (a) Identify dialogue acts expressed in the message. There may be multiple dialogue acts. (At this point, annotators do not assign the labels to the text yet.)

   (b) Divide the message into text segments based on the dialogue acts. Each text segment is specified by selecting the first word and the last word of the text segment.

(c) For each text segment, label the dialogue act, by specifying the category of the function and the domain.

5. Once all the messages are annotated, go back to the list of assigned interviews, and mark the interview as "completed".

Each instance of a dialogue act corresponded with one text segment, which must be a grammatical and interpretable utterance by itself. When needed, multiple dialogue acts could be specified for a single text segment. Specification of a dialogue act consisted of selecting one of the function categories and selecting one of the domain labels. For example, the text segment below is associated with two dialogue acts: an information request for the user's e-mail address and an information provision about the system's functionality (labeled as *Info Request / Info:Other:User* and *Info Provision / Info:Other:System*):

L: *If you would like a transcript of this session emailed to you, please type your full email address now.*

If a single dialogue act was divided into two or more messages, the subsequent messages were labeled as a continuation of the earlier message. This was done by an additional label called *adjacency relations*, based on the notion of *adjacency pairs* (Goffman, 1981) in discourse linguistics.[5] The following are examples of such messages, where two messages constitute one dialogue act, rapport building (*Social Mgmt / Social:Rapport*):

U: *You are the*

U: *BEST!*

The identification of the text segments and dialogue acts were mutually dependent, so annotators might need to go back-and-forth between the steps during the coding. This is due to the fact that dialogue acts, by definition, are not reducible to syntactic structure. If so, text segments could have been identified by some syntactic parsing mechanism. Two

[5] In discourse linguistics, an adjacency pair is a pair (or sometimes sequence) of utterances where the former (or the first) utterance expects a certain types of responses (or sub-sequences). Examples are a question and an answer, an apology and a downplay, and a greeting and a response. In the analysis, the notion was extended to recognize the common patterns across the utterances, such as paraphrasing, clarifying and correction. However, theorization of this notion is out of scope for this study and this label was used only to recognize dialogue acts that went across multiple messages in the document.

simpler alternative approaches to this issue were considered: 1) using sentences as the unit of dialogue acts, and 2) using messages as the unit of dialogue acts. Neither of the approaches, however, were determined to be adequate, because mismatches between the unit and dialogue act were observed in a substantial number of cases during the pilot study. Thus, while the current approach added another level of complexity to the study, it was considered necessary.

3.3.4 Annotation Environment

The annotation environment was developed using Java, with general web/database application development tools such as MySQL, Hibernate, and Spring MVC. The software environment allowed annotators to perform their annotation task through a web browser, with the following functionality:

Features for Annotation

- Browsing the data at multiple levels (data sets, sessions, and messages).
- Dividing a message into text segments.
- Labeling a text segment with a dialogue act.
- Labeling an interview session for a classification of reference questions.
- Browsing the list of annotation labels and their definitions (coding book).
- Searching previously annotated data with a specified annotation label or content.
- Editing the coding book, by adding a new code or removing a code, editing descriptions, and changing the labels.
- Adding or deleting annotation schemes, dimensions, and annotation labels.

Features for Managing Annotators and Annotation Tasks

- Adding and deleting annotators.
- Assigning annotators to interview sessions.

- Monitoring the progress of each annotation task.
- Monitoring the workload of each annotator.
- Browsing and comparing multiple annotators' coding.
- Exporting confusion matrices and agreement measures between annotators.
- Exporting data for machine learning.

The software was hosted by the Syracuse University iSchool, using their virtual machine clusters.

3.3.5 Evaluation

3.3.5.1 Overview

As described earlier, the coding task for this study involved two distinct types of subtasks: 1) segmenting texts based on dialogue acts (text segmentation) and 2) annotating dialogue acts (content analysis). While these two subtasks are not independent in this study, each of them has different challenges in evaluating outcomes and is associated with a separate set of evaluation methods. Thus, in this study, the evaluation of the discourse analysis stage was performed in two separate stages: 1) the text segmentation task was evaluated using *WindowDiff* (Pevzner and Hearst, 2002); and 2) the content analysis was evaluated using Cohen's *kappa* (Cohen, 1960).

3.3.5.2 Evaluation of Text Segmentation

Text segmentation is a task of partitioning a stream of text into one or more segments, based on some linguistic unit. The task is used in qualitative data analysis, such as content analysis of interview data or other text data, as well as various subfields of human language technologies including speech and dialogue, text summarization, question answering, and information retrieval.

A difficulty in evaluating agreement of text segmentation by multiple coders (or human coding and a machine learning system) comes from the fact tha the outcomes of tasks are not simple nominal or numerical values, but a set of text segments. Each set of text segments may be represented a set of pairs of numbers, each of which represents the starting and ending position of a text segment, or simply by a set of numbers, each of which represents the position of segment boundaries.

One of the simplest ways to evaluate these values is to reduce them to simple numerical values, such as the size of each set (the number of text segments in each instance) or the difference between the two numbers of each pair (the length of each text segment) as seen in Kurasaki (2000). Another way to evaluate the values is to use precision and recall, the common metrics for the information retrieval research. But such metrics obviously lack accuracy, because of their inability to recognize the structure (e.g. position and length) of text segments. Another challenge in evaluating text segmentation tasks is in how to recognize close misses (i.e. segment boundaries that are off only by one or a few words). It is natural to expect that close misses should be evaluated more positively than far misses, and exact matches should be evaluated more positively than close misses.

The P_k evaluation metric, proposed by Beeferman et al. (1999) was the first standard metric for the text segmentation task, attempting to overcome the difficulties mentioned above. Pevzner and Hearst (2002), however, claims the following as shortcomings of the method: 1) False negatives are penalized more than false positives; 2) The number of boundaries are ignored; 3) False boundaries are over- or under- penalized depending on the segment size; 4) Close errors are over-penalized; and 5) The scale of the score is not clear. Based on these observations, Pevzner and Hearst (2002) proposed an amended metric, called *WindowDiff*. *WindowDiff* is defined as follows:

$$WindowDiff(ref, hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|b_{ref}(i,i+k) - b_{hyp}(i,i+k)| > 0) \tag{3.1}$$

where:

ref and hyp represent two segmentations; [6]

N is the number of segments in the text;

$b(i, j)$ is the number of boundaries between positions i and j in the text; and

k is the "window" size representing the tolerance for near misses.

In Beeferman et al. (1999), k is set to the average segment length of the reference segmentation (in words). But in this study, it was set to the average of the segment lengths of both segmentations, since neither of the segmentations are considered as a reference nor a hypothesis.

The symmetry of the equation solves all of the shortcomings listed above, except 2) and 5). As for 2), Pevzner and Hearst (2002) argues that the outcomes of their evaluation did not change if they weighted the difference in the number of boundaries (specifically, using $|b_{ref}(i, i+k) - b_{hyp}(i, i+k)|$ instead of $|b_{ref}(i, i+k) - b_{hyp}(i, i+k)| > 0$). As for 5), it is clear from the equation that the scale represents the ratio of miss-matched boundaries at each point of the text with a certain tolerance.

3.3.5.3 Evaluation of Dialogue Act Annotations

When researchers started applying discourse linguistics to computational linguistics in the 1980s (e.g. Grosz and Sidner (1986)), analysis or annotations in many such studies depended on the researchers' interpretation and individual judgments. Such studies, thus, lacked generalizability and repeatability, making comparison between results from two studies difficult.

Carletta (1996) suggested that Cohen's *kappa* statistics (Cohen, 1960) could be used to measure the reliability of annotation objectively and for comparing different annotation studies in a standard manner, adopting the idea from the field of content analysis. Today, the validity of a coding scheme for a discourse annotation study is most-often evaluated using *kappa*. The concept of *kappa* (K), or more generally, the inter-coder reliability coefficients is

[6] The intention is that hyp is the segmentation to be evaluated against ref, but given the symmetry of the equation, the difference is ignored in this particular evaluation.

described by the following formula:

$$K = \frac{PA_O - PA_E}{1 - PA_E} \tag{3.2}$$

where the PA_O is the proportion of observed agreement and PA_E is the proportion of expected agreement (the probability of agreement by chance). In *kappa*, they are defined as follows:

$$PA_O = \frac{\text{Number of agreements}}{n}{}_7 \tag{3.3}$$

$$PA_E = \frac{1}{n^2} \sum pm_i \tag{3.4}$$

This measurement is considered to be more accurate than other methods, such as simply reporting the proportion of agreements, because it takes the probability of random agreement into account.[8] Thus, the measure enables comparisons among various studies, with different numbers of labels, and different numbers of annotators.

What constitutes the acceptable level of kappa is open to debate and researchers have proposed a wide range of levels as the acceptable level from 0.4 to 0.9 (Neuendorf, 2002). Many researchers, however, use somewhere between 0.75 and 0.8 as the "rule of thumb" (Ellis, 1994). Krippendorff (1980) considers $0.8 < K$ indicates good reliability, while $0.67 < K < 0.8$ allows only tentative conclusions. Carletta (1996) also points out that when the unit of analysis (segmentation) is not provided pre-theoretically, the task of coding may become inherently more difficult than content analysis studies. For example, in the dialogue act annotation study by Jurafsky et al. (1997), eight linguistics graduate students annotated 1,155 conversations with 42-label annotation scheme, resulting in $K = 0.80$. However, as Table A.4 in the Appendices shows, only two data sets out of 12 data sets that had been used in previous studies reported any inter-coder reliability measures, a much lower ratio

[7] The notations of standard variables are: n is the number of instances; pm_i is each product marginal; and p_i is each joint marginal proportion.

[8] Scott's *pi* (1955) is similar to *kappa* in that it also takes agreements by chance into account. However, Scott's *pi* defines PA_E as $\sum p_i^2$, assuming the equal distributions of labels across the annotators. Thus Cohen's *kappa* is an improvement over Scott's *pi*.

than even the early days of content analysis.[9]

In this study, the inter-coder agreements and inter-coder reliability coefficients were calculated at multiple levels, based on the hierarchical organization of the coding scheme. Specifically, the levels of evaluation performed were: 1) dialogue act functions only, 2) dialogue act domains at the higher (category) level only, 3) functions and domains at the higher (category) level, 4) dialogue act domains at the detail (sub-category) level only, and 5) functions and domains at the detail (sub-category) level.

3.4 Machine Learning

3.4.1 Approach

The task of a machine learning system is to induce a *model*, an abstract representation of a phenomenon in the world, based on data available to the system. While various machine learning algorithms have been applied for the dialogue act annotation task, they are mainly categorized into two approaches. The first approach is to treat the problem as a machine learning text classification problem, where a system attempts to learn a function that maps utterances to dialogue act labels based on a set of attributes (features) of each instance. The performance of such systems mainly depends on the algorithm that the system deploys and the features available to the system. Among the various algorithms, SVM (Cortes and Vapnik, 1995) has been one of the most-widely used algorithms for text classification tasks in recent years (Cohen et al., 2004; Carvalho and Cohen, 2006; Hu et al., 2009). The second approach is to induce an abstract temporal model of dialogues based on observed evidence. The most common example of this approach is the Hidden Markov Model (HMM), which constructs a probabilistic temporal model, represented by a weighted finite state automaton. The HMM has been widely used for speech recognition or hand-writing recognition

[9] According to Riffe and Freitag (1997) (cited in Neuendorf (2002)), only 56% of content analysis studies published in *Journalism and Mass Communication Quarterly* from 1971 to 1995 reported inter-coder reliability.

and applied to many of the early automatic dialogue act (DA) annotation studies done by researchers in the speech and dialogue communities (Kita et al., 1996; Reithinger et al., 1996; Reithinger and Klesen, 1997; Jurafsky et al., 1997; Stolcke et al., 2000).

For the purpose of DA annotation, the text classification approach has an advantage in providing various algorithms to model the relationship between utterances and DAs. This approach, however, is not well-suited for capturing patterns in DA sequences. The label sequence learning approach, on the other hand, tends to be weak in discriminative power, depending more on the explicit feature representations, rather than the algorithms (more on the shortcomings of the traditional approaches are found in Section 2.4.3). There have been efforts to combine the two approaches (e.g. Lafferty et al. (2001), McCallum et al. (2000), Punyakanok and Roth (2001), Altun et al. (2003) and Joachims (2008)), but it is yet to be seen if any of these methods are more effective when compared to existing methods for DA annotation tasks.

Thus the machine learning experiments in this study compared the effectiveness of two learning algorithms: SVM, from the text classification approach, and HM-SVM, from the combined approach.

3.4.2 Features

The machine learning experiments also tested the effects of various features. The literature review found over thirty features that have been employed for the previous studies that conducted machine learning experiments of dialogue act annotation. The features were often derived from different theoretical motivations (if any) and created from different parts of data or processes. Some of them were derived from the surface forms (e.g. word vector) while others were derived from syntactic analysis (e.g. part of speech), semantic analysis (e.g. "meaningful" word *n*-grams), phonological properties (e.g. acoustics and prosody), metadata (e.g. speaker, duration), etc. In this study, the features were selected based on empirical observations.

Feature	Description
Word vector	Word frequencies in the text segment.
Speaker	A binary feature that represents the speaker (either the librarian or the user).
Message sequence number	The number that is assigned to the message that represents the sequence in the conversation
Text segment length	The number of words in the text segment.
Word bigram vector	Frequencies of Word bigrams in the text segment
Message position	The relative position of the message in the dialogue, represented by the proportion of the sequence number of the message over the number messages in the conversation.

Table 3.10: Features

Specifically, dependencies between the dialogue acts and features of the data were analyzed by examining their cross distributions in the annotated data at the discourse analysis stage (See Section 4.1.5 for the analysis.). As the result of the analysis, six features (word vector, speaker, message sequence number, text segment length, word bigram vector, and message position, described in Table 3.10) were selected as they appeared promising to provide positive effects from the analysis described in Section 4.1.

3.4.3 Procedure

The overall process of the experiments is illustrated in Figure 3.5. The following are brief descriptions of the experiments.

1. The annotated data was exported from the database to a text file, formatted with all the features, creating the Gold Standard. For the sessions that were annotated by multiple annotators, the annotation by Annotator 1 or 3 was preferred over the annotation by Annotator 2 in order to achieve the highest consistency (Annotator 1 and 3 had higher agreements with each other than with Annotator 2. See Section 4.1.3 for the inter-coder agreements.).

2. The Gold Standard was then duplicated to eight data sets: F16, F17, F18, F20, F24, F32 and F48. Each of the datasets, except F16, included one additional feature compared to F16, which had only the word vector features. The names of the datasets represented the features that were included. Each feature was assigned to a binary

digit number (message sequence number: 1, speaker: 2, message length: 4, message position: 8, word vector: 16, and bigram vector: 32) and the number in each name was the sum of the numbers of features used for the data set. For example, F18 represented word vector + speaker (16 + 2 = 18).

3. Each dataset was split into ten sub-data sets for the 10-fold cross validation process.
4. For F16, training and testing were done using both SVM and HM-SVM.
5. For other data sets, training and testing were done with HM-SVM only.
6. For each experiment, confusion matrices and standard measures were generated and analyzed.
7. Based on the results from the previous step, an additional experiment was performed including all the features that improved the HM-SVM baseline system. (The dataset depicted as Fxx in the figure.)

The annotated data was stored in a database system using MySQL and processed by a set of Java programs that were developed by the researcher. As for the machine learning software, SVM^{multiclass} and SVM^{hmm}, both of which were implemented by Joachims (1998, 1999, 2008), were used. Since these two programs were developed upon the same platform, SVM^{struct}, the outcomes from the two programs enabled the direct comparison of the algorithms while ensuring the same conditions for the other parts of the implementation, such as the optimization of the kernel implementation or the algorithms for the multi-class selection. All the experiments were subjected to the 10-fold cross validation. Given the characteristics of the data (high-dimensional, sparse vectors), the linear kernel was used following Joachims (1998). The Epsilon parameter for the HM-SVM and SVM algorithms, which specifies the required precision to terminate the learning iteration, was set to 0.5, following Joachims (2008). The iterations were set to terminate after 500,000. Default values were used for all of the other parameters.

The preparation for the experiments, including the design, selection of the software, and formatting the data, started in December 2011. Three months, from February to April 2012,

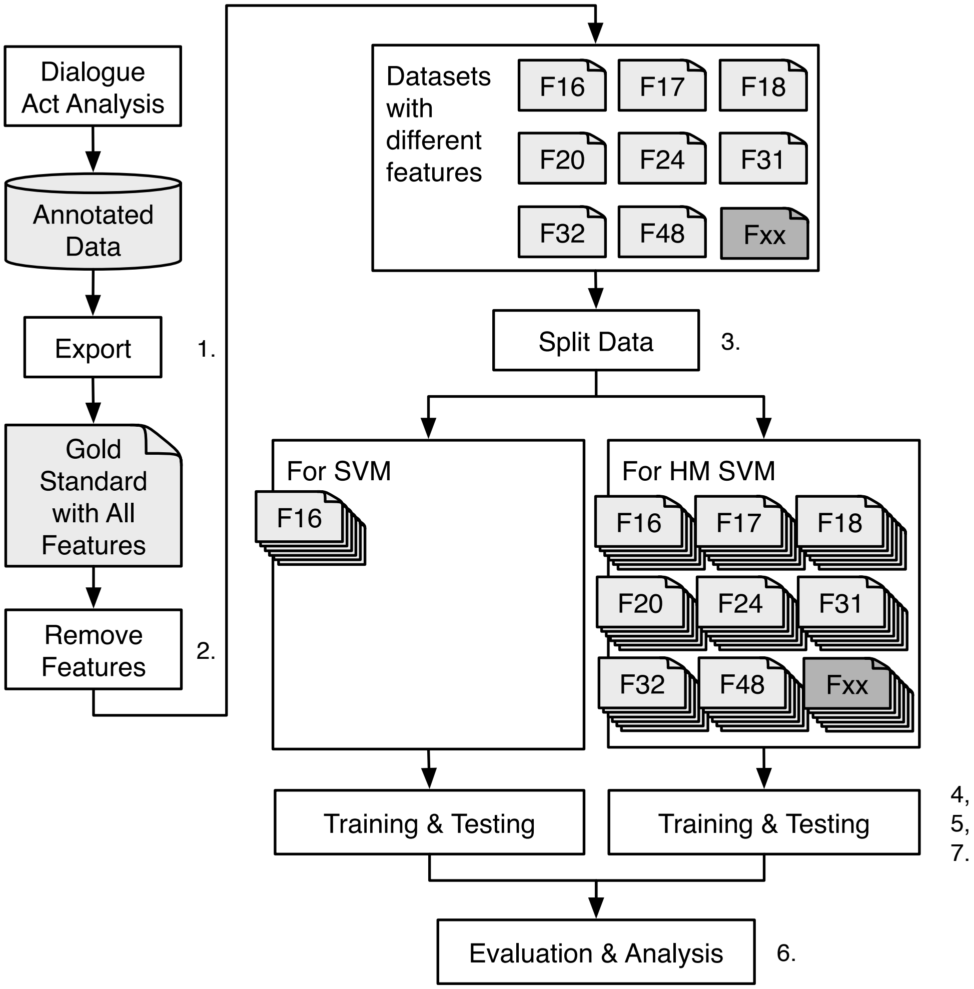


Figure 3.5: Experiment Overview

were spent developing Java programs to export the data and generate necessary features and Ruby scripts to provide an environment for the 10-fold cross-validation and analysis. The majority of the experiments were performed from May 2012 to June 2012.

The experiments were performed at two levels of annotation: 1) function categoy and 2) domain category, both of which produced an inter-coder reliability coefficient $K > 0.75$. Thus, the process depicted in Figure 3.5 was performed twice – once for learning the function categories and another time for learning the domain categories.

3.4.4 Manual Text Segmentation

The procedure of machine learning experiments described above assumed that proper text segmentation was done a priori to the classification task. This assumption is often justifiable for text classification tasks where the unit of analysis is the unit of data collection (e.g. a newspaper article) or consists a predefined syntactic unit (e.g. a paragraph or sentence). In this study, however, the unit of analysis was not predefined – it was rather discovered by the annotators who performed text segmentations. Thus in terms of proving the feasibility of the automated DA annotations, including the text segmentation task as part of the task of the machine learning would be desirable. Doing so, however, would increase the complexity of experiments to the point it would not be feasible. Thus, machine learning experiments of the text segmentation task for the DA annotation was left as a goal for future studies, and the focus of the experiments was on the annotation task.

Using manually-segmented texts as inputs for DA learning is also a common practice, as seen in Jurafsky et al. (1997) and Stolcke et al. (2000), where the Linguistic Data Consortium (Meteer and Taylor, 1995) provided data with predefined text segments.

Chapter 4

Findings and Outcomes

4.1 Discourse Analysis

4.1.1 Overview

4.1.2 Volume of Annotation

The annotation mainly took place from April to August 2011 (16 weeks), following two weeks of training in March, and was followed by the "brush-up" period in October and November. Each week, annotators were given roughly the same amount of data based on the number of messages (approximately 100 - 150 messages, or five to ten conversations). Table 4.1 summarizes the volumes of annotation done by each annotator – Annotator 1 annotated 103 conversations, Annotator 2 annotated 62 conversations, and Annotator 3 annotated 97 conversations. Annotator 2 had to leave the project due to a job opportunity at the end of May, which made her annotation volume lower than the others. While the overall volumes were different, the proportions of each unit (the number of messages, segments, dialogue acts, or words) to another unit was consistent across the annotators. On average, each conversation had approximately 26 messages, and each message was segmented into about 1.5 text segments. The length of each text segment averaged 8 to 9 words. The number of dialogue acts per text segment was consistently very close to 1, ranging from 1.03 to 1.06.

Annotator	**Sessions**	**Messages (per ses.)**	**Segments (per mes.)**	**DA (per seg.)**	**Words** (per DA)**
1	103	2621 (25.45)	4002 (1.53)	4192 (1.05)	37879 (9.04)
2	62	1634 (26.35)	2498 (1.53)	2575 (1.03)	19764 (7.68)
3	97	2555 (26.34)	3886 (1.52)	4135 (1.06)	32558 (7.87)
Total	209*	5441* (26.03)	10386	10902	72155*

* The numbers in the column do not add up because of overlaps in the annotations.

** The numbers are the numbers of space-separated terms in texts.

Table 4.1: Volumes of Annotation

4.1.3 Evaluation

4.1.3.1 Overview

The evaluation of the annotation was done by using standard pair-wise evaluation measures applied to data that two or more annotators annotated, and by analyzing disagreements among the annotators. The following are the volumes of overlap in annotation among the annotators, in the numbers of interview sessions.

- Annotator 1 and 2: 23 (22% for Annotator 1 and 37% for Annotator 2)
- Annotator 1 and 3: 21 (20% for Annotator 1 and 22% for Annotator 3)
- Annotator 2 and 3: 19 (31% for Annotator 2 and 20% for Annotator 3)
- All: 10 (10% for Annotator 1, 16% for Annotator 2 and 10% for Annotator 3)

Overall, the proportion of data annotated by two annotators were maintained at about 20%, and the proportion of data annotated by all the three annotators were maintained at 10%. This was based on Wimmer and Dominick (1997), which suggested 10% to 20% overlap. Given the average number of messages (approximately 25), each annotator had approximately 500 messages in common with another annotator, and 250 messages in common with all the other annotators. These sub-sample sizes also satisfy the general guideline by Neuendorf (2002), which states the sub-sample size should be more than 50 and should rarely need to exceed 300.

The text segmentation and annotation of the dialogue act labels were evaluated separately because of the complexity of evaluating the dialogue act annotation as described below. The text segmentation was evaluated using *WindowDiff*, and the annotation was evaluated with percentile inter-coder agreements and inter-coder coefficients, *kappa* and *pi* (See Section 3.3.5 for description of the standard evaluation measures).

4.1.3.2 Evaluation of Text Segmentation

The evaluation of text segmentation was done using *WindowDiff*. As described in Section 3.3.5.2, *WindowDiff* is a pair-wise evaluation metric for text segmentation based on P_k.

For each pair of annotators, the average *WindowDiff* was calculated over all the messages that the two annotators annotated. Following are the results:

- **Annotator 1 and 2:** 0.039
- **Annotator 1 and 3:** 0.035
- **Annotator 2 and 3:** 0.038

The very low values above indicate excellent agreement between the annotators (the value of *WindowDiff* varies from 0.0 to 1.0, 0.0 indicating the perfect agreement). Annotator 1 and Annotator 3 had a slightly higher agreement, which seemed consistent throughout the study. This may be largely because of the lack of the "brush-up" period by Annotator 2. Even so, all three pairs showed excellent agreement. These very good results are partly due to the fact that the texts to be segmented in this task were much shorter than typical text segmentation tasks. In many cases, each text (message) contained only one segment and if two annotators agreed on it, the *WindowDiff* for that message was 0.0, which lowered the average *WindowDiff*.

4.1.3.3 Evaluation of Dialogue Act Annotation

The evaluation of annotation was performed at each level of the hierarchy of the annotation scheme structure (described in Section 3.3.2). For each message, agreements (or disagreements) were identified by comparing the dialogue acts that were annotated by two annotators in the order of occurrence. The dialogue acts that were in the same of order of occurrence as another dialogue act were called "corresponding dialogue acts". Some dialogue acts had corresponding dialogue acts while others didn't – when two annotators label a message with different numbers of dialogue acts, say n and $n+k$, last k dialogue acts annotated in the second annotated data do not have corresponding dialogue acts. Thus, inter-coder agreements were measured at two levels: 1) all the dialogue acts and 2) all the dialogue acts that had corresponding dialogue acts.

Table 4.2 shows an example of different annotations by two annotators. In the example, Annotator 2 grouped two sentences into one dialogue act text segment. In this case, three comparisons should be drawn for evaluation:

1. the first dialogue act by Annotator 1 and the first dialogue act by Annotator 2,
2. the second dialogue act by Annotator 1 and the second dialogue act by Annotator 2, and
3. the third dialogue act by Annotator 1 and the third dialogue act by Annotator 2.

However, since Annotator 2 labeled only two dialogue acts, the third comparison cannot be done. In other words, the third dialogue act by Annotator 1 does not have a corresponding dialogue act. In this example, the agreement is 0.33 at level 1) and .50 at level 2).

Original Message:	
Thank you very much for using the service. Please come again. Bye!	
Annotator 1:	
Text Segment	**Function / Domain**
Thank you very much for using the service.	Social Rel. Mgmt / Gratitude
Please come again.	Social Rel. Mgmt / Rapport Building
Bye!	Social Rel. Mgmt / Valediction
Annotator 2:	
Text Segment	**Function / Domain**
Thank you very much for using the service.	Social Rel. Mgmt / Gratitude
Please come again. Bye!	Social Rel. Mgmt / Closing Ritual

Table 4.2: An Example of Different Text Segmentations

Although the annotators could label multiple dialogue acts for one text segment, the evaluation and analysis of dialogue acts primarily concerned one dialogue act per text segment (the reasons for this decision is discussed in Section 4.1.4, which concerns the text segments with multiple dialogue acts). More specifically, the following rules were used to select one dialogue act from multiple dialogue acts that are annotated for the same text segment.

For functions:

f.1) If all the functions were the same, that function was selected.

f.2) Otherwise, priority was given in the order of: 1. *Info Provision*, 2. *Info Request*, 3. *Task Mgmt*, 4. *Relationship Mgmt*, and 5. *Comm Mgmt*.

For domains:

d.1) If all the domains were the same at the subcategory level, that sub-category was selected.

d.2) If all the domains were the same at the category level but were different at the subcategory level, the most general label for the category was selected. For example, a domain label *Information Problem:Topic* was considered more general than others, e.g. *Information Problem:People* and *Information Problem:Location.* When such labels could not be identified, a label was randomly selected from the ones that were annotated by one of the annotators.

d.3) If the domains were in different categories, a category was selected in the order of: 1. *Information*, 2. *Task Mgmt*, 3. *Social Rel Mgmt*, and 4. *Communication Mgmt*, and the most general label for the category was selected. If the general label was not available for the particular category, the subcategory was randomly selected from the category.

During the execution of the experiments, these rules were rarely used, since most (96%) of the text segments were labeled with one dialogue act. Furthermore, even if they were labeled with multiple dialogue acts, most of them were labeled with *Info Provision* or *Info Request* for functions and a subcategory of *Info:Problem* or *Info:Object* for domains. Overall, the random selection (d.2, and d.3) was used only to 149 text segments (less than 2%). Thus, while the process could be improved by incorporating priories for selecting all the labels or having general labels for all the subcategories, the added processes and complexities would not have benefited the accuracy of the evaluation.

Table 4.3 lists the inter-coder agreements and coefficient for five levels of analysis. The evaluation measures for the annotation showed a similar pattern as the evaluation measures for the text segmentation: the agreement between Annotator 1 and Annotator 3 was higher than the agreement between Annotator 1 and Annotator 2 or the agreement between Annotator 2 and Annotator 3. Again, this is considered due to the lack of the "brush-up" period by Annotator 2, as mentioned earlier. Over all, the reliability met the "rule of thumb", $0.75 \leq K \leq 0.80$ (Ellis, 1994), for the first three levels of analysis (Function, Domain, and Function + Domain). Especially, the Function annotation yielded $K > 0.8$, which is the standard for a good reliability according to Krippendorff (1980). The last two levels, Domain Subcategory and Function + Domain Subcategory, yielded notably worse agreements compared to the first three. Thus, in this document, observations from the first three levels of analysis are reported as part of the findings of the study. Observations from the subcategories are reported supplementally.

4.1.3.4 Common Disagreements for Functions

The most common disagreements at the function level were 1) *Info Provision* and *Social Rel Mgmt*, 2) *Info Provision* and *Comm Mgmt* and 3) *Info Provision* and *Task Mgmt*. These disagreements were proportional to the frequencies of the labels (especially the dominance of the *Info Provision* function), and while the rankings based on frequencies were different among the annotators, the top three remain the same, and their differences were relatively small.

Common Disagreements for Domains

Disagreements for the domain level appeared when the meanings of two labels are similar or tend to co-occur. In the following, two of the most common disagreements are described.

1. *Info:Problem* and *Info:Object*

Level of Analysis	**Agreement**	**Kappa**	**Pi**	**Reliability**
Annotator 1 and 2:				
Function	0.82 (0.88)	(0.83)	(0.83)	Good
Domain	0.71 (0.77)	(0.74)	(0.74)	Tentative
Function + Domain	0.69 (0.75)	(0.73)	(0.73)	Tentative
Domain Subcategory	0.60 (0.65)	(0.63)	(0.63)	
Function + Domain Subcategory	0.59 (0.64)	(0.62)	(0.62)	
Annotator 1 and 3:				
Function	0.87 (0.92)	(0.88)	(0.88)	Good
Domain	0.80 (0.84)	(0.82)	(0.82)	Good
Function + Domain	0.79 (0.83)	(0.81)	(0.81)	Good
Domain Subcategory	0.67 (0.70)	(0.69)	(0.69)	Tentative
Function + Domain Subcategory	0.66 (0.70)	(0.69)	(0.69)	Tentative
Annotator 2 and 3:				
Function	0.81 (0.88)	(0.82)	(0.82)	Good
Domain	0.70 (0.75)	(0.73)	(0.73)	Tentative
Function + Domain	0.79 (0.74)	(0.72)	(0.72)	Tentative
Domain Subcategory	0.63 (0.68)	(0.66)	(0.66)	
Function + Domain Subcategory	0.62 (0.66)	(0.65)	(0.65)	
Average:				
Function	0.83 (0.89)	(0.84)	(0.84)	Good
Domain	0.73 (0.79)	(0.76)	(0.76)	Tentative
Function + Domain	0.76 (0.78)	(0.75)	(0.75)	Tentative
Domain Subcategory	0.63 (0.68)	(0.66)	(0.66)	
Function + Domain Subcategory	0.62 (0.67)	(0.65)	(0.65)	

* The numbers in the parentheses are based on corresponding dialogue acts only.

Table 4.3: Inter-coder Agreement

Info:Problem refers to a description of the information problem, while *Info:Object* refers to a description of an information object for solving the information problem. In some cases, both of them were specified in one utterance that was inseparable. Below is an example of such utterances by users:

> *Are there any books about Vietnamese Buddhism in the aspen hill library* [1]

In this utterance, three domains of information were specified: 1) the form of the information object (*"book"*), 2) the location of the information object (*"the aspen hill library"*) and 3) the topic of the information problem (*"Vietnamese Buddhism"*). In such situations, the annotators used the domain of information that the speaker intended to specify most in their interpretation for annotation. In the utterance above, some of the information could have been provided or implied previously. For example, the user could have implied that she/he needed something from the local library earlier. Otherwise, the annotators could choose to label both domains. In some cases, annotators disagreed in choosing one label from possible domains.

2. *Info:Search* and *Info:Other*

   Info:Search refers to information related to search processes, e.g. a current progress or strategies, while *Info:Other* refers to information that is not directly related to the information problem, information object or search process, e.g. information about the library or librarian him/herself. In some cases, the difference between the two were unclear. Below is an example of such utterances by a librarian:

   > *I can help you find some websites on leis. But without knowing which library you are near, I will not be able to locate books very effectively.*

   The utterance may be interpreted in two ways: 1) informing the user that the search strategy has not been effectively formed (i.e. *Info Provision / Search:Strategy*) and 2)

[1] A punctuation (question mark) was absent in the original data.

informing the user that the librarian will not be able to search effectively (i.e. *Info Provision / Info:Other:Librarian*) [2].

4.1.4 Text Segments with Multiple Dialogue Acts

4.1.4.1 Distribution of Number of Dialogue Acts per Text Segment

As stated earlier, the average number of dialogue acts per text segment was consistently very close to 1 regardless of annotator. In fact, the vast majority of the text segments were labeled with only one dialogue act, despite the fact that the annotators were allowed to label multiple dialogue acts to one text segment. Table 4.4 shows the numbers of text segments based on the number of dialogue acts per text segment by annotator. The table shows that the distributions are highly skewed but consistent among the annotators. One notable exception (although it is proportionally very small), is that Annotator 2 did not have any text segments with three or more dialogue acts, while other two annotators had about 1%. There are three possible reasons for this.

1. The learning curve of the annotation process

   Although the annotators went through a two-week training period at the onset of the annotation process, dealing with text segments that have multiple dialogue acts may have required more training to increase proficiency. Supporting this, almost half (45%, or 36 out of 80) of the text segments with three or more dialogue acts were because of dialogue acts that were added in the last revision period, a period in which Annotator 2 did not participate.

2. The data

   While it is unlikely that it explains the case completely, given that Annotator 2 worked on a fewer number of sessions, it might be the case that the data Annotator 2 worked on had fewer text segments with multiple dialogue acts. This explanation is supported

[2] Additionally, it may be interpreted as a request to the user for specifying the nearest library for the user (i.e. *Info Provision / Info:Object:Location*).

by the fact that text segments with three or more dialogue acts were very rare (about 1%) and Annotator 2 had similar distributions (in terms of percentile) for the text segments with two dialogue acts or less.

3. The styles of annotation

   While annotators were encouraged to include as many details as possible, individual preference may have resulted in more general alternatives, and thus fewer labels. If this were the case, it should have been corrected by going through the revision stage, but unfortunately, as described earlier, Annotator 2 did not go though the process.

	1	**2**	**3**	**4**	**5+**
Annotator 1	3811 (96%)	114 (3%)	26 (1%)	8 (0%)	0 (0%)
Annotator 2	2393 (97%)	77 (3%)	0 (0%)	0 (0%)	0 (0%)
Annotator 3	3640 (95%)	149 (4%)	38 (1%)	8 (0%)	0 (0%)
Total	9844 (96%)	333 (3%)	64 (1%)	16 (0%)	0 (0%)

Table 4.4: Number of Text Segments vs. the Number of Dialogue Acts

4.1.4.2 Distribution of Dialogue Acts

Among the text segments with three or more dialogue acts, the vast majority was labeled with one function, *Info Provision*, and very few of them were labeled with different functions. Specifically, out of the 80 text segments (64 + 16 + 0, in the bottom row of Table 4.4), 69 (86%) were labeled with *Info Provision*, five (6%) were labeled with *Info Request* and the remaining six (8%) were labeled with different functions. Among the remaining six, two were labeled with *Info Provision* and *Info Request*, which means 76 (69 + 5 + 2, 95% of 80) text segments were entirely about information exchange.

As for the distribution of domain labels, among the text segments with three or more dialogue acts, 75 (94%) were labeled with subcategories of *Info:Problem* entirely.

These observations indicate that it is rare for librarians or users to express *Social Rel Mgmt* or *Communication Mgmt* functions with another function in a single utterance (e.g. asking a question while thanking, or suggesting a book while saying good-bye), which may be intuitive. They also reflect that the classification scheme that was used for the study had more detailed notions in the aspects of information exchanges of descriptions of users' information problems.

4.1.4.3 Treating Text Segments with Multiple Dialogue Acts

In summary, there were three primary observations regarding text segments with multiple dialogue acts.

1. Text segments with multiple dialogue acts, especially ones with more than two dialogue acts, are very rare.
2. Among the text segments with multiple dialogue acts, the vast majority were labeled with the same function label, primarily with *Info Provision* or *Info Request*.
3. They are most often labeled with the domain labels of one category, namely *Info:Problem*.

Given these observations, combined with the fact that the evaluation of the annotation is already complex, the researcher decided to treat each text segment as always labeled with one function and one domain.[3] The rules for selecting the function and domain for a single text segment with multiple DAs are described in Section 4.1.3.3.

4.1.5 Analysis

4.1.5.1 Distribution of Dialogue Acts

The distribution of dialogue acts was analyzed based on the three aspects: 1) dimensions (functions and domains), 2) the speakers (librarians and users) and 3) the relative positions of dialogue acts in conversations.

The distribution of function labels based on speaker is reported in Table A.6 in the Appendices and is summarized in the pie charts in Figure 4.1.

Detailed analysis of the *Info Transfer* and *Social Rel Mgmt* functions (two of the most

[3] These findings potentially disagree with the multi-dimensionality of dialogue act functions that Bunt (2007) claims. This issue is left for future studies (See Section 5.1.3.).

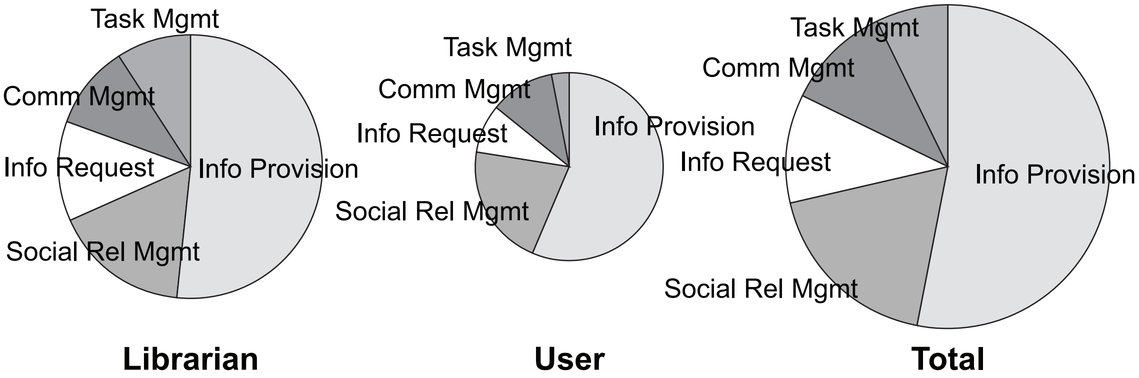


Figure 4.1: Distribution of Functions by Speaker

common types) are presented in the following subsections. Below are primary observations from Table A.6 and the pie charts in Figure 4.1.

First, librarians contribute to the conversations roughly twice as much as users in overall volume, as well as in every category of functions except *Task Mgmt* functions, which were expressed by librarians as six times often as users. Thus, the distributions of dialogue act functions by librarians and the distributions of dialogue act functions by users are very similar, except for the *Task Mgmt* functions. The asymmetry in the distribution of *Task Mgmt* functions can be explained by the nature of reference sessions – librarians are often the ones that carry out search tasks and explain the tasks to users or suggest tasks to users. The symmetry in the distributions of the other functions support the theories that explain the duality of the communication discussed earlier (Watzlawick et al., 1967; Bunt, 1994).

Secondly, *Info Provision* was the dominant function for utterances by both librarians (51%) and users (55%), by a large margin. This was expected, given that the reference encounters are "goal-oriented information-seeking environments" (Radford, 2006b), where the primary goal of the communication is to provide information to satisfy the user's information need. [4]

The dialogue acts for social relationship were used frequently (the second most frequently

[4] The relative proportion of each dialogue act should not be confused with the overall volumes of dialogue acts that speakers contribute to the conversation. The overall volume of the information provision was higher in utterances by librarians (3788) than in utterances by users (2104), as Table A.6 shows.

used by both librarians and users). Roughly, one in six utterances by librarians (17%) and one in five utterances by users (21%) were labeled with some kind of social management function. This is consistent with the observations from previous studies (Ruppel and Fagan, 2002; Nilsen, 2004; Radford, 2006b), which stated the importance of such functions for the success of online reference interviews.

The relative position of a dialogue act was determined by the proportion of the sequence number of the message that contained the dialogue act to the total number of messages in the conversation. Five positions were defined by dividing the messages by the proportions : *Beginning* (from 0 to 0.2), *Beginning-mid* (from 0.2 to 0.4), *Middle* (from 0.4 to 0.6), *Mid-ending* (from 0.6 to 0.8), and *Ending* (from 0.8 to 1.0). The number, five, was chosen after experimenting with three, five, and ten positions, as a good compromise between precision (i.e. how repeatable the measurement is) and accuracy (i.e. how detail the measurement is). The distribution of function labels over the positions are listed in three tables (Table A.9 in the Appendices for dialogue acts by librarians, Table A.10 for dialogue acts by users and Table A.8 for the total), however, since they are overwhelming to interpret, area charts are presented in Figure 4.2 to summarize the tables and illustrate how the contents of the information exchanges between a librarian and a user changes over the progress of a reference interview. These trends are discussed in detail further below.

Functions and Domains for Information Transfer

As Figure 4.1 (or Table A.6) shows, the most dominant dialogue act function for both librarians and users was *Info Provision*, which presented 51% of the librarians' utterances and 55% of the users' utterances. The slightly higher rate for the users may be explained by the nature of reference interviews – users' utterances are primarily for describing their information needs, while librarians' utterances are for requesting further descriptions of users' information needs, and for providing the information that satisfies the users' needs. Supporting this, the most common dialogue act domain by users was, by far, *Info:Problem* (45%), while the most common dialogue act domain by librarians was *Info:Object* (25%), as

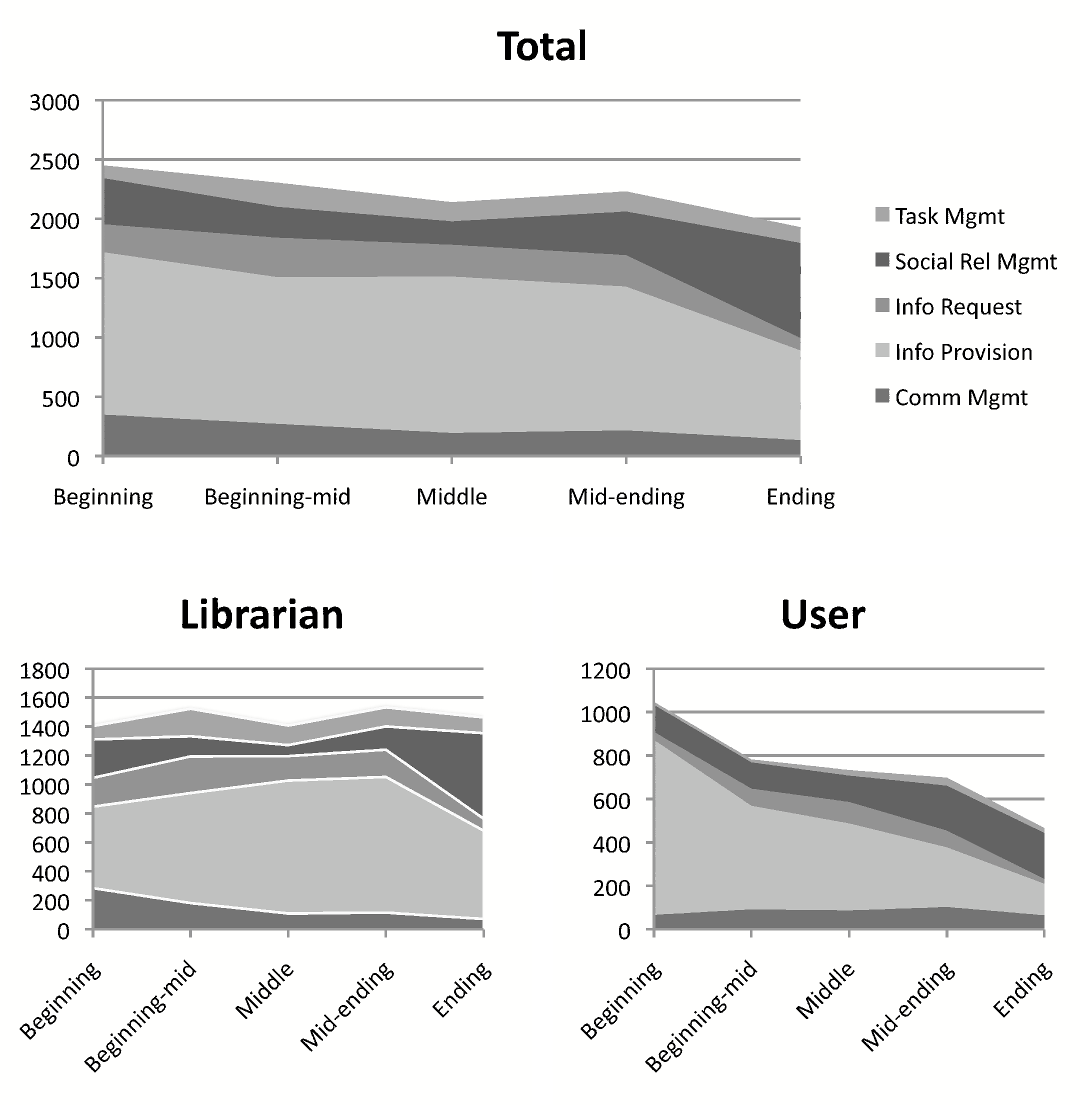
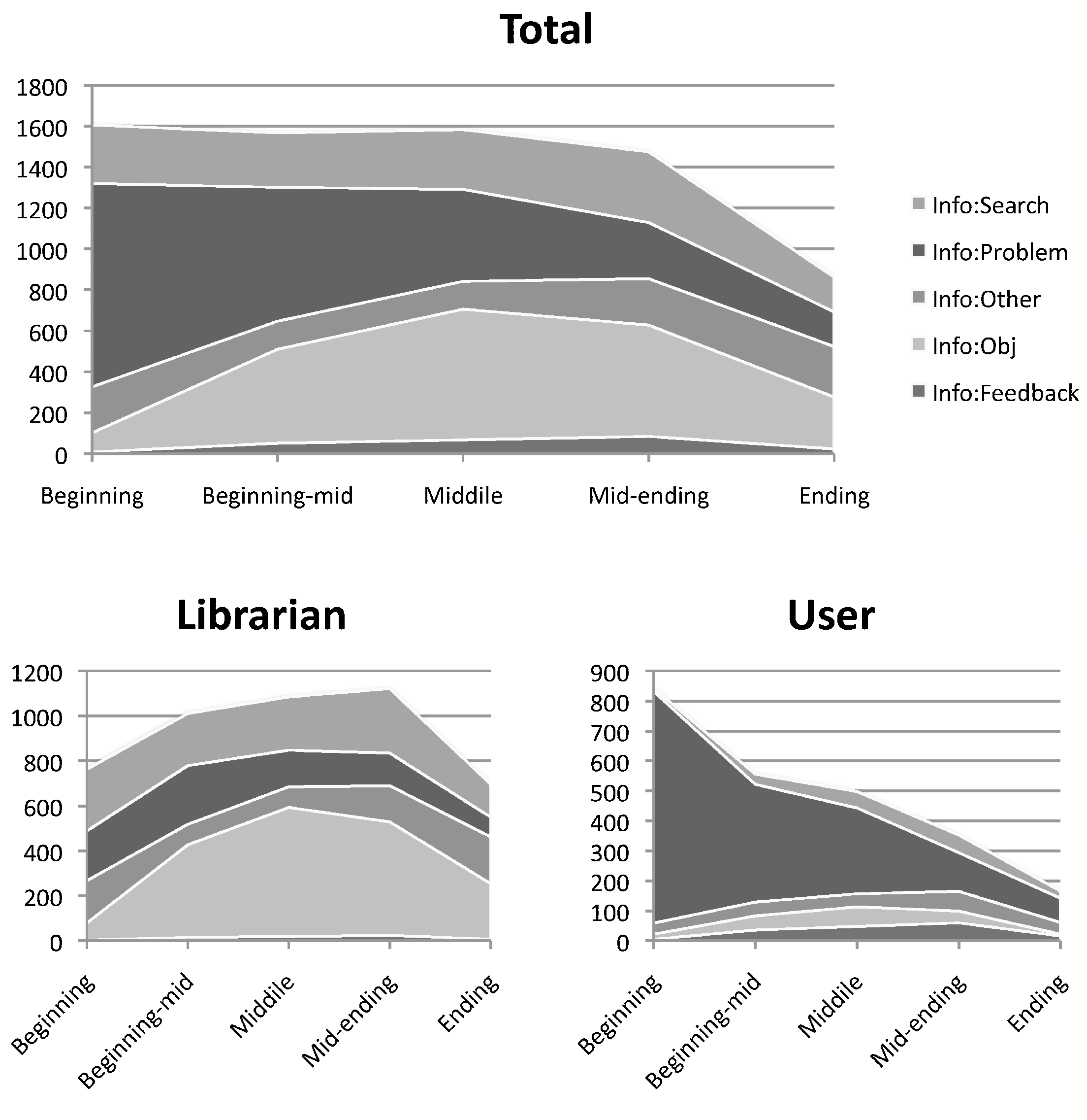


Figure 4.2: Distribution of Functions by Position



Total

Librarian

User

Figure 4.3: Distribution of Information Domains by Position

shown in Table A.7. In addition, the proportion of *Info Request* was higher in librarians' utterances (12%) than in users' utterances (8%), as shown in Table A.6. Table 4.5 shows an example of information exchanges during a reference interview.

	Content	**Function /**	**Domain**
U	why did pakistan change for worse during war on terror?	IP /	Info:Prob:Topic
L	How are you hoping to use this information?	IR /	Info:Prob:Background
L	Five page report? Discussion with friends?	IR /	Info:Prob:Background
U	five page report	IP /	Info:Prob:Background
U	it is requirs 6-8 pages	IP /	Info:Prob:Background
L	http://mblc.state.ma.us/books/magazine/gale.php	IP /	Info:Obj:Source

Speakers: L: Librarian, U:User

Fuctions: IP: Info Provision, IR: Info Request

Table 4.5: Examples of Information Exchanges

In terms of the distribution of information transfer dialogue acts over positions, the following observations were made from Figure 4.2 and Figure 5.1.3.2 (and confirmed with Tables A.9, A.10 and A.8 in the Appendices).

- The use of *Info Provision* by librarians gradually increases until the mid-ending of the conversation and starts decreasing towards the ending.
- The use of *Info Provision* by users is dominant at the beginning and decreases drastically over time.
- The use of *Info Request* is consistent over time for both librarians and users.
- The use of *Info:Problem* domains decreases over time, both by librarians and users, but more drastically by users.
- The use of *Info:Object* domains by librarians increases over time until the middle of the conversation, and decreases towards the ending.
- Other information transfer domains (*Info:Search*, *Info:Feedback*, and *Info:Other*) are fairly consistent over time.

These observations collectively suggest two hypotheses. First, reference interviews involve exchanges of pieces of information, repeatedly or iteratively, which has been suggested by previous studies such as *berrypicking* by Bates (1989) and *micro-level information seeking* by Wu (2005). Second, these information exchanges have a general tendency over time, namely, an increase of provision of information regarding an information object from librarians and a decrease of provision of problem description from users. Since other kinds of information exchanges are fairly consistent, these two variables may be used to characterize the information-seeking interactions. Further studies are desired to investigate how these variables are related to the characteristics of the interviews (See more discussion on this in Section 5.1.3.2.).

Functions and Domains for Relationship Management Functions

The most common examples of *Social Rel. Mgmt* dialogue acts by librarians were in order: 1) *Greeting* (e.g. "Hello!"), 2) *Gratitude* (e.g. "Thank you for using the 24/7 Reference Service."), and 3) *Rapport Building* (e.g. "Best wishes on your exams."). And the most common examples in this category by users, in order, were: 1) *Gratitude* (e.g. "thanks for your help"), 2) *Rapport Building* ("you've been a great help thanks alot for your help"), and 3) *Greeting* (e.g. "hi"). While the rankings were similar, the distributions of these dialogue acts by librarians and users were not exactly parallel (Table A.7). Users tended to show gratitude about 75% more than librarians, but in every other category, librarians contributed more dialogue acts by a large margin: *Apologies* (2.6 times as often), *Closing Ritual* (12 times), *Downplay* (2.5 times), *Greeting* (4 times), *Rapport Building* (2.3 times), and *Valedictions* (1.7 times). This asymmetry may be explained by the following reasons. First, as mentioned above, previous studies have shown the importance of the interpersonal aspects of communication for the success of online reference interviews, thus librarians are trained to use these dialogue acts during the interviews. Second, given the nature of the context, where a librarian is serving a user, it is more likely that the librarian is in a position where she/he needs to apologize to the user (e.g. for lack of resources to help the user) or to declare the end of the session. Third, librarians often use "scripts", prefixed messages that are frequently used in reference sessions. These scripts make it easy to send greetings or closing rituals for librarians. Following are examples of such scripts

> **Example of greeting scripts:**
>
> "Hello! Welcome to the Washington State Library's Ask a Librarian service! My name is [Librarian Name]. I am reading your question right now; it will be just a minute ... "
>
> **Example of closing scripts:**
>
> "Thank you for using Maryland AskUsNow! If you have any further questions, please contact us again."

And lastly, users rarely send verbose closing messages, preferring simple valedictions (e.g. "Bye!") or simply disconnecting the session with no notice.

Table 4.6 shows the numbers of conversations with the *Social Rel Mgmt* dialogue acts for comparison. The higher occurrence of *gratitude* by users coincides with results from the study by Radford (2006b), while the lower occurrence of *apology* differs from Radford's observations. In Radford's study, users showed their respect to their librarians by use of gratitude, praise, apologies, etc., more so than the librarians did to the users. Radford explained the phenomena by hypothesizing that librarians in reference interviews were in a higher "societal status" (p. 1051). This asymmetry between librarians and users, however, was less clear in the data in this study.

Domain	**Number of Interviews**	
	By Librarian	**By User**
Apology	40 (19%)	16 (8%)
Closing Ritual	117 (55%)	12 (6%)
Downplay	46 (22%)	15 (7%)
Exclamation	31 (15%)	32 (15%)
Gratitude	134 (63%)	152 (72%)
Greeting	186 (88%)	57 (27%)
Rapport	115 (55%)	49 (23%)
Valediction	83 (39%)	39 (18%)

Table 4.6: Interviews with the Relationship Management Dialogue Acts

4.1.5.2 Transition of Dialogue Acts

Transitions of dialogue acts were analyzed in terms of a simple conditional probability, i.e. which dialogue acts are how likely to be followed by which dialogue acts. The intention of this analysis was two-fold: 1) to identify structural patterns in the online reference conversations if there are any; and 2) to explore the possibility of improving the machine learning performance by incorporating an algorithm that utilizes conditional probabilities, such as HMM.

Transition of Dialogue Act Functions

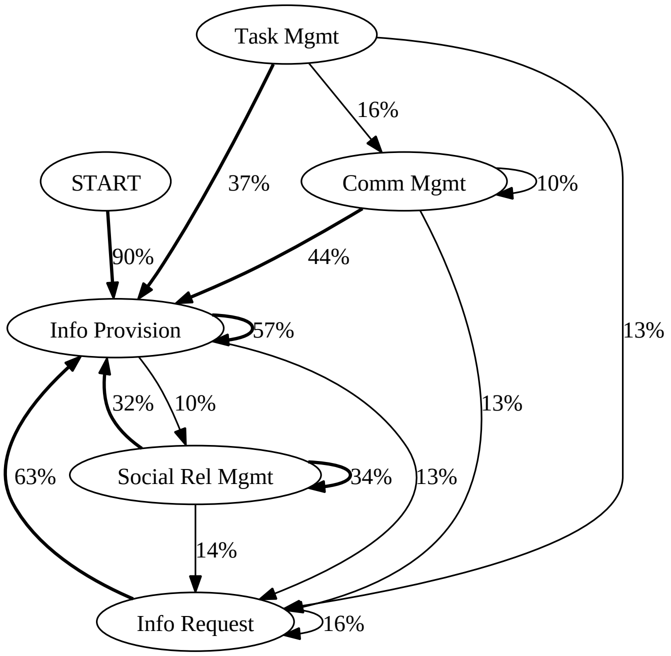
Table A.11 in the Appendices shows the raw frequencies and percentiles of transitions from one dialogue act function to another and Figure 4.4 summarizes the table by visualizing it as a directional graph structure. Although these data represent simple, short sequences of only two dialogue acts, they clearly display some facts about information-seeking dialogues. The following are the observations made from the table and the graph:

1) Four out of five categories (*Info Provision*, *Info Request*, *Task Mgmt*, and *Comm Mgmt*) are followed by *Info Provision*. This reflects that the central task of reference interviews is information provision.

2) The *Social Rel Mgmt* function was followed by the *Social Rel Mgmt* function itself more often. This was expected, since the *Social Rel Mgmt* function includes social gestures such as greetings, apologies, gratitude, and valedictions, which are believed to put pressure to the receiver to respond with a downplay or a proper response, to complete the *adjacency pair* (Levinson, 1983).

3) A reference conversation starts with *Info Provision* most (90%) of the time. This is because the first utterance is usually the description of the information problem or the context by the user.

4) The most common two-dialogue act sequence was from *Info Request* to *Info Provision*, which is most likely represent a question and an answer to it.



* Transitions with a distribution less than 10% were omitted.

Figure 4.4: Transition of Dialogue Act Functions

Transition of Dialogue Act Domains

A similar analysis was also performed for the domains (Table A.12 in the Appendices and Figure 4.5). Transitions of domains were far more complex then one of functions, due to the increased number of labels. The following observations were made:

1) Four information domains (*Info:Problem*, *Info:Object*, *Info:Search*, and *Info:Other*) followed most utterances.

2) *Info:Object* and *Info:Problem* were most often repeated (48% and 46% respectively).

3) *Social: Closing Ritual* and *Social Valediction* most often appeared right before the termination of a conversation.

4) *Comm:Pausing* were exclusively preceded by *Task:Librarian*. This was because, in many cases, librarians explained the task (e.g. searching) they were about to do and paused the conversation while working on it.

5) All the social domains, except *Social:Apology*, were followed by another social domain or the termination of the conversation.

6) Although it did not happen very often, when *Comm:Channel* (e.g. "Are you still there?") appeared in a conversation, it did not lead to any other domains.

The observations above reflect the underlying structure of online reference conversations. Some of the patterns were linguistically motivated (e.g. adjacent pairs), while others were motivated by the nature of the reference conversation. Some of the identified structure at the function level did not provide much more information than what could have been deduced from the frequency distributions (i.e. the dominance of *Info Provision*). Other observations, however, indicate the advantages of utilizing a sequence-learning algorithm for dialogue act annotation.

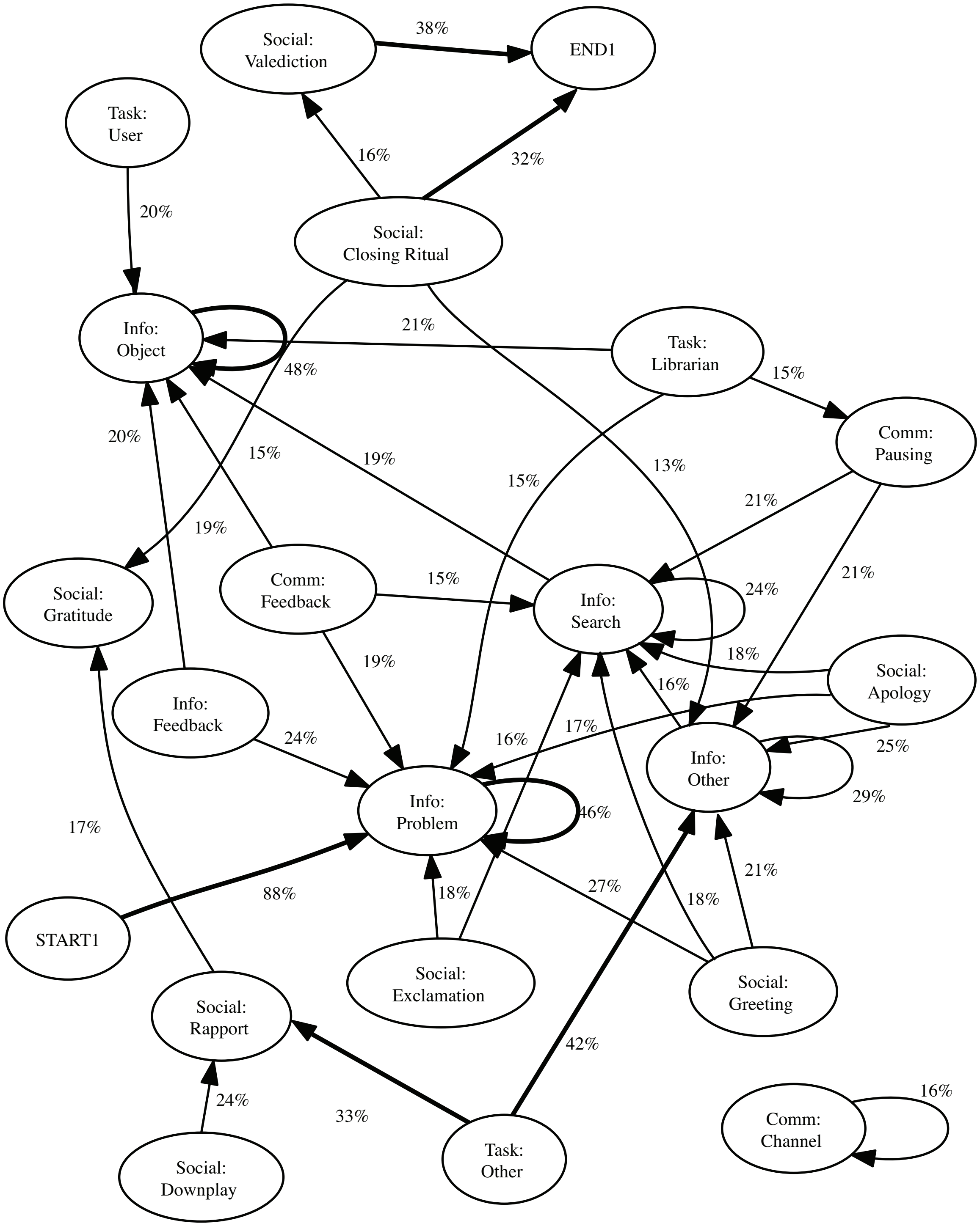
4.1.5.3 Linguistic Forms of Dialogue Acts

Some of the aspects of the linguistics forms of the text segments were analyzed in order to identify the features that are useful for machine learning and why they are so. The following aspects of text segments for each label were analyzed: lengths of text segments (Table A.13), segment frequencies[5] of terms (Table A.14 and A.15), segment frequencies of term bigrams (Table A.16 and A.17), term bigrams at the beginning of text segments (Table A.18 and A.19), term tri-grams at the beginning of text segments (Table A.20 and A.21), and punctuation for text segments (Table A.22 and Table A.23). The tables for the raw figures are all in the Appendices. The following sections outline and discuss on the observations.

Length

The length of a text segment was measured by the number of space-separated tokens in the segment. While it is a simple feature, length has been commonly used as a feature for text classification tasks (e.g. Hu et al. (2009)). The goal of analyzing length was to determine if certain dialogue acts (functions or domains) were labeled to text segments with a certain length consistently.

[5] The "segment frequency" of a term refers to the number of segments in the data that contains the terms in this document.



* Transitions with a distribution less than 10% were omitted

Figure 4.5: Transition of Dialogue Act Domains

As Table A.13 shows, the average lengths of text segments for the underlying task related functions (*Info Provision*, *Info Request*, and *Task Mgmt*) were eight to nine words, while the communicative task related functions (*Social Rel Mgmt* and *Comm Mgmt*) were considerably shorter with a three to four word length. The standard deviations were fairly proportional to the average deviations throughout the categories, except *Info:Object* and *Comm Mgmt*, which had relative standard deviation[6] close to one (1.18 and 0.96, respectively). The reason why the relative standard deviation of *Info:Object* segments was high is that the text segments that were labeled with *Info:Object* were, by and large, of two kinds: provision of the reference information of an information object, and description of an information object. The former was most typically a provision of a URL, and thus only consisted of one term, while describing an information object tended to be longer. Thus, the lengths of *Info:Object* segments varied largely, relative to the average length of the segments. Similarly, the relative standard deviation of *Comm Mgmt* was higher than others since two of the communication domains were relatively long – the average length of *Comm:Channel* was 5.01 and the average length of *Comm:Pausing* was 6.33 – while the average length of *Comm:Feedback* was only 1.07 (because the vast majority of *Comm:Feedback* segments was one word, e.g. "ok"). Most of the social domains were short (one to three words), except *Social:Closing*. This was because, as previously mentioned, librarians often used prepared scripts for closing the interview sessions, which tended to be long.

Table 4.7 summarizes the general trends in terms of the relationship between the length of a text segment and its domain label from Table A.13.

Segment Frequency of Term by Function

Segment frequencies of terms for each function were counted after being filtered by a stop list. The terms were then sorted based on the frequency and the top ten terms for each function were analyzed. The goal of the analysis was to find out if there are terms that

[6] The relative standard deviation, also called coefficient of variance (CV), is calculated by $CV = \frac{\sigma}{\mu}$ where σ is the standard deviation and μ is the average.

indicate certain functions (clues). Stemming was not applied for two reasons: 1) because of the nature of the conversations, which describe information needs or information-seeking processes, tenses may reflect important information in the semantics (e.g. "i'm searching" vs "i searched"); and 2) because of the short length of the text segments, it was expected that the syntactic constructions of texts were limited, and thus, use of derivational or inflectional morphemes would also be limited.

Table 4.8 summarizes Table A.14 by listing the terms that most frequently occurred for a unique function. The two terms for *Info Provision* in Table 4.8 (*transcript* and *joined*) were most often used in a specific way, as part of routines. The word *transcript* was used in a script that informs the user that the transcript could be emailed to the user after the session ("If you would like a transcript of this session emailed to you, please type your full email address now."). The word *joined* was used to inform the user that a librarian has entered the reference session ("24/7 Librarian [Librarian Initials] - A librarian has joined the session."). The three terms for *Info Request* (*school*, *give*, *looked*) represent some of the typical clarifying questions i.e. asking if the information need is for a school project ("Is this for a school project?"), asking for more details ("Could you please give me more information?"), and asking where the user has already looked ("Where have you looked so far?"). *Task Mgmt* functions were naturally represented by text segments with verbs that describe the task (e.g. "I am going to send you a couple of URLS that may help...", "let me check if they have it at the university"). The top terms for the *Comm. Mgmt* functions

Length	**Labels**
Long (8-10)	Info:Problem, Info:Search, Info:Others, Task:Librarian, Task:User, Task:Other, Social:Closing
Medium (5-6)	Comm:Channel, Comm:Pausing
Short (1-3)	Social:Apology, Social:Downplay, Social:Exclamation, Social:Gratitude, Social:Greeting, Social:Rapport, Social:Valediction, Comm:Feedback
Inconsistant	Info:Object

The numbers in the parentheses are average numbers of words in the segments.

Table 4.7: Typical Length for Domain-Level Labels

Function	**Terms**
Info Provision	transcript, joined
Info Request	school, give, looked
Task Mgmt	send, check
Comm Mgmt	minute, reading, moment, minutes, wait, hold
Social Rel Mgmt	askusnow, bye, service, assistance, contact, _patron_name_, maryland, goodbye

Table 4.8: Terms that occur most frequently to a unique function

clearly were used to ask the receiver (most likely the users) to wait (*Comm:Pausing*, e.g. "Hold on a minute"), with time-related words (*minute*, *moment*) or verbs such as *wait* and *hold*. The top terms for *Social Rel Mgmt* are harder to interpret. Terms such as (*askusnow*, *maryland*, *service*) were used in scripts for opening or closing a session. *Bye*, *goodby*, and *[Patron Name]*[7] were obviously used for *Social:Greeting* or *Social:Valediction*.

Although these terms most often represented particular functions and thus might be a good indication of those functions, many of the most-frequent terms for each function also occurred frequently with other functions. This tendency was particularly true among the functions for the underlying goals (*Info Provision*, *Info Request*, and *Task Mgmt*). As Table A.14 shows, more than half the top ten terms (*information*, *find*, *page*, *library*, *email*, *search*, and *session*) were shared among all the three categories. This indicated that while simple word vector features would help narrow down the possibilities, it is unlikely they would be sufficient for correctly predicting one of the three labels. On the other hand, *Social Rel Mgmt* and *Comm Mgmt* had more unique terms for each function. Given that they were also shorter in length, it was expected that machine learning of these two labels would be more feasible than the other functions – the length and word vector features may be sufficient for predicting these two labels.

[7] As described earlier, any words/terms that would identify the users or librarians were replaced with a place holder by the OCLC before the release of the data. *[Patron Name]* is one of such place holders.

Segment Frequency of Term by Domain

Segment frequencies of terms for domains were treated in a similar way as the terms for functions. The terms were counted after being filtered by a stop list, sorted based on frequency, and the top five were selected.

Table 4.9 summarizes Table A.15 by listing the terms that occur most frequently only to a unique domain. The table supports the observations presented above with the terms for functions – *Social* and *Comm* domains seem to have more clear single-term clues than *Information* and *Task* domains.

Domain	**Terms**
Info:Problem	books
Info:Search	joined, search
Info:Feedback	work, helpful
Info:Other	page, online, library
Task:Librarian	send
Task:User	type, click
Task:Other	librarians, respoinse, research
Social:Apology	wrong, time, apoligize, long
Social:Closing	questions, free, feel
Social:Downplay	problem, patient, fault
Social:Exclamation	hmm, wow, wesome
Social:Gratitude	service, 24/7
Social:Rapport	hope, luck
Social:Valedication	cheers, bye, goodbye, night
Comm:Channel	heard, connection
Comm:Pausing	minute, minutes, hold, moment
Comm:Feedback	yeah, correct

Table 4.9: Terms that occur most frequently to a unique domain

Segment Frequency of Term Bigrams

Segment frequencies of term bigrams were counted, much like the segment frequencies of terms, except that a stop list was not used. Term bigrams have been known to be effective features for machine learning dialogue acts (Reithinger and Klesen, 1997; Samuel et al., 1998b; Stolcke et al., 2000). Table A.16 and A.17 show that many of the most frequent bigrams for each label are indeed unique to the label, indicating they are likely to be good additional features for identifying dialogue act labels than using the single word vector alone.

Another observation from the tables is that some of the bigrams in the table typically occur at the beginning of a sentence, and they characterize the segment. For example, if a sentence starts with “i will” or “let me”, it may be expected that the sentence is about task management, e.g. “I will send your request to the business librarian.” or “Let me check the legal periodicals database.”, while if a sentence starts with “are you” or “have you”, it may be expected that the sentence is about an information request, e.g. “Are you looking at US Universities?” or “Have you used any databases?”

Words at the beginning of a sentence often indicate the syntactic construction of the sentence in English,[8] (e.g. whether it is a question, assertion, imperative, etc.) and have been used as features for dialogue act machine learning experiments (Hu et al., 2009). It is also intuitive to expect that certain syntactic constructions are more likely to yield certain dialogue acts in the context of the reference conversation, e.g. imperative sentences most likely are assigning a task to the receiver or sender him/herself; and a question is more likely requesting information. These were not assumptions of this study, since there are many cases where such generalizations don’t apply, for example, rhetorical questions. An important case in this study’s data where the generalization above does not apply is a sentence that starts with “Do you”. Such sentences are most often used in clarifying questions from a librarian (e.g. “Do you mean the new development project for the Woodwards building?”), and thus yield *Info:Request*. But they are also used by a user to specify their information need (e.g. “Do you have a record if JOHN ARNDELL boarded at Portsmouth?”) and thus yield *Info:Provision*.

Increasing the degree of *n*-gram may help in differentiating the sentences (by recognizing the main verbs “mean” and “have” in the case above). It will, however, increase the potential number of features exponentially and may create a feature-set that is too sparse to analyze or to use effectively for machine learning. For comparison, the distributions of the first two words and the first three words were examined. Table A.18 and A.19 show common segment-

[8] This is because English is more of a right-branching language and the constructions are more often head-initial. Thus the syntactic tree can be constructed as a sentence is read (or heard) from the left (or the beginning).

initial bigrams by function or domain, while Table A.20 and A.21 show common segment-initial tri-grams by function or domain. While it is hard to interpret, a few observations are possible. First, for most *Social:* and *Comm:* labels, tri-grams seemed to be overkill, because they simply show the variations of the bigrams, e.g. “thank you for”, “thank you so”, “thank you very”, etc. Also, as reported above, many of them use only one to three words as a whole, so even the single word vector features may be enough to capture whole utterances. And second, for some labels where two words were not enough, three words indeed seem to help differentiate them. For example, “do you have” (*Info:Problem*) vs. “do you think” (*Info:Feedback*) and “i am trying” (*Info:Problem*) vs. “i am looking” (*Info:Search*). Thus, while it was expected that assigning separate features for the sentence initial *n*-gram would help machine learning dialogue act classification, further experiments are required to determine what would be the optimal degree for the task.

Punctuation Mark

The forms of punctuation, especially use of the question mark (“?”) or the exclamation mark (“!”) have been used as features for dialogue act classification tasks (Hu et al., 2009). Table A.22 and A.23 show the distribution of punctuation marks at the end of the text segments by function and domain. The following are notable observations from Table A.22.

First, the most common form of punctuation was not using any punctuation marks, which characterized about 40% of the data. The second most common form was a period (approx. 33%), and the third was a question mark (approx. 12%). In general, these three forms were distributed among different function labels proportionally to their overall distribution, with one exception. *Info Request* functions were mostly represented with a question mark, and conversely, questions marks were most often used for *Info Request.* Second, some punctuation marks (e.g. “!”, “!!”, and emoticons “:-)”) were almost exclusively used for *Social Rel Mgmt* functions. Third, colons (“:”) were most often used for *Info Provision.* Ellipses (“...”) and commas (“,”) were distributed to the functions proportionally, and other forms of punctuations (two or more exclamation marks (“!!”), a combination of an exclamation mark and a question mark ("!?"), a hyphen ("-"), and two or more question marks ("??")) did not happen frequently enough to make any useful observations.

As described in the previous subsection, the form of an English sentence can often be determined by the first few words of the sentence. Thus punctuation marks such as periods or question marks are redundant in terms of representing the semantics of the sentence. Moreover, punctuation is often omitted in the chat environment as described above. Thus, while some of the punctuation marks are expected to be helpful for identifying dialogue acts, it is also possible that using all the punctuation marks as features may do more harm than good.

4.2 Machine Learning Experiments

This section presents the outcomes of the machine learning experiments. The goal of the machine learning experiments was two-fold: 1) to find the optimal algorithm for recognizing dialogue acts in the data, and 2) to find the optimal attributes for recognizing dialogue acts in the data. Specifically, the experiments examined two algorithms, SVM and SVM-HMM, and seven attributes that were identified as potentially useful from the analysis presented in the previous section: the sequence number, speaker, message length, message position, word vector, and two-word (bigram) vector. Although punctuation, segment-initial bigram/trigram, and higher-degree n-grams showed some potential in the previous study, they were left for future studies. As explained in Section 3.4.3, the experiments were performed at two levels of annotation: 1) function and 2) high-level domains. For each level, eight experiments were performed:

1. Standard SVM with the word vector feature only (S-16)
2. HM-SVM with the word vector feature only (H-16)
3. HM-SVM with the additional message sequence number feature (H-17)
4. HM-SVM with the additional speaker feature (H-18)

5. HM-SVM with the additional message length feature (H-20)
6. HM-SVM with the additional message position feature (H-24)
7. HM-SVM with the additional bigram features (H-48)
8. HM-SVM with all of the additional features that had positive effects (H-XX, where XX represents the features used as described below).

The experiments 1) and 2) were to examine if HM-SVM would perform better at this task as expected. The experiments 2) through 7) were to verify the individual effect of each additional feature. And lastly, the experiment 8) was an attempt to maximize the performance of the machine learning and to examine the interactive effects of the all additional features. Each configuration was named by a prefix representing the algorithm (S for standard SVM and H for HM SVM) and a number representing the features. Each feature was assigned to a binary digit number (message sequence number: 1, speaker: 2, message length: 4, message position: 8, word vector: 16, and bigram vector: 32) and the number in each name was the sum of the numbers of features used for that setup. For example, H-17 represents the combination of the HM-SVM algorithm and the message sequence number feature (1) and the word vector feature (16). And H-63 would represent the combination of the HM-SVM algorithm with all the features ($1 + 2 + 4 + 8 + 16 + 32 = 63$).

Outcomes were reported using the standard measures and formatted following the outputs of Weka (Hall et al., 2009), a popular machine learning software package.

The **True Positive (TP) Rate** is the proportion of examples which are correctly classified as category x, among all examples which actually belong the category x, i.e. True Positive Rate = True Positives / (True Positive + False Negative).

The **False Positive (FP) Rate** is the proportion of examples which are wrongly classified as category x, among all examples which do not belong the category x, i.e. False Positive Rate = False Positive / (False Positive + True Negative).

Precision is the proportion of the examples which are correctly classified as category x, among all the examples that are classified as category x, i.e. Precision = True Positive /

(True Positive + False Positive).

Recall is equal to the True Positive Rate. (Listed for the sake of completeness.)

F-Measure is a harmonic mean of precision and recall, i.e. F-Measure = 2 * Precision * Recall / (Precision + Recall).

4.2.1 Learning Dialogue Act Functions

Table 4.10 summarizes the experiment results of learning dialogue act function annotation using SVM and HM-SVM. For comparison, the table also includes the performance measures for the majority classifier. The detailed results are shown in the Appendices: Tables A.24, A.25, A.26, A.27, A.28, A.29, A.30 and A.31 show the confusion matrices, and Tables A.32, A.33, A.34, A.35, A.36, A.37, A.38 and A.39 show the class-by-class measures for all the experiments.

4.2.1.1 SVM vs. SVM-HMM

Table 4.10 clearly shows the advantage of HM-SVM over standard SVM for this task. HM-SVM produced better results for all the measurements than SVM ($\delta = 0.1132$ for the TP rate, $\delta = -0.1503$ for the FP rate, $\delta = 0.1349$ for precision, and $\delta = 0.1737$ for recall). As described in Section 4.1.5.1, the distribution of the function labels were highly skewed, with *Info Provision* constituting more than 50% of all the dialogue acts. Therefore, the majority classifier produced a relatively high precision (0.7551), which was better than the precision

	Setup	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**
	Majority Classifier	0.5716	0.5716	0.7551	0.5716	0.4158
S-16	SVM + word vector	0.6816	0.3392	0.6606	0.6816	0.6115
H-16	HM-SVM + word vector	0.7982	0.1799	0.8055	0.7982	0.7852
H-17	H-16 + sequence number	0.7893	0.1969	0.7972	0.7893	0.7740
H-18	H-16 + speaker	0.8094	0.1801	0.8178	0.8094	0.7963
H-20	H-16 + message length	0.8075	0.1662	0.8108	0.8075	0.7951
H-24	H-16 + message position	0.7975	0.1762	0.8018	0.7975	0.7843
H-48	H-16 + bigram vector	0.8413	0.1361	0.8476	0.8413	0.8358
H-54	H-16,18,20,48	0.8492	0.1322	0.8577	0.8492	0.8437

Table 4.10: Results Summary (Function)

of standard SVM. However, for all the other measurements, SVM outperformed the majority classifier, and HM-SVM was superior to the majority classifier in every measurement.

The confusion matrix (Table A.24) and the class-by-class measures (Tables A.32) for S-16 show that most examples (approximately 82%) were categorized as *Information Provision*, a result of over-generalization given that *Information Provision* represented 53% of the function annotations (Table A.6). As a result, recall for the category was extremely high (0.9905). On the other hand, the *Information Request* category, which tends to have a similar word vector as *Information Provision* (Table A.14), had an extremely low precision (0.037) and recall (0.0028). Thus, the improvement over the outcomes of S-16 could be made, by and large, by decreasing the false positives for *Information Provision* and increasing the true positives for the other categories, especially for *Information Request*.

The class-by-class measures for H-16 (Tables A.33) exactly show the expected improvement. While the recall value for the *Information Provision* category was decreased slightly ($\delta = -0.0314$), the value for the *Information Request* category rose from 0.0028 to 0.4280 ($\delta = 0.4252$), and the precision for both of the categories improved ($\delta = 0.1153$ and $\delta = 0.7198$, respectively). The overall distributions of the output labels were, as shown in the confusion matrix (Table A.33), closer to the overall distributions, and the F-measures for all the categories showed improvement. The most notable improvement was made for the *Information Request* category, for which the true positive count went up from 2 to 333. This improvement in the *Information Request* category can be explained by the observation in the earlier analysis (section 4.1.5.2), which found the asymmetry in the transition between *Information Provision* and *Information Request* (i.e. *Information Request*s are most often followed by *Information Provision*s, but not vice-versa), given the SVM-HMM is designed to learn such dependencies.

The output for the *Task Mgmnt* category showed a similar improvement from S-16 to H-16. Table A.24 shows 395 (approximately 94%) of 421 *Task Mgmnt* examples were mislabeled as *Information Provision* with S-16. Table A.25 shows that the same error was reduced to 208 with H-16.

4.2.1.2 Additional Attributes

Among the five additional features, three features, speaker (H-18), message length (H-20), and word bigram vector (H-48), improved performance over the second baseline system (H-16) while the other two features, sequence number (H-17) and message position (H-24) slightly hurt performance. Among the three features that improved performance, the improvements that were made by the speaker and message length features were marginal (0.0111 and 0.0099 respectively), possibly due to the fact that baseline performance was already very high (precision: 0.8055, recall: 0.7982), especially given the inter-coder agreement was 0.87. The bigram vector features made by far the most improvement over the baseline system as an individual feature[9] ($\delta = 0.0506$ for the F-Measure), which was expected given that two-word sequences often characterize utterances, as described in in Section 4.1.5.3.

The combination of the three features that individually improved performance provided the best improvement ($\delta = 0.0585$) overall, although the improvement was not as good as the simple sum of the individual improvements. Because the three features are not completely independent from each other (e.g. librarians tend to send longer messages and longer messages tend to have higher bigram feature values), it was expected that only marginal improvements would be made by adding up all the features.

4.2.1.3 Overall Performance

Overall, the results showed that HM-SVM successfully learned the annotation of dialogue act functions with appropriate features. The best result the experiment produced was indeed respectable: 0.8784 for precision, 0.8492 for recall, and 0.8637 for the F-measure. These measures are most comparable to the outcomes of the experiments by Hu et al. (2009), which were similar to the experiments in this study in terms of the number of the classes (seven), the nature of the classes ("dialogue function units" in e-mails, which were similar to the dialogue act functions in this study), and the learning algorithms (SVM^{struct}). Hu

[9] The bigram vector or word vector is a set of features that represents the vector, but it is counted as a single feature in the document.

et al.'s experiments produced F-measures of 0.7026 for one data set and 0.8871 for another.

The agreement between the output from the best-performing machine learning system (H-54) and the Gold Standard was 0.8251, which was at the same level as the average intercoder agreement between two annotators (Table 4.3). The agreement coefficient, *kappa*, between H-54 output and the Gold Standard was, on the other hand, 0.7124, which was slightly lower than *kappa* between two annotators (0.75, Table 4.3). This was due to the fact that the algorithm learned the skewed distribution. Overall, the outcomes indicate that the best-performing machine learning system performed the dialogue act function annotation task at the same level as the human annotators.

As mentioned earlier, the analysis of the annotated data found characteristics of the distributions of dialogue act functions that were suited for machine learning: 1) Some functions (*Social Rel Mgmt* and *Communication Mgmt*) had many terms that were used uniquely for those functions, and thus word vector features could effectively identify them; 2) Some labels had order dependencies (e.g. *Information Request* was most often followed by *Information Provision*) and thus HMM could learn these dependencies to predict the next dialogue act function; and 3) Although *Information Request*, *Information Provision*, and *Task Management* were hard to distinguish, *Information Provision* was dominant in terms of the distribution, and thus machine learning systems could default to label any instances with *Information Provision* as a safe bet. It is likely that these characteristics contributed to the high performance of machine learning. Although the details are not available, the distributions of labels in experiments by Hu et al. appear to be similar to the ones of the dialogue act functions in this study. For example, the *Inform* class in their study has a dominant distribution (50-61%), just like the *Info Provision* function has a 54% distribution in this study. Thus, it is also possible that similar characteristics were found with their annotation to contribute to high machine learning performance.

	Setup	TP Rate	FP Rate	Precision	Recall	F-Measure
	Majority Classifier	0.2531	0.2531	0.8109	0.2531	0.1023
S-16	SVM + word vector	0.4434	0.0514	0.5315	0.4434	0.4138
H-16	HM-SVM + word vector	0.6909	0.0576	0.6881	0.6909	0.6674
H-17	H-16 + sequence number	0.6815	0.0548	0.6741	0.6815	0.6604
H-18	H-16 + speaker	0.7046	0.0564	0.7176	0.7046	0.6826
H-20	H-16 + message length	0.6836	0.0555	0.6856	0.6836	0.6608
H-24	H-16 + message position	0.6946	0.0510	0.6797	0.6946	0.6722
H-48	H-16 + bigram vector	0.7185	0.0523	0.7189	0.7185	0.6996
H-58	16,18,24,48	0.7400	0.0461	0.7379	0.7400	0.7272

Table 4.11: Results Summary (Domain)

4.2.2 Learning Dialogue Act Domains

The experimentation for annotating dialogue act domains was done in the same fashion as the one for dialogue act functions. Given the higher number of labels (nineteen, compared to five) and the lower inter-coder agreement rate (0.80, compared to 0.87), a lower performance was expected. Table 4.11 summarizes the experiment results for machine learning of dialogue act domains, again with the performance measures for the majority classifier for comparison. The detailed results are shown in the Appendices: Tables A.40, A.41, A.42, A.43, A.44, A.45, A.46 and A.47 show the confusion matrices and Tables A.48, A.49, A.50, A.51, A.52, A.53, A.54 and A.55 show the class-by-class measures for all the experiments.

As Table 4.11 shows, the overall performance for learning the domains was indeed lower than the performance for learning the functions (the best F-measure: 0.7272 compared to 0.8437). Because of the larger number of categories (18) compared to the functions (5), precision for the majority classifier was very high (0.8109). This was higher than the precision of any of the experimented configurations. However, as in the case with the function annotation task, SVM and HM-SVM were superior in every other measurement.

4.2.2.1 Baseline System: S-16

The confusion matrix for the baseline system S-16 (Tables A.40) shows that false positives were clustered around the information domains (*Info:Object*, *Info:Problem*, *Info:Search* and *Info:Other*). Approximately 75% of all the false positives (2769 among 3715) were associated with those four categories. Nearly 30% of all the false positives (1100 among 3715) were with the *Info:Problem* category. This is, again, the result of over-generalization, based on the overall distribution of the categories. As Table A.7 shows, 62% of the examples were labeled with one of those four categories.

Class-by-class measures (Table A.48) show that most of the social domains (*Social:Apology*, *Social:Closing Ritual*, *Social:Exclamation*, *Social:Gratitude*, *Social:Greeting*, and *Social:Valediction*) performed at a high level, while *Social:Downplay* and *Social:Rapport* performed very poorly. As the earlier analysis in Section 4.1.5.3 revealed, all the high-performance social domains were very short, ranging from one to three words on average, except *Social:Closing Ritual*, which was an average of 9.05-words long. Despite its length, *Social:Closing Ritual* performed well because it was predictable, thanks to the use of scripts – many closing utterances were prescribed scripts and thus had fixed sentences (e.g. "If you have any further questions, please contact us again."). This was also indicated by the segment frequencies of terms, bi-grams, and tri-grams for the domain (Tables A.15, A.17, A.19 and A.21). The *Social:Rapport* utterances were long (averaging 5.8 words) and often personalized to the responder (e.g. "Best wishes on your exams."). The *Social:Downplay* utterances were relatively short (averaging 2.46 words) but also very rare (only 1% of all the examples), which made the learning difficult. In general, categories that performed poorly had one or all of the characteristics which made the learning difficult, e.g. *Comm:Channel* (5.01 words, 1% of all the examples) and *Task:User* (9.25 words, 3% of all the examples).

4.2.2.2 SVM vs. HM-SVM

The comparison between S-16 and H-16 shows, once again, how HM-SVM was more effective than standard SVM for the task. HM-SVM improved the TP rate from 0.4434 to 0.6906 ($\delta = 0.2465$), reducing the overall FP count by 1703 from 3715 to 2012. The FP count among the four information categories listed above were reduced from 2769 to 1261, with the *Info:Problem* category alone, from 1100 to 239. *Social:Downplay* and *Social:Rapport* both performed better with H-16 than with S-16 ($\delta = 0.7865$ and $\delta = 0.2887$ respectively), possibly due the structural dependencies that were identified in Section 4.1.5.2.

4.2.2.3 Additional Attributes

Among the five additional features, three features, speaker (H-18), message position (H-20) and word bigram (H-48), improved the results while the other two, sequence number (H-17) and message length (H-24), slightly hurt the outcome. Unlike the case for function learning, this time, the speaker feature made a substantial improvement ($\delta = 0.152$) over the baseline. This can be explained by two reasons: 1) the domain learning had lower overall performance as compared to function learning and 2) the distribution of domains by speaker depended on the speaker more than the distribution of functions did. For example, Table A.7 shows utterances by users were labeled with *Info:Problem* nearly three times as often as utterances by librarians. On the other hand, utterances by librarians were labeled with *Info:Object* approximately five times as often as utterances by users.

The earlier analysis also showed a relationship between the relative position of a message and the dialogue act domain of the message, which explains the improvement by the message position feature.

The bigram vector feature (H-48), once again, made the most improvement over the baseline system (H-16) as an individual feature ($\delta = 0.0269$ for the F-Measure). Interestingly, the combination of the three features (H-58) provided more improvement ($\delta = 0.0491$) than the sum of all the features improvements (0.0443), indicating there was a positive interactive effect. This actually resonates with the analysis in the previous section, i.e. the distribution of the dialogue act domains depend on the relative position of a message more when the speaker is identified. For example, Figure 5.1.3.2 shows that while *Information Problem* utterances by librarians are relatively consistent throughout a reference session, the distribution of the same kind of utterances by users are noticeably skewed – starting with a very high distribution and rapidly decreasing towards the end.

4.2.2.4 Overall Performance

Overall, the results showed that HM-SVM successfully learned the annotation of dialogue act domains with appropriate features. The best measures, which were produced by the H-58 system, were 0.7837 for precision, 0.7400 recall, and 0.7272 for the F-measure. Although the performance figures for the domain labeling task were lower than the figures for the function labeling task, they were still respectable, given the increased number of labels (from five to twenty), and are comparable to similar previous studies (Surendran and Levow, 2006; Lan et al., 2008; Hu et al., 2009). The performance measures of the H-58 system also displayed a positive interactive effect of individual features. As in the case of function annotation, observations made during the analysis of the annotated data provided explanations for the positive effects of some features, including the interactive effect.

The agreement between the output from the best-performing machine learning system (H-58) and the Gold Standard was 0.7370, which was at the same level as the average inter-coder agreement between two annotators (0.73, Table 4.3). The agreement coefficient, *kappa*, between H-58 output and the Gold Standard was 0.6985, which was slightly lower than *kappa* between two annotators (0.73, Table 4.3). These two measures followed the same patterns as found in the case of the function annotation task. The outcomes indicate that the best-performing machine learning system performed the dialogue act domain annotation task at the same level as the human annotators.

Chapter 5

Discussion and Conclusion

5.1 Discussion

5.1.1 Limitations of the Discourse Analysis

This section describes limitations of the discourse analysis stage of this research, which were derived from the nature of the data or the analytical methods used. The following descriptions are to clarify the generalizability of the findings and to present directions for future studies.

Access to the information objects exchanged

During an online chat reference session, a librarian and a user exchange web pages via the co-browsing feature of the system. In the data used in this study, these web pages were represented as URLs (followed by an indication of the co-browsing *[page sent]*). This means that the annotators did not have access to the contents of the web pages exchanged if the URLs were not valid at the time of analysis, or if they required access from a certain domain (e.g. access to a database of a certain institution). Such URLs may have prohibited the annotators from accurately coding utterances in the data. Specifically, the annotators did not know the information content of the exchanged web page, e.g. an information resource such as an online database, (*Info:Object:Resource*), a reference information of a specific information object such as as an online article *Info:Object:Reference*, etc. When to the page access was not available, the annotators were forced to guess the category of the information object based on the domain name and path. This weakened the meaning of the subcategories of the *Info:Object* domain, and most likely lowered the agreement at that level. In future studies, access to the URLs that are exchanged in interviews is desired for analysis.

Lack of direct input (survey) from the participants

The idea of this study was developed with the intention of integrating linguistic analysis, human language technologies, and information-seeking behavior research. In particular, the linguistic analysis was designed so that the findings could be applied to machine learning, in regard to selecting the machine learning features. Thus the study did not employ surveys or any other methods that direclty elicited input from the users or librarians, even though such elicitation might have been useful for interpreting the data or confirming observations. In this regard, the researcher plans to analyze the outcomes of this study in light of the findings from the research project led by Connaway and Radford (2011). Connaway and Radford have conducted a series of studies investigating digital reference services using multiple data sources, including the same data sets as the ones this study use, employing multiple methods (focus group, interviews, survey and content analysis). Among many findings, they claim that accuracy, a positive attitude by the librarian, and good communication are critical for the success of the reference service, and that query clarification is the key for accuracy and effectiveness. If there is a correlation between the accuracy of digital reference interviews from Connaway and Radford's study and the distribution of dialogue acts (DAs) from this study, it will be another evidence for the claim.

Other forms of Digital Reference

While emails have been another popular medium for digital reference services, analysis of such data was outside the scope of this study. This was because the following observations were made while the researcher examined email reference data from the Internet Public Library.

1. Lengths of messages:

   Descriptions of information needs that were sent to the reference services were, in general, much longer than the messages in online chat reference. The forms for the service consisted of multiple text fields which represented typical clarifying questions (e.g. “Is this a school project?”, “Which sourced have you already consulted?”, etc.).

2. Lack of question negotiation:

   Questions that were asked in email reference services tended to be better-defined compared to the messages that were sent to the online chat reference services, and thus clarifications were rare.

3. Lack of interactions:

   In most cases, there was no exchange of messages between the librarian and the user, other than the initial question and answers.

The decision was made not to include email data since it was clear that these characteristics would make the integration of the data into the study extremely difficult. Analysis of email-based reference services can be found elsewhere (e.g. Carter and Janes (2000)). Comparing the two modes of communication at reference services is left for a future study.

5.1.2 Challenges of Machine Learning Experiments

As described in Section 3.4, the experimentation was designed based on previous studies and the results of the discourse analysis stage, in order to maximize the performance of machine leaning for the automatic DA annotation task. The approach was, however, by no means exhaustive, in order to keep the number of experiments at a manageable level for reliable execution and meaningful analysis. Specifically, experimenting with more alternatives, such as the following aspects, were left as topics for future studies.

- Algorithms

  Only two algorithms, SVM and HM-SVM, were used in the experiment, as the two of the most promising approaches available today (described in Section 3.4.1). Although the systems in this study performed better than or comparable to the systems that were tested previously, this does not eliminate the possibility that other systems could outperform the HM-SVM for the same task.

- Features

  Similarly, there are features that were previously shown to be effective for DA annotations such as segment-initial n-gram, information gain, named entities (names of locations, people, or organizations), etc, but were not used in the experiments (The motivations for selecting the features were described in Section 3.4.2.).

- Parameters

  As described in Section 3.4.3, parameters for the machine learning were selected based on suggestions by Joachims (1998, 2008). When suggestions were not available, default values were used.

- Interactive Effects

  Interactive effects between the attributes (algorithms, features, and parameters) were tested minimally. For example, the features that did not yield positive effect in the individual experiments were not tested for their interactive effects with other features. While this saved many iterations of experiments, some experiments might have been worth the attempt. For example, the message position feature, which had a positive interactive effect with the speaker feature for the domain annotation task (described in Section 4.2.2), could also have been tested for the positive interactive effect for the function annotation task.

5.1.3 Future Studies

This section describes two topics for future studies that were derived from the outcomes of the study.

5.1.3.1 New Application and Further Development of Methods

While the data used in this study was from a digital reference service, a specific form of online information-seeking interaction, the employed methods could be applied to other forms of information-seeking interaction, such as online forums and community-based question answering services, which are rapidly gaining in popularity. Applying the methods used in this study to such data will further verify the current findings and refine the understanding of information-seeking interactions.

In addition, such studies are desired in order to further develop the methods. For example, as described in Section 3.3.2, the development of the annotation scheme was one of the major challenges during the preparation of the study. The annotation scheme used in this study was an outcome of an attempt to incorporate the theoretical framework of discourse linguistics and models of information-seeking behavior while maintaining the applicability of the annotation scheme to the data, and organization and structure that allow multi-level analysis. But as the annotation progressed, it became apparent that the design of the annotation scheme has a great impact on the feasibility and quality of coding (i.e. inter-coder agreement) and on maintaining linguistic attributes or properties that enable machine learning. While the study produced satisfactory outcomes in many aspects, there is also room for improvement. First, the dimensionality of the DAs need further investigation, as the observations in this study seemingly disagree with some of the previous studies, e.g. multifunctionality of utterances, as mentioned in Section 4.1.4.3. Second, the detail-level codes (subcategories) need refinement, as the annotators could not agree on their coding at a satisfactory level. These two issues are intertwined, as the current coding scheme serves two relevant purposes (dimensionality and detailed categories) with one structure. The coding scheme needs to be reorganized for both ease of coding as well as ease of analysis. Third, the coding scheme also needs additional development to accommodate data outside digital reference. Enhancing the coding scheme while including such revisions and maintaining the theoretical framework (i.e. maintaining assumptions and theoretical motivations) will be a challenging task, but is a sure way to contribute to the body of the knowledge following this study.

5.1.3.2 Evaluation of Information-seeking Interactions using Dialogue Acts

One of the most interesting findings revlealed by this study was the potential of DA annotation to analyze and understand the progress of information-seeking interactions for evaluation and theorization. As described in Section 4.1.5.1, the analysis of the distribution of DAs based on the relative position of the message identified two general tendencies regarding the volume of information exchanges during a reference interview: 1) *Information Provision* of *Info:Object* by librarians (LIO) increases gradually from the beginning towards the middle of the interview, and decreases towards the end; and 2) *Information Provision* of *Info:Problem* by users (UIP) starts with a very high volume and rapidly decreases towards the end (illustrated in Figure in Chapter 4, also in Graph a. in Figure 5.1). Given that the other kinds of information exchanges are fairly consistent throughout an interview session, one may hypothesize that the changes in these two volumes reflect the progress of an information-seeking process. For example, the increase of LIO may indicate that the librarian is providing the user with a variety of information objects, to better understand the broad view of the information need. The decrease of LIO, on the other hand, may indicate that the librarian started narrowing down the user's information need. The ratio of decrease of UIP may indicate how well the librarian is understanding the information need of the user. Thus the patterns created by the changes of the volumes of these two types of DA may suggest the success of the reference interviews, the quality of the service, and the user's satisfaction. If so, what would be the possible patterns? Figure 5.1 illustrates some of the possible patterns LIO and UIP may create: Graph A shows the prototypical pattern that was found in the current data;

Graph B shows a pattern where the users may not be liking the information objects that the librarian is providing while the librarian reformulates the searching strategy repeatedly; Graph C shows a pattern where the librarian may be struggling in finding information for the user and taking longer before narrowing down the information objects; and Graph D shows a pattern where the librarian was quick to figure out the user's information need (or deciding a direction in haste) and thus the volume reached its peak early. If such patterns can indeed be observed, it will be possible to examine the correlations between the patterns and the success of the interviews (e.g. accuracy, effectiveness, user satisfaction), by using the data that was created by Connaway and Radford (2011). Such studies will contribute to the evaluation of reference services (or other information services) as well as theorizing the process to the sucess of online information-seeking interaction using the DAs.

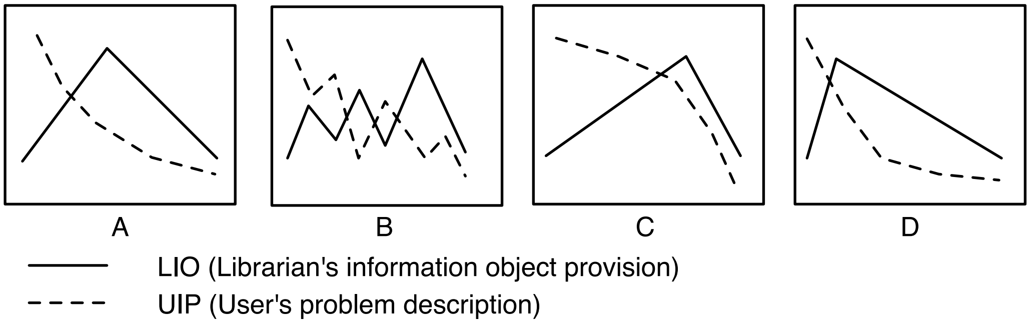


Figure 5.1: Examples of Possible Patterns of Information Provisions

5.2 Conclusion

This study investigated the information-seeking behavior of digital reference services and experimented with automatic identification of DAs using machine learning techniques. The first stage of the investigation was an analysis of the discourse properties of interactions of digital reference services using DA annotation. The annotated data was analyzed along the following dimensions: 1) the distribution of DAs by speaker, 2) the distribution of DAs by the relative position of a message in the conversation, 3) the transitions of DAs, 4) the lengths of the text segments, 5) words and word sequences in the text segments, 6) initial word sequences of the text segment, and 7) punctuation used in the text segments. Observations yielded the following generalizations that confirmed existing theories of communication or the information-seeking behavior:

1. During a reference conversation, the librarian and the user are similarly engaged in the underlying tasks (information exchanges) as well as the communicative tasks (keeping the communication). This confirms an underlying assumption of the study, which originated from the Dynamic Interpretation Theory (Bunt, 1994), that is, speakers need to carry out these two types of tasks during a conversation.
2. Reference interviews involve exchanges of pieces of information, repeatedly or iteratively, which has been suggested by Bates (1989) and Wu (2005).
3. Librarians and users pay close attention to social obligations and relationships throughout the interviews, which has been emphasized by researchers of online reference interviews, e.g. Ruppel and Fagan (2002), Nilsen (2004) and Radford (2006b).

These findings, together with more detailed observations presented in Section 4.1, answered the first research question: What is the discourse of question negotiation in digital reference?

The analysis also found some potential disagreements with previous studies. First, the analysis of distributions of DAs found the text segments with multiple DA functions were extremely rare (less than 1%), while Allwood (1992) and Bunt (2007) claim that utterances are necessarily multifunctional. Second, the analysis of distributions of *Social Rel Mgmt* and *Communication Mgmt* functions by librarians and users disagreed with the observations made in the previous study by Radford (2006b). In Radford's study, users showed their respect to their librarians more so than the librarians did to the users. This was, however, not the case in this study. These issues are left as subjects of future studies, as discussed in Section 5.1.3.

The analysis of the transitions between DAs identified the structural characteristics of the reference interviews. Some of the characteristics were linguistically explained while others were motivated by the nature of the reference conversation. Some of the findings indicated the advantages of utilizing a sequence-learning algorithm for DA annotation, which was later confirmed in the machine learning stage.

The analysis also identified attributes of messages that were more indicative of particular DAs than the others. Based on the findings, features for machine learning were selected and examined in the second stage.

The second part of the study, the machine learning experiments, provided the proof of concept by confirming that there is linguistic evidence representing the discourse semantics, that linguistic analysis in this study could capture it, and that the semantics could be learned by following certain procedures (algorithms). The experimentation employed semi-factorial combinations of different algorithms and features, showing that the automation of DAs annotation was achievable.

Overall, the results showed that HM-SVM successfully learned the annotation of DAs with appropriate features. The best results the experiments produced were 0.8784 (precision) and 0.8492 (recall) for the function labeling task, and 0.7837 (precision) and 0.7400 (recall) for the domain labeling task, both of which are comparable, if not better than, the previous similar studies.

The experiments also demonstrated the possibility of practical applications of the DA analysis for further research across disciplines, such as 1) a new measurement for evaluating virtual reference services, 2) new data attributes for information extraction / retrieval algorithms (document models), and 3) a prototypical dialogue model for constructing fully-automated dialogue systems.

Appendices

Dimension	Category		Description
Function	Information Transfer	Request	Requesting information from the recipient.
		Provision	Providing information to the recipient.
	Task Management	Request	Requesting task to the recipient, e.g. asking for an instruction.
		Provision	Assigning a task to the recipient or committing oneself to a task.
	Communication Management		Managing physical aspects of communication, such as the channel, place, etc.
	Social Relationship Management		Managing socio-emotional aspects of communication.
Domain	Information	Problem	Description of the user's problem or information need.
		Search Process	Description of the search process and related issues.
		Object	Description of a particular information object.
		Feedback	Feedback for the info. object, librarian, search strategy etc.
		Other	Contact information of the participant, etc.
	Task	Librarians	Description of a task for the librarian.
		Users	Description of a task for the user.
	Communicative	Feedback	Confirming the reception of the previous utterance, e.g. *I got it.*
		Pausing	Indicating an interruption of conversation, e.g. *Let me see...*
		Channel	Checking the communication channel, e.g. *Are you still there?*
	Social	Gratitude	Showing an appreciation, e.g. *Thank you!*
		Apology	Showing an apology, e.g. *I'm sorry.*
		Downplay	Downplaying a gratitude or apology, e.g. *You are welcome!*
		Greeting	Saying or responding to a greeting, e.g. *Hello.*
		Valediction	Saying or responding to a valediction, e.g. *Bye.*
		Exclamation	A remark of surprise, frustrations, joy, etc., e.g. *OMG!*
		Rapport	Other expressions for rapport building, such as humor.

Table A.1: Category-Level Annotation Labels

Category		Subcategory	Description
Information	Problem	Lib. Knowledge	What the librarian knows about the information problem and related issues.
		Background	Background of the problem, e.g. the goal of the information seeking or intended use of the information object, the reason why the information is needed, etc.
		Info. Type	The types of information (numbers, person, locations, definition, etc.) or the forms of information (books, articles, web pages) needed.
		Lib. Understanding	The librarian's understanding of the information problem.
		Location	Geographical background of the problem, e.g. place, weather, etc.
		People	Names of people or organizations associated with the information problem.
		Previous Search	History of the user's search process.
		Status	Asking or telling if the user's information need is satisfied.
		Time	Temporal information associated with the information problem, e.g. date, time, year, etc.
		Topic	Topic, subject area, or general description of the problem.
		User Knowledge	What the user knows about the information problem.
	Search Process	Lib Progress	The librarian's activity related to the search for information needed.
		User Progress	The user's activity related the search for information needed.
		Strategy	A general strategy for searching the information needed.
	Object	Access	Accessibility or availability of the information object.
		Description	Specification of an information object, e.g. size, color, location, etc.
		Direct Answer	Utterances where the librarian directly answers the information problem.
		Excerpt	An excerpt, quotation, or extraction from an information source. This also includes summary or paraphrasing.
		Interpretation	Speaker's opinion, impression, or observations about an information object.
		Reference	A reference to an information object, e.g. URL, book title, etc.
		Source	The source of the information, e.g. database.
	Feedback	Info Object	Feedback on an information object.
		Librarian	Feedback on the librarian.
		Problem	Feedback on the description of the information problem.
		Strategy	Feedback on the search strategy.
	Other	Librarian	Information about the librarian him/herself, e.g. personal interest, etc.
		Library	Information about a library, e.g. phone number, hours, etc.
		System	Information about the system, e.g. sending out the chat log to the user.
		Other	Other information.

Table A.2: Subcategory-Level Annotation Labels for the Information Domain

	Techniques	**Utterance Features**	**Structural Features**	**Labels**	**Performance**	**Dataset**
Kita et al. (1996)	HMM	NA	NA	9	NA	ATR Model
Reithinger et al. (1996)	HMM	speaker	DA n-gram	18	71-.76	VERBMOBIL1
Reithinger and Klesen (1997)	HMM	word bigram	da n-gram	18/43	.65/.67, NA/.75	VERBMOBIL2 (Germ. & Eng.)
Samuel et al. (1998a)	TBL	DA Cue, Speaker, word n-gram			.71-.75	VERBMOBIL (English)
Stolcke et al. (2000)	HMM	prosody, acoustic, word n-gram, speaker	DA n-gram	43	.65	SWBD
Fernandez and Picard (2002)	SVM	prosody		43	.65	SWBD
Cohen et al. (2004)	VT, DT, AB, SVM	BOG, TFIDF, temporal expressions, POS, proper noun phrase		5	.68-.88/.70-.89	NF, PW CALO
Carvalho and Cohen (2006)	SVM	meanigful word n-gram	NA	43		NF, PW CALO
Verbree et al. (2006)	J48	length, top POS n-gram, top word n-gram, question mark, or,	previous da, order specific,	42	.70	ICSI, SWBD, AMI
Surendran and Levow (2006)	SVM + HMM	prosody, acoustic, word n-gram, speaker	DA n-gram	13	.66	HCRC
Lan et al. (2008)	ME	DA cues, POS, disfluency, exclamations	DA n-gram, speaker change, length, position	44	.74	SWBD
Hu et al. (2009)	SVM/Str.SVM	1st 3 POSs, exclamations, wh-word, length, head, body, tail, BOG, content words, fillers, acoustic	position	7	.68/.70, .87/.89	Loqui, Enron

Table A.3: Comparison of Machine Learning Experiments for Dialogue Act Annotation

Corpus	**Kappa**	**Type**	**Size**	**Source**	**Studies**
ATR Model	NA	phone	10 dlgs/25 sentences	ATR Corpus	Kita et al. (1996)
ATR Keyboard	NA	email	50 dlgs/1,686 sentences	ATR Corpus	Kita et al. (1996)
VERBMOBIL1	NA	face-to-face	150 dlgs	VERBMOBIL	Reithinger et al. (1996)
VERBMOBIL2 (Germ.)	NA	face-to-face (Germ.)	350 + 87 dlgs	VERBMOBIL	Reithinger and Klesen (1997)
VERBMOBIL2 (Eng.)	NA	face-to-face (Eng.)	143 + 20 dlgs	VERBMOBIL	Reithinger and Klesen (1997), Samuel et al. (1998a)
SWBD1	.80	phone	1155 dlgs	Switchboard Corpus	Jurafsky et al. (1997), Stolcke et al. (2000) Lan et al. (2008)
HCRC		dialogue	128 dlgs	Map Task	Carletta et al. (1997)
NF	.72-.83	email	1716 msgs	CSpace Corpus	Cohen et al. (2004), Carvalho and Cohen (2006)
PW CALO	NA	email	222 msgs	NA	Cohen et al. (2004), Carvalho and Cohen (2006)
Loqui	NA	ref. dlgs	48 dlgs/3845 DAs	Loqui	Hu et al. (2009)
Enron	NA	email	122 threads/1,400 DAs	Enron	Hu et al. (2009)

Table A.4: Data for Machine Learning Dialogue Act Annotation

.

No.	Speaker	Text	Function	Domain
1	User	Hi.	RM	Social:Greeting
		I would like to ask about monkey animal.	IP	Problem.: Topic
		Monkeys from jungle nearby often go to vege garden and pluck off vege plants or fruits available before they ripe.	IP	Problem.: Background
		What are the ways to scre off moneky from vege garden, what they fears anyway...?	IP	Problem.: Topic
		Thanks.	RM	Social: Gratitude
2	User	anyone there able to help?	TP	Task: Librarian
3	Librarian	Hi, [Patron Name]	RM	Social: Greeting
		– let me see what I can find for you.	TP	Task: Librarian
		Do you know what kind of monkeys they are?	IR	Problem: Topic
		If not, could you give me your general geographic location?	IR	Problem: Topic (Location)
		oh yes,	RM	Social: Exclamation
		the jungle is high land about 300m away from the vege garden.	IP	Problem: Topic (Location)
		in front of the garden there is a large pond.	IP	Problem: Topic (Location)
		this is Malaysia it is a tropical country.	IP	Problem: Topic (Location)
5	User	the monkeys are quite large in size.	IP	Problem: Topic
		they have greyish color fur. long tail, long arms	IP	Problem: Topic
6	Librarian	Hmm.	RM	Social: Exclamation
		I am trying to find some tips for you	IP	Search: Progress (Librarian)
		– here is one: "...one initiative that has been taken to scare off monkeys is by tying a toy dog to the jackfruit tree. According to a neighbour, it really works."	IP	Info. Object: Direct answer
		This is a suggestion from a newspaper article that can be found at: [URL]	IP	Info. Object: Source
		... I am still looking for more suggestions...	IP	Search: Progress (Librarian)
7	User	they can pluck up potato plants up to get the potatos below to eat and pluck brinjal to eat.	IP	Problem: Topic
		Thanks... for your kind helps...	RM	Social: Gratitude
		i could think about toy dog...	IP	Feedback: Info. Object

Table A.5: Examples of annotations

Label	**Librarian**	**User**	**Total**
Information Provision	3788 (51%)	2104 (55%)	5892 (53%)
Social Relationship Mgmt	1231 (17%)	785 (21%)	2016 (18%)
Information Request	892 (12%)	314 (8%)	1206 (11%)
Communication Mgmt	760 (10%)	414 (11%)	1174 (10%)
Task Mgmt	659 (9%)	109 (3%)	68 (7%)
Log:Disconnect	47 (1%)	45 (1%)	92 (1%)
Uninterpretable	19 (0%)	19 (0%)	38 (0%)
Total	**7396**	**3790**	**11186**

The percentiles are based on the overall volume per annotator categories (libarian, user, and total).

Table A.6: Distribution of functions by speaker

Label	**Librarian**	**User**	**Total**
Information:Problem	878 (12%)	1662 (45%)	2540 (23%)
Information:Object	1814 (25%)	174 (5%)	1988 (18%)
Information:Search	1177 (16%)	183 (5%)	1360 (12%)
Information:Other	740 (10%)	232 (6%)	972 (9%)
Information:Feedback	68 (1%)	165 (4%)	233 (2%)
Task:Librarian	439 (6%)	38 (1%)	477 (4%)
Task:User	210 (3%)	72 (2%)	282 (3%)
Social:Gratitude	249 (3%)	434 (12%)	683 (6%)
Social:Greeting	310 (4%)	76 (2%)	386 (3%)
Social:Rapport	212 (3%)	92 (2%)	304 (3%)
Social:Valediction	111 (2%)	64 (2%)	175 (2%)
Social:Closing Ritual	160 (2%)	13 (0%)	173 (2%)
Social:Exclamation	56 (1%)	56 (2%)	112 (1%)
Social:Apology	73 (1%)	28 (1%)	101 (1%)
Social:Downplay	64 (1%)	25 (1%)	89 (1%)
Communication:Feedback	243 (3%)	379 (10%)	622 (6%)
Communication:Pausing	431 (6%)	10 (0%)	441 (4%)
Communication:Channel	86 (1%)	26 (1%)	112 (1%)
Total	**7321**	**3729**	**11050**

The percentiles are based on the overall volume per annotator categories (libarian, user, and total).

Table A.7: Distribution of domains by speaker

Label	**Beginning**	**Beg.-mid**	**Middle**	**Mid-end.**	**Ending**
Comm Mgmt	351 (30%)	273 (23%)	196 (17%)	218 (19%)	136 (12%)
Info Provision	1369 (23%)	1237 (21%)	1319 (22%)	1212 (21%)	755 (13%)
Info Request	236 (20%)	332 (28%)	268 (22%)	265 (22%)	105 (9%)
Social Rel Mgmt	389 (19%)	260 (13%)	197 (10%)	368 (18%)	802 (40%)
Task Mgmt	106 (14%)	203 (26%)	160 (21%)	168 (22%)	131 (17%)

Table A.8: Distribution of functions by message position (overall)

Label	**Beginning**	**Beg.-mid**	**Middle**	**Mid-end.**	**Ending**
Comm Mgmt	284 (37%)	181 (24%)	109 (14%)	115 (15%)	71 (9%)
Info Provision	563 (15%)	760 (20%)	918 (24%)	937 (25%)	610 (16%)
Info Request	199 (22%)	253 (28%)	169 (19%)	188 (21%)	83 (9%)
Social Rel Mgmt	265 (22%)	139 (11%)	76 (6%)	161 (13%)	590 (48%)
Task Mgmt	93 (14%)	190 (29%)	135 (20%)	132 (20%)	109 (17%)

Table A.9: Distribution of functions by message position (librarian)

Label	**Beginning**	**Beg.-mid**	**Middle**	**Mid-end.**	**Ending**
Comm Mgmt	67 (16%)	92 (22%)	87 (21%)	103 (25%)	65 (16%)
Info Provision	806 (38%)	477 (23%)	401 (19%)	275 (13%)	145 (7%)
Info Request	37 (12%)	79 (25%)	99 (32%)	77 (25%)	22 (7%)
Social Rel Mgmt	124 (16%)	121 (15%)	121 (15%)	207 (26%)	212 (27%)
Task Mgmt	13 (12%)	13 (12%)	25 (23%)	36 (33%)	22 (20%)

Table A.10: Distribution of functions by message position (user)

	Comm Mgmt	**Info Provison**	**Info Request**	**Social Rel Mgmt**	**Task Mgmt**
Comm Mgmt	120* (10%)	520 (44%)	147 (13%)	214 (18%)	156 (13%)
Info Provision	513 (9%)	3369 (57%)	786 (13%)	608 (10%)	505 (9%)
Info Request	69 (6%)	754 (63%)	195 (16%)	98 (8%)	68 (6%)
Social Rel Mgmt	104 (5%)	649 (32%)	196 (10%)	680 (34%)	156 (8%)
Task Mgmt	126 (16%)	286 (37%)	99 (13%)	163 (21%)	68 (9%)
START**	8 (2%)	436 (90%)	11 (2%)	28 (6%)	3 (1%)

* The number of instances where dialogue acts on the left hand side of the row was followed by the dialogue act on the top of the column.

** START denotes that the dialogue acts on the top was at the begining of interview session.

Table A.11: Transition of Dialogue Act Functions

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
1=Comm:Channel	18*	11	1	0	5	7	10	12	0	10	2	1	7	2	8	2	4	0	7	1	4	0
2=Comm:Feedback	1	12	24	21	94	40	119	91	1	3	4	5	86	0	17	14	54	0	32	4	0	0
3=Comm:Pausing	5	30	17	2	34	91	45	94	1	1	4	0	35	6	4	1	41	0	23	7	0	0
4=Info:Feedback	2	7	2	25	45	15	55	23	3	3	0	3	23	0	6	0	9	0	9	2	1	0
5=Info:Obj	13	34	20	61	963	95	190	270	20	8	4	4	80	0	31	12	77	0	66	26	14	0
6=Info:Other	11	26	26	6	96	282	77	151	9	22	8	8	48	10	27	40	61	3	32	2	27	0
7=Info:Problem	28	49	201	27	236	190	1179	212	10	5	5	6	107	26	32	12	131	5	39	8	32	0
8=Info:Search	11	45	105	30	264	92	194	323	10	7	2	8	54	34	21	3	85	0	52	10	10	0
9=Social:Apology	3	0	1	3	7	25	17	18	2	4	1	0	2	0	9	3	2	1	2	0	1	0
10=Social:Closing Ritual	0	0	0	0	1	26	2	2	0	11	0	0	34	0	6	27	0	0	3	6	55	0
11=Social:Downplay	2	1	4	1	1	6	6	9	3	6	1	0	11	0	21	7	3	0	3	2	2	0
12=Social:Exclamation	0	3	2	9	9	6	20	18	0	2	1	2	13	0	10	1	10	0	3	2	1	0
13=Social:Gratitude	2	8	19	26	76	53	76	54	2	52	18	2	31	2	69	69	41	2	29	16	36	0
14=Social:Greeting	2	2	35	1	7	81	103	68	5	0	0	0	20	22	13	2	23	0	2	0	0	0
15=Social:Rapport	3	0	11	2	22	38	30	13	4	35	4	0	52	0	30	12	12	1	9	7	19	0
16=Social:Valediction	0	5	0	2	3	6	1	2	0	13	2	3	22	0	4	21	2	0	7	16	66	0
17=Task:Librarian	9	32	70	5	98	42	60	34	1	7	3	1	59	1	8	4	19	3	12	4	5	0
18=Task:Other	0	0	0	1	1	5	0	0	0	0	0	0	0	0	4	0	0	1	0	0	0	0
19=Task:User	1	11	4	7	56	20	19	37	4	10	0	3	38	0	7	11	22	0	15	2	15	0
20=NA	2	4	0	1	23	5	3	4	0	9	2	0	5	1	1	12	2	0	5	11	34	0
21=END**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22=START***	0	3	4	1	7	4	427	8	3	0	0	0	1	24	1	0	2	0	1	1	0	0

* The number of instances where dialogue acts on the left hand side of the row was followed by the dialogue act on the top of the column.

** END denotes that the dialogue acts on the left was at the end of interview session.

*** START denotes that the dialogue acts on the top was at the begining of interview session.

Table A.12: Transition of Dialogue Act Domains

Function	Avg.	Std. Dev.	Relative Std. Dev.
Info Provision	8.96	6.81	0.76
Info Request	8.32	4.28	0.51
Task Mgmt	8.99	4.33	0.48
Social Rel Mgmt	3.72	3.1	0.84
Comm Mgmt	3.31	3.19	0.96

Domain	Avg.	Std. Dev.	Relative Std. Dev.
Info:Problem	9.52	5.95	0.62
Info:Object	8.15	9.65	1.18
Info:Search	9.37	5.5	0.59
Info:Feedback	6.48	4.2	0.65
Info:Other	9.7	6.12	0.63
Task:Librarian	8.85	3.94	0.45
Task:User	9.25	5.08	0.55
Task:Other	9.33	2.87	0.31
Social:Gratitude	3.57	2.37	0.66
Social:Apology	3.16	2.74	0.87
Social:Downplay	2.46	0.95	0.38
Social:Greeting	2.2	1.7	0.77
Social:Valediction	1.81	1.37	0.75
Social:Closing	9.05	3.1	0.34
Social:Exclamation	1.32	0.81	0.62
Social:Rapport	5.8	3.81	0.66
Comm:Channel	5.01	3.28	0.65
Comm:Pausing	6.33	3.17	0.5
Comm:Feedback	1.07	0.3	0.28

Table A.13: Number of terms in text segment by label
* Top and bottom 1% were trimmed before the calculation.

Function		Common Terms
Info Provision	6030*	page (9%), session (6%), librarian (6%), library (4%), find (4%), joined (3%), information (3%), question (3%), email (3%), transcript (3%)
Info Request	1011	information (6%), library (5%), page (4%), find (3%), question (3%), search (3%), looked (3%), school (3%), give (2%), email (2%)
Task Mgmt	575	find (14%), check (8%), email (8%), send (7%), page (4%), session (4%), question (4%), search (4%), library (4%), information (3%)
Social Rel Mgmt	411	contact (5%), [PATRON NAME] (5%), service (5%), bye (4%), goodbye (4%), good (3%), askusnow (3%), maryland (3%), assistance (3%), great (3%)
Comm Mgmt	154	librarian (12%), hold (9%), moment (7%), minute (6%), minutes (3%), wait (3%), question (3%), reading (2%), good (1%), great (1%)

* The numbers in the second column are the numbers of unique terms.

Table A.14: Terms based on document frequency by function

Domain		Common Terms
Info:Problem	2530*	information (7%), find (6%), books (2%), question (2%), info (2%)
Info:Object	5394	page (21%), site (4%), library (3%), online (3%), information (3%)
Info:Search	1261	librarian (14%), session (13%), joined (13%), question (7%), search (6%)
Info:Feedback	211	good (8%), helpful (6%), work (4%), great (4%), site (4%)
Info:Other	856	email (15%), session (14%), transcript (14%), librarian (11%), address (11%)
Task:Librarian	368	find (21%), send (11%), check (9%), email (6%), question (5%)
Task:User	312	click (8%), email (8%), check (7%), type (6%), session (6%)
Social:Greeting	49*	[PATRON NAME] (23%), maryland (7%), askusnow (7%), patron (5%), reference (3%)
Social:Gratitude	132	service (10%), askusnow (5%), maryland (5%), reference (4%), 24/7 (3%)
Social:Rapport	185	good (13%), contact (11%), hope (10%), luck (7%), assistance (7%)
Social:Downplay	17	problem (18%), patient (3%), [PATRON NAME] (3%), alright (3%), fault (1%)
Social:Valediction	29	bye (47%), goodbye (37%), good (5%), cheers (1%), night (1%)
Social:Exclamation	34	great (27%), wow (8%), hmmm (5%), good (4%), awesome (3%)
Social:Closing	86	contact (42%), assistance (20%), free (20%), feel (20%), questions (13%)
Social:Apology	45	wrong (4%), long (4%), apologize (4%), time (3%), disconnected (3%)
Comm:Channel	47	patron (9%), heard (8%), [PATRON NAME] (7%), disconnected (7%), connection (6%)
Comm:Pausing	83	librarian (33%), hold (25%), moment (20%), minute (17%), minutes (9%)
Comm:Feedback	35	good (3%), great (3%), alright (1%), yeah (1%), correct (1%)

* The numbers in the second column are the numbers of unique terms.

Table A.15: Terms based on document frequency by domain

Function		Common Bi-grams
Info Provision	22403*	page sent (6%), i am (5%), the session (4%), sent - (4%), session (3%)
Info Request	3942	do you (15%), are you (10%), can you (7%), have you (5%), is this (5%)
Task Mgmt	2554	let me (20%), i can (15%), i will (15%), can find (11%), see what (10%)
Social Rel Mgmt	1505	thank you (17%), for using (8%), you for (8%), if you (6%), us again (5%)
Comm Mgmt	629	will be (11%), with you (10%), be with (10%), you in (8%), librarian will (7%)

* The numbers in the second column are the numbers of unique bi-grams.

Table A.16: Term bi-grams based on document frequency by function

Domain		Common Bi-grams
Info:Problem	9684*	do you (6%), i am (4%), i need (4%), looking for (4%), can you (4%)
Info:Object	15048	page sent (16%), sent - (10%), here is (4%), of the (4%), is a (3%)
Info:Search	5242	joined the (13%), session (13%), the session (13%), has joined (13%), a librarian (11%)
Info:Other	3506	if you (10%), to you (9%), a transcript (8%), this session (8%), transcript of (8%)
Task:Librarian	1613	let me (27%), i can (24%), can find (18%), see what (15%), i will (15%)
Task:User	1210	if you (16%), i will (15%), let me (8%), you can (7%), you need (7%)
Task:Other	60	email you (44%), you with (33%), the librarians (22%), will get (22%), to your (22%)
Social:Gratitude	377	thank you (49%), you for (22%), for using (21%), thanks for (13%), service (9%)
Social:Greeting	129	[PATRON NAME] (21%), welcome to (17%), hi [PATRON NAME] (14%), hello [PATRON NAME] (8%), maryland askusnow (7%)
Social:Apology	159	i'm sorry (12%), i am (8%), i apologize (4%), for the (4%), sorry we (3%)
Social:Valediction	72	goodbye (36%), bye (27%), bye for (18%), for now (18%), good bye (4%)
Social:Closing	301	if you (45%), us again (43%), contact us (41%), you need (28%), please contact (27%)
Social:Rapport	710	if you (15%), us again (11%), you need (11%), contact us (10%), need further (8%)
Social:Downplay	62	you're welcome (48%), problem (14%), no problem (13%), not a (6%), a problem (6%)
Social:Exclamation	60	great (27%), wow (6%), hmmm (5%), awesome (3%), this is (2%)
Comm:Channel	206	are you (31%), you still (22%), still there (20%), you there (9%), i haven't (9%)
Comm:Pausing	369	will be (30%), with you (27%), be with (27%), librarian will (19%), you in (19%)
Comm:Feedback	72	great (3%), good (3%), i see (1%), alright (1%), yeah (1%)

* The numbers in the second column are the numbers of unique bi-grams.

Table A.17: Term bi-grams based on document frequency by domain

Function		Commmon Bi-grams at Sentence Beginning
Info Provision	2050	page sent (6%), i am (3%), yes (2%), if you (2%), this is (2%), i have (2%), here is (1%), i need (1%), i'm reading (1%), i think (1%),
Info Request	342	do you (10%), are you (6%), can you (6%), did you (4%), is this (4%), have you (3%), is there (3%), would you (2%), is that (2%), can i (2%),
Task Mgmt	214	let me (17%), i will (12%), if you (4%), i am (3%), i'll see (3%), while i (2%), i'm going (2%), you can (2%), i can (2%), we will (1%),
Social Rel Mgmt	349	thank you (15%), thanks (8%), hello (5%), if you (5%), thanks for (4%), hi (3%), goodbye (3%), hi [PATRON NAME] (3%), you're welcome (2%), sorry (2%),
Comm Mgmt	170	ok (28%), a librarian (7%), yes (6%), please hold (6%), okay (6%), are you (2%), please wait (2%), just a (2%), it will (2%), one moment (2%),

* The numbers in the second column are the numbers of unique bi-grams at sentence beginnings.

Table A.18: First two words of text segments by function

Domain		Commong Bi-grams at Sentence Beginning
Info:Problem	846*	do you (4%), i am (3%), can you (3%), yes (3%), i need (3%)
Info:Object	945	page sent (16%), here is (4%), this is (1%), there is (1%), here's a (1%)
Info:Search	541	i am (5%), i'm reading (4%), 24/7 librarian (3%), i'm looking (2%), did you (2%)
Info:Feedback	141	this is (6%), do you (3%), yes (3%), is that (3%), does that (2%)
Info:Other	378	if you (9%), this is (5%), i am (4%), do you (3%), yes (2%)
Task:Librarian	116	let me (25%), i will (11%), i am (4%), i'll see (4%), while i (4%)
Task:Other	8	we will (22%), someone from (11%), and they (11%), let me (11%), the librarians (11%)
Task:User	116	i will (13%), if you (8%), let me (5%), you can (4%), you may (3%)
Social:Apology	31	sorry (39%), i'm sorry (12%), i am (6%), i apologize (4%), sorry about (3%)
Social:Closing	43	if you (37%), have a (6%), goodbye, and (5%), it's been (5%), goodbye and (4%)
Social:Downplay	21	you're welcome (48%), no problem (13%), not a (6%), your welcome (4%), you are (4%)
Social:Exclamation	47	great (25%), well (8%), wow (6%), hmmm (5%), awesome (3%)
Social:Gratitude	61	thank you (45%), thanks (22%), thanks for (11%), and thank (2%), thanks so (2%)
Social:Greeting	37	hello (27%), hi (18%), hi [PATRON NAME] (14%), hello [PATRON NAME] (9%), welcome to (7%)
Social:Rapport	134	if you (12%), i hope (7%), have a (5%), good luck (5%), :) (2%)
Social:Valediction	28	goodbye (36%), bye (23%), bye for (17%), good bye (4%), bye now (2%)
Comm:Channel	43	are you (22%), i haven't (9%), hello (7%), can you (4%), r u (3%)
Comm:Pausing	69	a librarian (19%), please hold (16%), just a (6%), please wait (6%), it will (5%)
Comm:Feedback	62	ok (52%), yes (11%), okay (10%), good (3%), sure (2%)

* The numbers in the second column are the numbers of unique bi-grams at sentence beginnings.

Table A.19: First two words of text segments by function

Function		**Commmon Tri-grams at Sentence Beginning**
Info Provision	3007*	page sent - (4%), yes (2%), if you would (1%), this is the (1%), i'm reading your (1%), here is a (0%), i am looking (0%), i'm looking at (0%), i am not (0%), if you provided (0%),
Info Request	567	do you have (3%), can you tell (2%), would you like (2%), do you know (1%), is there anything (1%), is this for (1%), did you get (1%), are you looking (1%), are you able (1%), do you think (1%),
Task Mgmt	376	let me see (7%), let me check (3%), i'm going to (2%), i'll see what (2%), i am going (2%), if you would (2%), let me know (2%), we will email (1%), let me look (1%), i will send (1%),
Social Rel Mgmt	460	thanks (8%), thank you for (7%), thank you (5%), hello (5%), hi (3%), hi [PATRON NAME] (3%), if you need (3%), goodbye (3%), you're welcome (2%), sorry (2%),
Comm Mgmt	208	ok (28%), a librarian will (7%), yes (6%), okay (6%), please hold for (5%), it will be (2%), are you still (2%), good (1%), just a moment (1%), great (1%),

* The numbers in the second column are the numbers of unique bi-grams at sentence beginnings.

Table A.20: First three words of text segments by function

Domain		**Commmon Tri-grams at Sentence Beginning**
Info:Problem	1226*	yes (3%), do you have (2%), can you tell (1%), do you know (1%), i am trying (1%)
Info:Object	1241	page sent - (10%), here is a (1%), here is another (1%), here is the (1%), yes (1%)
Info:Search	782	i'm reading your (3%), yes (2%), i'm looking at (2%), 24/7 librarian [LIBRARIAN INITIALS] (1%), 24/7 librarian [NAME] (1%)
Info:Feedback	166	yes (3%), do you think (3%), this is good (2%), what do you (2%), is that what (2%)
Info:Other	527	if you would (5%), this is the (4%), yes (2%), if you provided (2%), my name is (2%)
Task:User	172	if you would (4%), let me know (3%), if you need (2%), you need to (2%), please let me (2%)
Task:Librarian	215	let me see (12%), let me check (4%), i'm going to (3%), i'll see what (3%), i am going (2%)
Task:Other	8	we will email (22%), someone will get (11%), let me know (11%), i meant you (11%), someone from uw (11%)
Social:Greeting	49	hello (27%), hi [PATRON NAME] (18%), hi (18%), hello [PATRON NAME] (8%), hi [NAME] (3%)
Social:Valediction	32	goodbye (36%), bye (23%), bye for now (16%), good bye (3%), bye bye (1%),
Social:Rapport	164	if you need (9%), i hope this (4%), good luck with (3%), have a good (3%), if you have (2%),
Social:Gratitude	95	thanks (22%), thank you for (20%), thank you (16%), thank you so (4%), thanks for using (4%)
Social:Apology	41	sorry (39%), i'm sorry (9%), i am sorry (3%), sorry you were (1%), (oh sorry (1%)
Social:Downplay	27	you're welcome (48%), no problem (8%), not a problem (6%), i am very (3%), you are welcome (3%)
Social:Closing	52	if you need (25%), if you have (12%), goodbye and thank (5%), it's been a (5%), goodbye and thanks (4%)
Social:Exclamation	48	great (25%), well (8%), wow (6%), hmmm (5%), oh (3%)
Comm:Channel	50	are you still (17%), hello (7%), i haven't heard (7%), are you there (6%), [PATRON NAME] (3%)
Comm:Pausing	99	a librarian will (19%), please hold for (13%), it will be (5%), just a moment (4%), please wait (3%)
Comm:Feedback	62	ok (52%), yes (11%), okay (10%), good (3%), sure (2%)

* The numbers in the second column are the numbers of unique tri-grams at sentence beginnings.

Table A.21: First three words of text segments by domain

Label	!!	-	!?	none	.	...	!	:-)*	?	:**	??	,
Info Provision	2	60	0	2127	1618	200	15	1	159	129	1	163
Info Request	0	1	0	107	38	8	0	0	776	0	4	4
Task Mgmt	0	9	0	184	315	59	7	0	19	12	0	16
Social Rel Mgmt	12	34	0	529	539	36	229	14	5	4	0	196
Comm Mgmt	0	14	0	358	326	40	1	0	55	1	2	139
Total	14	118	0	3378	2843	346	252	15	1018	146	9	518

Table A.22: Distribution of punctuation by function

* “:-)” included different forms of emoticons e.g. “:-)”, “:)”, “:P”, etc.

** “:” included a colon “:” and a semicolon “;”.

Label	!!	-	!?	none	.	...	!	:-)*	?	:**	??	,
Info:Problem	1	8	0	606	414	52	4	1	643	4	3	52
Info:Object	0	25	0	1009	370	58	1	0	41	84	1	35
Info:Search	0	13	0	367	478	83	4	0	112	30	1	37
Info:Feedback	1	3	0	73	43	7	6	0	41	3	0	17
Info:Other	0	13	0	228	423	24	1	0	108	12	1	32
Task:Librarian	0	4	0	110	201	42	2	0	12	7	0	9
Task:User	0	5	0	72	111	18	5	0	8	5	0	7
Task:Other	0	0	0	3	6	0	0	0	0	0	0	0
Social:Apology	1	3	0	30	23	2	1	0	0	0	0	17
Social:Closing	0	1	0	23	113	2	19	0	0	0	0	0
Social:Downplay	0	3	0	22	33	3	4	1	0	0	0	5
Social:Exclamation	3	2	0	28	16	5	19	0	1	0	0	19
Social:Gratitude	6	9	0	247	159	13	78	2	0	1	0	28
Social:Greeting	0	7	0	66	79	2	44	0	1	3	0	106
Social:Rapport	1	2	0	66	122	10	33	11	3	0	0	12
Social:Valediction	2	7	0	54	23	1	37	0	0	0	0	11
Comm:Channel	0	1	0	19	16	0	0	0	48	0	2	3
Comm:Pausing	0	3	0	74	234	18	0	0	6	0	0	7
Comm:Feedback	0	10	0	266	75	21	1	0	2	1	0	130
Total	15	119	0	3363	2939	361	259	15	1026	150	8	527

Table A.23: Distribution of punctuation by domain

* “:-)” included different forms of emoticons e.g. “:-)”, “:)”, “:P”, etc.

** “:” included a colon “:” and a semicolon “;”.

	Output:							
	1	**2**	**3**	**4**	**5**	**Size**	**Correct**	**Wrong**
1 = Information Request	2	656	5	5	35	54	2	701
2 = Information Provision	13	3424	0	8	12	5343	3424	33
3 = Task Mgmt	0	395	12	1	3	17	12	399
4 = Communication Mgmt	7	377	0	238	21	261	238	405
5 = Social Rel Mgmt	32	491	0	9	756	827	756	532
Total						6502	4432	2070

Table A.24: Confusion Matrix for S-16 (Function)

	Output:							
	1	**2**	**3**	**4**	**5**	**Size**	**Correct**	**Wrong**
1 = Information Request	333	410	7	15	13	440	333	445
2 = Information Provision	63	3543	33	29	26	4625	3543	151
3 = Task Mgmt	17	208	190	11	10	247	190	246
4 = Communication Mgmt	19	194	6	530	52	595	530	271
5 = Social Rel Mgmt	8	270	11	10	989	1090	989	299
Total						6997	5585	1412

Table A.25: Confusion Matrix for H-16 (Function)

	Output:							
	1	**2**	**3**	**4**	**5**	**Size**	**Correct**	**Wrong**
1 = Information Request	370	502	7	17	8	484	370	534
2 = Information Provision	74	3744	38	21	41	4944	3744	174
3 = Task Mgmt	14	228	168	10	10	221	168	262
4 = Communication Mgmt	18	197	6	474	52	535	474	273
5 = Social Rel Mgmt	8	273	2	13	1010	1121	1010	296
Total						7305	5766	1539

Table A.26: Confusion Matrix for H-17 (Function)

	Output:							
	1	**2**	**3**	**4**	**5**	**Size**	**Correct**	**Wrong**
1 = Information Request	331	401	4	15	15	433	331	435
2 = Information Provision	60	3653	28	19	26	4716	3653	133
3 = Task Mgmt	11	217	188	15	13	229	188	256
4 = Communication Mgmt	21	189	7	504	26	563	504	243
5 = Social Rel Mgmt	10	256	2	10	1036	1116	1036	278
Total						7057	5712	1345

Table A.27: Confusion Matrix for H-18 (Function)

	Output:							
	1	**2**	**3**	**4**	**5**	**Size**	**Correct**	**Wrong**
Correct:								
1 = Information Request	319	344	8	14	28	433	319	394
2 = Information Provision	61	3439	21	11	50	4369	3439	143
3 = Task Mgmt	17	186	184	12	8	231	184	223
4 = Communication Mgmt	29	168	11	498	66	550	498	274
5 = Social Rel Mgmt	7	232	7	15	991	1143	991	261
Total						6726	5431	1295

Table A.28: Confusion Matrix for H-20 (Function)

	Output:							
	1	**2**	**3**	**4**	**5**	**Size**	**Correct**	**Wrong**
1 = Information Request	290	383	11	15	12	430	290	421
2 = Information Provision	81	3632	31	32	30	4726	3632	174
3 = Task Mgmt	11	254	210	7	15	266	210	287
4 = Communication Mgmt	33	180	7	488	53	551	488	273
5 = Social Rel Mgmt	15	277	7	9	1140	1250	1140	308
Total						7223	5760	1463

Table A.29: Confusion Matrix for H-24 (Function)

	Output:							
	1	**2**	**3**	**4**	**5**	**Size**	**Correct**	**Wrong**
1 = Information Request	485	282	10	4	3	601	485	299
2 = Information Provision	64	3748	38	27	10	4613	3748	139
3 = Task Mgmt	4	176	291	8	10	352	291	198
4 = Communication Mgmt	38	161	8	563	45	624	563	252
5 = Social Rel Mgmt	10	246	5	22	1122	1190	1122	283
Total						7380	6209	1171

Table A.30: Confusion Matrix for H-48 (Function)

	Output:							
	1	**2**	**3**	**4**	**5**	**Size**	**Correct**	**Wrong**
1 = Information Request	458	248	10	2	7	543	458	267
2 = Information Provision	39	3445	24	20	10	4238	3445	93
3 = Task Mgmt	6	179	271	4	10	317	271	199
4 = Communication Mgmt	28	137	5	518	29	561	518	199
5 = Social Rel Mgmt	12	229	7	17	1069	1125	1069	265
Total						6784	5761	10

Table A.31: Confusion Matrix for H-54 (Function)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Info Request	0.0028	0.0090	0.0370	0.0028	0.0053
2=Info Provision	0.9905	0.6302	0.6408	0.9905	0.7782
3=Task Mgmt	0.0292	0.0008	0.7059	0.0292	0.0561
4=Comm Mgmt	0.3701	0.0039	0.9119	0.3701	0.5265
5=Social Rel Mgmt	0.5870	0.0136	0.9141	0.5870	0.7149
Weighted Average	0.6816	0.3392	0.6606	0.6816	0.6115

Table A.32: Measurements for S-16 (Function)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Info Request	0.4280	0.0172	0.7568	0.4280	0.5468
2=Info Provision	0.9591	0.3276	0.7661	0.9591	0.8518
3=Task Mgmt	0.4358	0.0087	0.7692	0.4358	0.5564
4=Comm Mgmt	0.6617	0.0105	0.8908	0.6617	0.7593
5=Social Rel Mgmt	0.7679	0.0177	0.9073	0.7679	0.8318
Weighted Average	0.7982	0.1799	0.8055	0.7982	0.7852

Table A.33: Measurements for H-16 (Function)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Info Request	0.4093	0.0178	0.7645	0.4093	0.5331
2=Info Provision	0.9556	0.3543	0.7573	0.9556	0.8450
3=Task Mgmt	0.3907	0.0077	0.7602	0.3907	0.5161
4=Comm Mgmt	0.6345	0.0093	0.8860	0.6345	0.7395
5=Social Rel Mgmt	0.7734	0.0185	0.9010	0.7734	0.8323
Weighted Average	0.7893	0.1969	0.7972	0.7893	0.7740

Table A.34: Measurements for H-17 (Function)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Info Request	0.4321	0.0162	0.7644	0.4321	0.5521
2=Info Provision	0.9649	0.3250	0.7746	0.9649	0.8593
3=Task Mgmt	0.4234	0.0062	0.8210	0.4234	0.5587
4=Comm Mgmt	0.6747	0.0094	0.8952	0.6747	0.7695
5=Social Rel Mgmt	0.7884	0.0139	0.9283	0.7884	0.8527
Weighted Average	0.8094	0.1801	0.8178	0.8094	0.7963

Table A.35: Measurements for H-18 (Function)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Info Request	0.4474	0.0190	0.7367	0.4474	0.5567
2=Info Provision	0.9601	0.2958	0.7871	0.9601	0.8650
3=Task Mgmt	0.4521	0.0074	0.7965	0.4521	0.5768
4=Comm Mgmt	0.6451	0.0087	0.9055	0.6451	0.7534
5=Social Rel Mgmt	0.7915	0.0278	0.8670	0.7915	0.8276
Weighted Average	0.8075	0.1662	0.8108	0.8075	0.7951

Table A.36: Measurements for H-20 (Function)

	TP Rate	FP Rate	Precision	Recall	F-Measure
1=Info Request	0.4079	0.0215	0.6744	0.4079	0.5083
2=Info Provision	0.9543	0.3202	0.7685	0.9543	0.8514
3=Task Mgmt	0.4225	0.0083	0.7895	0.4225	0.5505
4=Comm Mgmt	0.6413	0.0097	0.8857	0.6413	0.7439
5=Social Rel Mgmt	0.7873	0.0190	0.9120	0.7873	0.8451
Weighted Average	0.7975	0.1762	0.8018	0.7975	0.7843

Table A.37: Measurements for H-24 (Function)

	TP Rate	FP Rate	Precision	Recall	F-Measure
1=Info Request	0.6186	0.0176	0.8070	0.6186	0.7004
2=Info Provision	0.9642	0.2476	0.8125	0.9642	0.8819
3=Task Mgmt	0.5951	0.0089	0.8267	0.5951	0.6920
4=Comm Mgmt	0.6908	0.0093	0.9022	0.6908	0.7825
5=Social Rel Mgmt	0.7986	0.0114	0.9429	0.7986	0.8647
Weighted Average	0.8413	0.1361	0.8476	0.8413	0.8358

Table A.38: Measurements for H-48 (Function)

	TP Rate	FP Rate	Precision	Recall	F-Measure
1=Info Request	0.6317	0.0140	0.8435	0.6317	0.7224
2=Info Provision	0.9737	0.2443	0.8129	0.9737	0.8861
3=Task Mgmt	0.5766	0.0073	0.8549	0.5766	0.6887
4=Comm Mgmt	0.7225	0.0071	0.9234	0.7225	0.8106
5=Social Rel Mgmt	0.8013	0.0103	0.9502	0.8013	0.8695
Weighted Average	0.8492	0.1322	0.8577	0.8492	0.8437

Table A.39: Measurements for H-54 (Function)

	Output:																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**	**TP**	**FP**
1=Comm:Channel	4	1	0	0	3	5	5	1	0	1	0	0	28	4	0	0	11	0	63	4	59
2=Comm:Feedback	2	273	0	2	7	4	69	6	0	0	0	7	0	0	12	0	6	0	388	273	115
3=Comm:Pausing	0	3	221	0	1	9	6	0	0	1	0	0	6	0	0	0	15	2	264	221	43
4=Info:Feedback	0	1	6	0	5	43	24	0	0	0	0	0	7	0	1	0	19	0	106	0	106
5=Info:Obj	0	1	78	3	791	219	34	3	0	1	0	0	31	0	2	0	72	0	1235	791	444
6=Info:Other	2	3	57	0	43	252	51	6	0	2	0	0	60	7	0	0	140	0	623	252	371
7=Info:Problem	0	2	110	0	107	296	290	6	1	5	0	0	203	2	1	0	376	1	1400	290	1110
8=Info:Search	4	2	43	0	58	239	66	21	0	5	0	0	88	0	0	3	345	1	875	21	854
9=Social:Apology	0	0	0	0	3	2	6	0	24	0	0	0	1	0	0	0	24	0	60	24	36
10=Social:Closing Ritual	0	0	10	0	3	8	0	0	0	84	0	0	17	0	0	0	4	0	126	84	42
11=Social:Downplay	0	2	6	0	0	5	8	0	0	0	2	0	41	5	1	0	0	0	70	2	68
12=Social:Exclamation	0	9	4	0	2	1	21	0	0	0	0	36	4	0	0	0	4	0	81	36	45
13=Social:Gratitude	0	0	0	0	0	1	2	0	1	0	0	0	425	0	0	0	4	0	433	425	8
14=Social:Greeting	0	0	0	0	0	10	2	1	0	0	0	0	0	229	6	0	2	0	250	229	21
15=Social:Rapport	0	1	44	0	8	22	21	0	0	14	0	0	29	2	10	0	44	1	196	10	186
16=Social:Valediction	0	0	0	0	6	0	3	0	0	0	0	0	16	0	0	75	5	0	105	75	30
17=Task:Librarian	0	0	11	0	0	16	1	0	0	0	0	0	5	0	0	0	221	0	254	221	33
19=Task:User	1	0	11	0	9	23	3	0	0	7	0	0	19	0	0	0	71	1	145	1	144
Total																			6674	2959	3715

Table A.40: Confusion Matrix for S-16 (Domain)

	Output:																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**	**TP**	**FP**
1=Comm:Channel	12	2	1	0	3	8	24	6	0	0	0	0	1	6	0	0	2	0	65	12	53
2=Comm:Feedback	0	324	3	0	5	4	32	6	0	0	1	6	5	3	2	0	7	0	398	324	74
3=Comm:Pausing	0	3	250	0	4	0	4	1	0	0	0	0	1	0	0	0	3	0	266	250	16
4=Info:Feedback	0	3	1	1	31	4	54	5	0	0	0	2	1	0	2	0	2	0	106	1	105
5=Info:Obj	0	6	9	0	830	32	160	51	0	0	0	0	3	2	0	0	1	0	1094	830	264
6=Info:Other	1	13	16	0	45	316	193	60	0	7	0	0	3	7	1	1	8	0	671	316	355
7=Info:Problem	0	16	33	1	67	27	1214	68	0	1	0	0	7	7	1	0	11	0	1453	1214	239
8=Info:Search	1	5	10	0	73	33	233	419	0	6	0	0	3	7	1	0	28	3	822	419	403
9=Social:Apology	0	0	1	0	0	0	3	1	33	0	0	0	0	0	0	0	0	0	38	33	5
10=Social:Closing Ritual	0	0	4	0	0	7	4	2	0	82	0	0	17	0	4	1	0	0	121	82	39
11=Social:Downplay	0	2	0	0	0	0	4	3	0	0	32	0	0	2	0	0	0	0	43	32	11
12=Social:Exclamation	0	19	3	0	3	0	19	3	0	0	0	12	2	3	1	0	0	0	65	12	53
13=Social:Gratitude	0	1	0	0	0	1	2	5	2	1	0	0	400	0	0	0	0	0	412	400	12
14=Social:Greeting	0	2	0	0	0	1	1	0	0	0	0	0	2	220	0	0	0	0	226	220	6
15=Social:Rapport	0	8	11	0	12	6	66	7	0	5	0	0	8	4	42	0	2	0	171	42	129
16=Social:Valediction	0	0	0	0	0	2	2	2	0	0	0	0	2	4	2	100	1	0	115	100	15
17=Task:Librarian	0	1	12	0	1	24	20	25	0	1	0	0	0	0	0	0	207	2	293	207	86
19=Task:User	1	2	5	0	14	19	26	18	0	10	0	0	0	1	0	0	51	3	150	3	147
Total																			6509	4497	2012

Table A.41: Confusion Matrix for H-16 (Domain)

	Output:																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**	**TP**	**FP**
1=Comm:Channel	11	4	0	0	4	16	21	10	0	0	0	0	1	10	0	0	0	0	77	11	66
2=Comm:Feedback	0	297	1	1	9	2	30	6	3	1	0	2	9	4	7	2	8	0	382	297	85
3=Comm:Pausing	0	5	232	0	5	2	7	8	1	0	0	0	0	0	0	0	3	0	263	232	31
4=Info:Feedback	0	3	2	5	20	2	38	6	1	0	1	1	8	0	5	0	2	0	94	5	89
5=Info:Obj	2	8	6	3	895	24	189	31	0	1	0	0	6	0	1	0	8	2	1176	895	281
6=Info:Other	0	16	8	1	50	331	162	82	1	4	4	0	10	10	2	0	11	3	695	331	364
7=Info:Problem	0	19	25	7	76	23	1192	72	0	3	0	2	22	7	3	2	11	2	1466	1192	274
8=Info:Search	1	12	13	0	85	35	226	439	0	9	0	0	24	1	5	2	28	1	881	439	442
9=Social:Apology	0	2	0	0	0	1	5	0	49	0	1	0	3	0	0	0	0	0	61	49	12
10=Social:Closing Ritual	0	3	2	0	2	6	5	3	0	87	0	0	11	0	11	2	0	0	132	87	45
11=Social:Downplay	0	2	0	1	1	0	4	0	0	0	39	0	0	0	0	0	0	0	47	39	8
12=Social:Exclamation	0	17	0	1	6	0	17	8	0	0	0	19	11	2	2	2	0	0	85	19	66
13=Social:Gratitude	0	0	0	0	0	0	6	2	1	1	0	0	417	0	0	0	0	0	427	417	10
14=Social:Greeting	0	0	0	0	0	0	1	0	0	0	0	0	2	247	0	0	0	0	250	247	3
15=Social:Rapport	1	7	4	5	6	2	51	19	0	13	0	0	13	3	56	3	3	1	187	56	131
16=Social:Valediction	0	1	0	0	0	5	3	3	0	0	0	0	5	2	0	112	0	1	132	112	20
17=Task:Librarian	0	1	5	0	1	21	19	27	0	0	0	0	0	0	0	0	195	2	271	195	76
19=Task:User	0	1	6	2	19	24	30	21	0	10	0	0	1	0	1	1	47	11	174	11	163
Total																			6800	4634	2166

Table A.42: Confusion Matrix for H-17 (Domain)

	Output:																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**	**TP**	**FP**
1=Comm:Channel	14	5	1	0	1	12	21	7	0	0	0	0	0	2	1	0	3	0	67	14	53
2=Comm:Feedback	0	297	2	0	12	4	65	2	0	0	0	5	5	2	4	0	11	0	409	297	112
3=Comm:Pausing	0	6	248	0	2	0	8	1	0	0	0	0	0	0	0	0	7	0	272	248	24
4=Info:Feedback	0	2	0	5	9	1	89	2	0	0	1	0	1	0	0	0	2	0	112	5	107
5=Info:Obj	2	4	2	1	1044	18	148	33	0	0	0	0	5	4	0	0	7	0	1268	1044	224
6=Info:Other	0	9	18	0	57	344	201	70	0	4	1	0	1	9	0	0	15	0	729	344	385
7=Info:Problem	0	13	9	1	73	11	1289	42	0	1	0	0	3	9	2	0	7	2	1462	1289	173
8=Info:Search	0	5	13	0	93	34	212	442	0	6	0	0	5	10	4	0	44	5	873	442	431
9=Social:Apology	0	1	4	0	0	1	7	2	37	0	0	0	0	0	0	0	0	0	52	37	15
10=Social:Closing Ritual	0	0	4	0	2	4	11	3	0	95	0	0	8	0	3	3	0	0	133	95	38
11=Social:Downplay	0	1	0	0	1	0	3	0	0	0	33	0	0	2	0	0	0	0	40	33	7
12=Social:Exclamation	0	19	2	0	10	0	30	1	0	0	0	13	0	2	1	0	0	0	78	13	65
13=Social:Gratitude	0	0	0	0	0	2	13	0	0	0	0	0	410	0	1	0	0	0	426	410	16
14=Social:Greeting	0	1	0	0	0	0	3	0	0	0	0	0	1	253	0	0	0	0	258	253	5
15=Social:Rapport	0	2	3	0	22	2	77	14	0	10	0	0	9	5	46	2	6	0	198	46	152
16=Social:Valediction	0	2	0	0	0	2	6	1	0	0	0	0	0	2	0	101	0	0	114	101	13
17=Task:Librarian	0	2	7	0	1	29	21	20	0	0	0	0	0	1	0	0	186	2	269	186	83
19=Task:User	0	1	7	0	17	14	41	12	0	7	0	0	1	1	0	0	36	9	146	9	137
Total																			6906	4866	2040

Table A.43: Confusion Matrix for H-18 (Domain)

	Output:																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**	**TP**	**FP**
1=Comm:Channel	20	2	0	0	7	9	21	5	0	0	0	0	1	11	0	0	4	0	80	20	60
2=Comm:Feedback	0	330	1	0	7	5	34	2	0	0	0	7	2	6	3	0	3	0	400	330	70
3=Comm:Pausing	0	1	251	0	0	2	4	6	0	0	0	0	3	0	0	0	10	1	278	251	27
4=Info:Feedback	0	3	1	3	35	1	59	2	0	0	0	0	1	0	2	0	4	0	111	3	108
5=Info:Obj	0	5	7	2	972	31	204	41	0	1	0	0	3	1	0	0	7	1	1275	972	303
6=Info:Other	3	25	10	0	56	378	175	57	0	2	0	0	2	10	0	0	16	1	735	378	357
7=Info:Problem	1	28	26	0	82	21	1226	52	0	3	0	0	13	3	1	1	13	1	1471	1226	245
8=Info:Search	3	8	7	0	88	47	252	438	0	6	0	0	1	7	1	0	42	1	901	438	463
9=Social:Apology	0	3	1	0	0	1	9	3	41	0	1	0	1	0	0	0	0	0	60	41	19
10=Social:Closing Ritual	0	2	3	0	3	4	2	2	0	85	0	0	16	0	3	4	0	0	124	85	39
11=Social:Downplay	0	4	0	0	0	3	8	1	0	0	33	0	1	2	0	0	0	0	52	33	19
12=Social:Exclamation	0	35	4	1	3	0	17	5	0	0	0	13	7	3	1	0	1	0	90	13	77
13=Social:Gratitude	0	1	0	0	0	3	2	5	0	1	0	0	449	3	0	0	0	0	464	449	15
14=Social:Greeting	0	1	0	0	0	5	0	0	0	0	0	0	1	230	3	0	0	0	240	230	10
15=Social:Rapport	0	5	2	0	12	5	68	22	0	14	0	0	10	2	44	0	2	1	187	44	143
16=Social:Valediction	0	2	0	0	0	1	3	0	0	0	0	0	2	2	0	115	1	0	126	115	11
17=Task:Librarian	0	2	6	0	2	30	29	29	0	0	0	0	0	0	0	0	216	4	318	216	102
19=Task:User	2	1	10	0	17	25	31	21	0	12	0	0	0	0	0	0	60	11	190	11	179
Total																			7102	4855	2247

Table A.44: Confusion Matrix for H-20 (Domain)

	Output:																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**	**TP**	**FP**
1=Comm:Channel	11	2	0	0	2	9	16	4	0	0	0	0	1	3	1	0	0	0	49	11	38
2=Comm:Feedback	0	298	0	0	8	4	23	5	0	0	0	2	8	4	5	2	6	0	365	298	67
3=Comm:Pausing	0	0	235	0	5	2	4	0	0	0	0	0	5	0	0	0	2	0	253	235	18
4=Info:Feedback	0	5	3	1	36	6	32	2	0	0	0	3	1	0	11	0	1	0	101	1	100
5=Info:Obj	0	3	8	0	853	16	136	50	0	2	1	0	10	2	2	0	5	1	1089	853	236
6=Info:Other	0	23	10	0	49	337	130	77	0	3	3	0	10	13	4	2	8	2	671	337	334
7=Info:Problem	0	42	22	2	72	27	1212	66	1	3	0	0	11	4	9	6	10	2	1489	1212	277
8=Info:Search	0	21	13	0	70	50	204	411	0	4	0	0	4	2	8	1	29	0	817	411	406
9=Social:Apology	0	2	2	0	0	0	3	3	35	0	3	0	0	0	0	0	0	0	48	35	13
10=Social:Closing Ritual	0	0	1	0	1	5	0	2	0	85	0	0	14	0	10	5	0	1	124	85	39
11=Social:Downplay	0	2	0	0	2	1	4	1	0	0	46	0	0	3	1	0	0	0	60	46	14
12=Social:Exclamation	0	26	0	0	5	0	10	4	0	0	0	24	8	3	3	0	0	0	83	24	59
13=Social:Gratitude	0	1	0	0	2	0	2	7	0	1	0	0	476	1	0	0	0	0	490	476	14
14=Social:Greeting	0	0	0	0	0	1	4	0	0	0	0	0	0	243	0	0	0	0	248	243	5
15=Social:Rapport	0	7	1	0	13	5	53	11	0	14	1	1	18	3	56	3	6	0	192	56	136
16=Social:Valediction	0	0	0	0	0	2	3	3	0	0	0	0	3	2	0	91	0	0	104	91	13
17=Task:Librarian	0	0	9	0	2	35	26	26	0	0	0	0	2	0	0	0	207	4	311	207	104
19=Task:User	0	3	5	0	18	25	27	21	0	11	0	0	1	0	0	1	49	6	167	6	161
Total																			6661	4627	2034

Table A.45: Confusion Matrix for H-24 (Domain)

	Output:																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**	**TP**	**FP**
1=Comm:Channel	25	2	0	0	3	12	25	2	0	0	0	0	0	7	4	0	2	0	82	25	57
2=Comm:Feedback	0	298	1	0	3	5	43	2	0	0	1	9	1	11	2	0	3	0	379	298	81
3=Comm:Pausing	0	2	216	0	0	1	4	2	0	0	0	0	2	0	0	0	0	0	227	216	11
4=Info:Feedback	0	2	0	3	30	12	57	8	0	0	0	1	0	0	2	0	5	0	120	3	117
5=Info:Obj	2	2	7	1	880	42	142	39	0	0	0	0	3	0	1	0	6	0	1125	880	245
6=Info:Other	2	14	11	0	43	383	181	41	0	4	1	0	0	3	0	2	7	1	693	383	310
7=Info:Problem	0	27	19	0	75	27	1275	56	1	1	2	0	5	6	3	0	11	0	1508	1275	233
8=Info:Search	1	11	7	0	47	39	189	451	0	6	0	0	0	6	6	1	25	3	792	451	341
9=Social:Apology	0	2	2	0	1	0	5	2	29	0	0	0	0	0	0	0	0	1	42	29	13
10=Social:Closing Ritual	0	0	8	0	1	3	1	5	0	99	0	0	7	0	9	6	0	0	139	99	40
11=Social:Downplay	0	0	0	0	1	0	7	1	0	0	43	0	0	4	0	0	0	0	56	43	13
12=Social:Exclamation	0	16	2	0	2	0	27	2	0	0	0	31	0	1	2	3	0	0	86	31	55
13=Social:Gratitude	0	3	0	0	0	4	13	1	2	0	4	0	420	0	0	0	0	0	447	420	27
14=Social:Greeting	0	4	0	0	0	0	0	0	0	0	0	0	0	251	0	0	0	0	255	251	4
15=Social:Rapport	0	7	5	2	5	2	60	6	0	14	2	3	6	3	53	1	4	1	174	53	121
16=Social:Valediction	0	0	0	0	0	0	0	1	0	0	0	0	2	0	1	117	0	0	121	117	4
17=Task:Librarian	0	2	2	0	2	20	19	13	0	0	0	0	0	1	0	0	200	0	259	200	59
19=Task:User	0	5	4	0	9	25	25	23	0	8	0	0	0	0	0	0	44	10	153	10	143
Total																			6658	4784	1874

Table A.46: Confusion Matrix for H-48 (Domain)

	Output:																				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**	**TP**	**FP**
1=Comm:Channel	33	3	0	0	1	7	17	3	0	0	0	0	1	4	1	0	2	0	72	33	39
2=Comm:Feedback	0	299	1	0	4	2	46	10	0	0	0	10	0	1	10	4	0	1	388	299	89
3=Comm:Pausing	0	2	264	0	0	0	10	0	0	0	0	0	2	0	0	2	1	0	281	264	17
4=Info:Feedback	0	1	0	4	9	2	62	0	0	0	0	3	0	0	3	1	1	0	86	4	82
5=Info:Obj	0	2	2	1	954	20	110	60	0	0	6	1	1	0	4	4	10	5	1180	954	226
6=Info:Other	9	10	13	0	36	359	167	49	0	1	8	0	1	4	3	1	14	1	676	359	317
7=Info:Problem	1	11	5	3	75	20	1247	35	0	3	1	1	1	0	9	0	5	4	1421	1247	174
8=Info:Search	0	5	5	0	89	38	158	516	0	2	0	0	1	0	3	3	21	4	845	516	329
9=Social:Apology	0	2	5	0	1	0	4	3	41	0	0	0	0	0	1	0	0	0	57	41	16
10=Social:Closing Ritual	0	1	2	0	0	8	0	2	0	103	1	0	6	0	9	6	0	3	141	103	38
11=Social:Downplay	0	1	0	0	1	2	10	1	0	0	46	0	0	0	0	0	0	0	61	46	15
12=Social:Exclamation	0	14	1	0	3	0	24	1	0	0	0	24	0	1	1	1	0	0	70	24	46
13=Social:Gratitude	0	0	0	0	0	1	12	0	3	1	0	0	420	0	1	0	0	2	440	420	20
14=Social:Greeting	0	1	0	0	0	0	2	0	0	0	1	0	0	247	0	0	0	0	251	247	4
15=Social:Rapport	1	8	2	1	7	6	53	13	0	19	2	2	1	1	60	8	1	3	188	60	128
16=Social:Valediction	0	1	0	0	0	1	2	2	0	0	0	0	1	0	2	104	2	0	115	104	11
17=Task:Librarian	0	0	7	0	1	22	19	20	0	0	1	0	0	0	0	0	208	0	278	208	70
19=Task:User	0	1	4	0	18	20	18	31	0	6	1	0	0	0	1	1	20	29	150	29	121
Total																			6700	4958	1742

Table A.47: Confusion Matrix for H-58 (Domain)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Comm:Channel	0.0635	0.0014	0.3077	0.0635	0.1053
2=Comm:Feedback	0.7036	0.0040	0.9161	0.7036	0.7959
3=Comm:Pausing	0.8371	0.0594	0.3671	0.8371	0.5104
4=Info:Feedback	0.0000	0.0008	0.0000	0.0000	0.0000
5=Info:Obj	0.6405	0.0468	0.7562	0.6405	0.6936
6=Info:Other	0.4045	0.1491	0.2182	0.4045	0.2835
7=Info:Problem	0.2071	0.0610	0.4739	0.2071	0.2883
8=Info:Search	0.0240	0.0040	0.4773	0.0240	0.0457
9=Social:Apology	0.4000	0.0003	0.9231	0.4000	0.5581
10=Social:Closing Ritual	0.6667	0.0055	0.7000	0.6667	0.6829
11=Social:Downplay	0.0286	0.0000	1.0000	0.0286	0.0556
12=Social:Exclamation	0.4444	0.0011	0.8372	0.4444	0.5806
13=Social:Gratitude	0.9815	0.0888	0.4337	0.9815	0.6016
14=Social:Greeting	0.9160	0.0031	0.9197	0.9160	0.9178
15=Social:Rapport	0.0510	0.0035	0.3030	0.0510	0.0873
16=Social:Valediction	0.7143	0.0005	0.9615	0.7143	0.8197
17=Task:Librarian	0.8701	0.1785	0.1615	0.8701	0.2725
19=Task:User	0.0069	0.0008	0.1667	0.0069	0.0132
Weighted Average	0.4434	0.0514	0.5315	0.4434	0.4138

Table A.48: Measurements for S-16 (Domain)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Comm:Channel	0.1846	0.0005	0.8000	0.1846	0.3000
2=Comm:Feedback	0.8141	0.0136	0.7961	0.8141	0.8050
3=Comm:Pausing	0.9398	0.0175	0.6964	0.9398	0.8000
4=Info:Feedback	0.0094	0.0002	0.5000	0.0094	0.0185
5=Info:Obj	0.7587	0.0476	0.7629	0.7587	0.7608
6=Info:Other	0.4709	0.0288	0.6529	0.4709	0.5472
7=Info:Problem	0.8355	0.1675	0.5890	0.8355	0.6910
8=Info:Search	0.5097	0.0462	0.6144	0.5097	0.5572
9=Social:Apology	0.8684	0.0003	0.9429	0.8684	0.9041
10=Social:Closing Ritual	0.6777	0.0049	0.7257	0.6777	0.7009
11=Social:Downplay	0.7442	0.0002	0.9697	0.7442	0.8421
12=Social:Exclamation	0.1846	0.0012	0.6000	0.1846	0.2824
13=Social:Gratitude	0.9709	0.0090	0.8791	0.9709	0.9227
14=Social:Greeting	0.9735	0.0073	0.8271	0.9735	0.8943
15=Social:Rapport	0.2456	0.0022	0.7500	0.2456	0.3700
16=Social:Valediction	0.8696	0.0003	0.9804	0.8696	0.9217
17=Task:Librarian	0.7065	0.0187	0.6409	0.7065	0.6721
19=Task:User	0.0200	0.0008	0.3750	0.0200	0.0380
Weighted Average	0.6909	0.0576	0.6881	0.6909	0.6674

Table A.49: Measurements for H-16 (Domain)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Comm:Channel	0.1429	0.0006	0.7333	0.1429	0.2391
2=Comm:Feedback	0.7775	0.0157	0.7462	0.7775	0.7615
3=Comm:Pausing	0.8821	0.0110	0.7632	0.8821	0.8183
4=Info:Feedback	0.0532	0.0031	0.1923	0.0532	0.0833
5=Info:Obj	0.7611	0.0504	0.7591	0.7611	0.7601
6=Info:Other	0.4763	0.0273	0.6647	0.4763	0.5549
7=Info:Problem	0.8131	0.1524	0.5942	0.8131	0.6866
8=Info:Search	0.4983	0.0503	0.5957	0.4983	0.5426
9=Social:Apology	0.8033	0.0010	0.8750	0.8033	0.8376
10=Social:Closing Ritual	0.6591	0.0063	0.6744	0.6591	0.6667
11=Social:Downplay	0.8298	0.0009	0.8667	0.8298	0.8478
12=Social:Exclamation	0.2235	0.0007	0.7917	0.2235	0.3486
13=Social:Gratitude	0.9766	0.0197	0.7680	0.9766	0.8598
14=Social:Greeting	0.9880	0.0059	0.8636	0.9880	0.9216
15=Social:Rapport	0.2995	0.0056	0.6022	0.2995	0.4000
16=Social:Valediction	0.8485	0.0021	0.8889	0.8485	0.8682
17=Task:Librarian	0.7196	0.0190	0.6113	0.7196	0.6610
19=Task:User	0.0632	0.0018	0.4783	0.0632	0.1117
Weighted Average	0.6815	0.0548	0.6741	0.6815	0.6604

Table A.50: Measurements for H-17 (Domain)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**
1=Comm:Channel	0.2090	0.0003	0.8750	0.2090	0.3373
2=Comm:Feedback	0.7262	0.0112	0.8027	0.7262	0.7625
3=Comm:Pausing	0.9118	0.0109	0.7750	0.9118	0.8378
4=Info:Feedback	0.0446	0.0003	0.7143	0.0446	0.0840
5=Info:Obj	0.8233	0.0532	0.7768	0.8233	0.7994
6=Info:Other	0.4719	0.0217	0.7197	0.4719	0.5700
7=Info:Problem	0.8817	0.1756	0.5742	0.8817	0.6954
8=Info:Search	0.5063	0.0348	0.6779	0.5063	0.5797
9=Social:Apology	0.7115	0.0000	1.0000	0.7115	0.8315
10=Social:Closing Ritual	0.7143	0.0041	0.7724	0.7143	0.7422
11=Social:Downplay	0.8250	0.0003	0.9429	0.8250	0.8800
12=Social:Exclamation	0.1667	0.0007	0.7222	0.1667	0.2708
13=Social:Gratitude	0.9624	0.0060	0.9131	0.9624	0.9371
14=Social:Greeting	0.9806	0.0074	0.8377	0.9806	0.9036
15=Social:Rapport	0.2323	0.0024	0.7419	0.2323	0.3538
16=Social:Valediction	0.8860	0.0007	0.9528	0.8860	0.9182
17=Task:Librarian	0.6914	0.0208	0.5741	0.6914	0.6273
19=Task:User	0.0616	0.0013	0.5000	0.0616	0.1098
Weighted Average	0.7046	0.0564	0.7176	0.7046	0.6826

Table A.51: Measurements for H-18 (Domain)

	TP Rate	FP Rate	Precision	Recall	F-Measure
1=Comm:Channel	0.2500	0.0013	0.6897	0.2500	0.3670
2=Comm:Feedback	0.8250	0.0191	0.7205	0.8250	0.7692
3=Comm:Pausing	0.9029	0.0114	0.7629	0.9029	0.8270
4=Info:Feedback	0.0270	0.0004	0.5000	0.0270	0.0513
5=Info:Obj	0.7624	0.0535	0.7570	0.7624	0.7597
6=Info:Other	0.5143	0.0310	0.6562	0.5143	0.5767
7=Info:Problem	0.8334	0.1631	0.5713	0.8334	0.6779
8=Info:Search	0.4861	0.0407	0.6339	0.4861	0.5503
9=Social:Apology	0.6833	0.0000	1.0000	0.6833	0.8119
10=Social:Closing Ritual	0.6855	0.0056	0.6855	0.6855	0.6855
11=Social:Downplay	0.6346	0.0001	0.9706	0.6346	0.7674
12=Social:Exclamation	0.1444	0.0010	0.6500	0.1444	0.2364
13=Social:Gratitude	0.9677	0.0096	0.8752	0.9677	0.9191
14=Social:Greeting	0.9583	0.0073	0.8214	0.9583	0.8846
15=Social:Rapport	0.2353	0.0020	0.7586	0.2353	0.3592
16=Social:Valediction	0.9127	0.0007	0.9583	0.9127	0.9350
17=Task:Librarian	0.6792	0.0244	0.5654	0.6792	0.6171
19=Task:User	0.0579	0.0014	0.5238	0.0579	0.1043
Weighted Average	0.6836	0.0555	0.6856	0.6836	0.6608

Table A.52: Measurements for H-20 (Domain)

	TP Rate	FP Rate	Precision	Recall	F-Measure
1=Comm:Channel	0.2245	0.0000	1.0000	0.2245	0.3667
2=Comm:Feedback	0.8164	0.0218	0.6851	0.8164	0.7450
3=Comm:Pausing	0.9289	0.0115	0.7605	0.9289	0.8363
4=Info:Feedback	0.0099	0.0003	0.3333	0.0099	0.0192
5=Info:Obj	0.7833	0.0511	0.7496	0.7833	0.7661
6=Info:Other	0.5022	0.0314	0.6419	0.5022	0.5635
7=Info:Problem	0.8140	0.1309	0.6416	0.8140	0.7176
8=Info:Search	0.5031	0.0483	0.5931	0.5031	0.5444
9=Social:Apology	0.7292	0.0002	0.9722	0.7292	0.8333
10=Social:Closing Ritual	0.6855	0.0058	0.6911	0.6855	0.6883
11=Social:Downplay	0.7667	0.0012	0.8519	0.7667	0.8070
12=Social:Exclamation	0.2892	0.0009	0.8000	0.2892	0.4248
13=Social:Gratitude	0.9714	0.0156	0.8322	0.9714	0.8964
14=Social:Greeting	0.9798	0.0062	0.8587	0.9798	0.9153
15=Social:Rapport	0.2917	0.0083	0.5091	0.2917	0.3709
16=Social:Valediction	0.8750	0.0031	0.8198	0.8750	0.8465
17=Task:Librarian	0.6656	0.0183	0.6409	0.6656	0.6530
19=Task:User	0.0359	0.0015	0.3750	0.0359	0.0656
Weighted Average	0.6946	0.0510	0.6797	0.6946	0.6722

Table A.53: Measurements for H-24 (Domain)

	TP Rate	FP Rate	Precision	Recall	F-Measure
1=Comm:Channel	0.3049	0.0008	0.8333	0.3049	0.4464
2=Comm:Feedback	0.7863	0.0157	0.7506	0.7863	0.7680
3=Comm:Pausing	0.9515	0.0106	0.7606	0.9515	0.8454
4=Info:Feedback	0.0250	0.0005	0.5000	0.0250	0.0476
5=Info:Obj	0.7822	0.0401	0.7985	0.7822	0.7903
6=Info:Other	0.5527	0.0325	0.6638	0.5527	0.6031
7=Info:Problem	0.8455	0.1553	0.6142	0.8455	0.7115
8=Info:Search	0.5694	0.0347	0.6885	0.5694	0.6234
9=Social:Apology	0.6905	0.0005	0.9062	0.6905	0.7838
10=Social:Closing Ritual	0.7122	0.0051	0.7500	0.7122	0.7306
11=Social:Downplay	0.7679	0.0015	0.8113	0.7679	0.7890
12=Social:Exclamation	0.3605	0.0020	0.7045	0.3605	0.4769
13=Social:Gratitude	0.9396	0.0042	0.9417	0.9396	0.9406
14=Social:Greeting	0.9843	0.0066	0.8567	0.9843	0.9161
15=Social:Rapport	0.3046	0.0046	0.6386	0.3046	0.4125
16=Social:Valediction	0.9669	0.0020	0.9000	0.9669	0.9323
17=Task:Librarian	0.7722	0.0170	0.6472	0.7722	0.7042
19=Task:User	0.0654	0.0009	0.6250	0.0654	0.1183
Weighted Average	0.7185	0.0523	0.7189	0.7185	0.6996

Table A.54: Measurements for H-48 (Domain)

	TP Rate	FP Rate	Precision	Recall	F-Measure
1=Comm:Channel	0.4583	0.0017	0.7500	0.4583	0.5690
2=Comm:Feedback	0.7706	0.0100	0.8260	0.7706	0.7973
3=Comm:Pausing	0.9395	0.0073	0.8489	0.9395	0.8919
4=Info:Feedback	0.0465	0.0008	0.4444	0.0465	0.0842
5=Info:Obj	0.8085	0.0444	0.7957	0.8085	0.8020
6=Info:Other	0.5311	0.0247	0.7067	0.5311	0.6064
7=Info:Problem	0.8776	0.1353	0.6359	0.8776	0.7374
8=Info:Search	0.6107	0.0393	0.6917	0.6107	0.6486
9=Social:Apology	0.7193	0.0005	0.9318	0.7193	0.8119
10=Social:Closing Ritual	0.7305	0.0049	0.7630	0.7305	0.7464
11=Social:Downplay	0.7541	0.0032	0.6866	0.7541	0.7188
12=Social:Exclamation	0.3429	0.0026	0.5854	0.3429	0.4324
13=Social:Gratitude	0.9545	0.0024	0.9655	0.9545	0.9600
14=Social:Greeting	0.9841	0.0017	0.9574	0.9841	0.9705
15=Social:Rapport	0.3191	0.0074	0.5556	0.3191	0.4054
16=Social:Valediction	0.9043	0.0047	0.7704	0.9043	0.8320
17=Task:Librarian	0.7482	0.0120	0.7298	0.7482	0.7389
19=Task:User	0.1933	0.0035	0.5577	0.1933	0.2871
Weighted Average	0.7400	0.0461	0.7379	0.7400	0.7272

Table A.55: Measurements for H-58 (Domain)

Bibliography

Allan, J. (2000). Nlp for ir. In *Tutorial presented at the NAACL/ANLP Language Technology Joint Conference*, Seattle, WA.

Allen, B. L. (1988). Text structures and the user-intermediary interaction. *Reference Quarterly*, pages 535–541.

Allwood, J. (1992). On dialogue cohesion. *Gothenburg Papers in Theoretical Linguistics*, 65.

Altun, Y., Tsochantaridis, I., Hofmann, T., et al. (2003). Hidden markov support vector machines. In *ICML '03: Proceedings of 20th International Conference on Machine Learning*.

Arnold, J. and Kaske, N. (2005). Evaluating the quality of a chat service. *Libraries and the Academy*, 5(2):177–193.

Auster, E. and Lawton, S. B. (1984). Search interview techniques and information gain as antecedents of user satisfaction with online retrieval. *Journal of the American Society for Information Science*, 35(2):90–103.

Austin, J. L. (1975). *How to do things with words*. Harvard University Press, Cambridge, MA, USA, 2nd edition.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., and Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–328, New York, NY, USA. ACM Press.

Belkin, Seeger, and Wersig (1983). Distributed expert problem treatment as a model for information system analysis and design. *The Journal of Information Science*, 5:153–167.

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In *Information Retrieval*, pages 55–66.

Belkin, N. J., Borgman, C. L., Brooks, H. M., Bylander, T., and Croft, W. B. (1987a). Distributed expert-based information systems: an interdisciplinary approach. *Information Processing and Management*, 23(5):395–409.

Belkin, N. J., Brooks, H. M., and Daniels, P. J. (1987b). Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies*, 27(2):127–144.

Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). Ask for information retrieval. part i: Background and theory. *Journal of Documentation*, 38(2):61–71.

Bopp, R. E. and Smith, L. C. (2001). *Reference and information services: an introduction*. Libraries Unlimited, 3rd edition.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study of part-of-speech tagging. *Computational Linguistics*, 21(4):543–566.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7).

Bristow, A. (1992). Academic reference service over electronic mail. *College and Research Libraries News*, 53(10):631–637.

Brooks, H. M. (1986). *An intelligent interface for document retrieval systems: Developing the problem description and retrieval strategy components*. PhD thesis, Department of Information Science, The City University, London, London, UK.

Brooks, H. M. and Belkin, N. J. (1983). Using discourse analysis for the design of information retrieval interaction mechanisms. In *SIGIR '83: Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 31–47, New York, NY, USA. ACM Press.

Brooks, H. M., Daniels, P. J., and Belkin, N. J. (1986). Research on information interaction and intelligent information provision mechanisms. *Journal of Information Science*, 12(1-2):37–44.

Bunt, H. (2000). Dialogue pragmatics and context specification. In *In Abduction, Belief and Context in Dialogue; studies in computational*, pages 81–150. John Benjamins.

Bunt, H. (2006). Dimensions in dialogue annotation. In *LREC '06: Proceedings of the 5th International Conference on Language Resources and Evaluation*, Paris, France. European Language Resources Association.

Bunt, H. (2007). Multifunctionality and multidimensional dialogue act annotation. In et al., E. A., editor, *Communication - Action - Meaning, A Festschrift to Jens Allwood*, pages 237–259. Gothenburg University Press.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., and Traum, D. (2010). Towards an iso standard for dialogue act annotation. In *Proceedings of LREC 2010, the Seventh International Conference on Language Resources and Evaluation*, pages 16–23, Malta.

Bunt, H. C. (1994). Context and dialogue control. *THINK Quarterly*, 3:19–31.

Cardoso-Cachopo, A. and Oliveira, A. (2003). An empirical comparison of text categorization methods. In *String Processing and Information Retrieval*, pages 183–196. Springer.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., and Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Comput. Linguist.*, 23(1):13–31.

Carter, D. S. and Janes, J. W. (2000). Unobtrusive data analysis of digital reference questions and service at the internet public library: An exploratory study. *Library Trends*, 49(2).

Carvalho, V. R. and Cohen, W. W. (2006). Improving "email speech acts" analysis via n-gram selection. In *ACTS '09: Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, pages 35–41, Morristown, NJ, USA. Association for Computational Linguistics.

Case, D. O. (2002). *Looking for information: a survey of research on information seeking, needs and behavior.* Academic Press, San Diego, CA, USA.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines.*

Clark, A. and Popescu-Belis, A. (2004). Multi-level dialogue act tags. In Strube, M. and Sidner, C., editors, *SIGDIAL '04: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 163–170, Cambridge, MA. Association for Computational Linguistics.

Cleverdon, C. (1967). The cranfield tests on index language devices. *Aslib Proceedings*, 19:173–194.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, R. (1987). Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13:I l–24.

Cohen, W. W., Carvalho, V. R., and Mitchell, T. M. (2004). Learning to classify email into "speech acts". In Lin, D. and Wu, D., editors, *Proceedings of EMNLP '04*, pages 309–316, Barcelona, Spain. Association for Computational Linguistics.

Cong, G., Wang, L., Lin, C.-Y., Song, Y.-I., and Sun, Y. (2008). Finding question-answer pairs from online forums. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 467–474, New York, NY, USA. ACM.

Connaway, L. S. and Radford, M. L. (2007). Service sea change: Clicking with screenagers through virtual reference. In *Presented at the Association of College and Research Libraries 13th National Conference, "Sailing into the Future – Charting Our Destiny,"*, Baltimore, MD, USA.

Connaway, L. S. and Radford, M. L. (2011). Seeking synchronicity: Revelations and recommendations for virtual reference. Technical report, OCLC Research, Dublin, OH.

Core, M. G. and Allen, J. F. (1997). Coding dialogues with the DAMSL annotation scheme. In Traum, D., editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. American Association for Artificial Intelligence.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Croft, W. B. and Thompson, R. H. (1987). Ir3: A new approach to the design of document retrieval system. Technical report, University of Massachusetts, Amherst, MA, USA.

Crouch, W. W. and Lucia, J. (1980). Analysis of verbal behaviors in the presearch interview: codebook with instructions for transcribing interviews. Report of the Presearch Interview

Project ED 205 184, School of Information Studies, Syracuse University, Syracuse, NY, USA.

Crouch, W. W. and Lucia, J. (1981). Analysis of verbal behaviors in the presearch interview. Report of the Presearch Interview Project ED 205 184, School of Information Studies, Syracuse University, Syracuse, NY, USA.

Daniels, P. J., Brooks, H. M., and Belkin, N. J. (1985). Using problem structures for driving human-computer dialogues. In *In RIAO 85, Acts of the Conference: Recherche d'Informations Assistee par Ordinateur*, pages 645–660, Grenoble. I.M.A.G.

Dervin, B. (1983). An overview of sense-making research: Concepts, methods, and results to date. In *Presented at the Annual Meeting of the International Communication Association*, Dallas, TX.

Dervin, B. and Dewdney, P. (1986). Neutral questioning: A new approach to the reference interview. *Research Quarterly*, 25(4):506–513.

Dewdney, P. and Ross, C. (1994). Flying a light aircraft: Reference service evaluation from the user's viewpoint. *Reference Quarterly*, 34(2):217–230.

Diamond, W. and Pease, B. (2001). Digital reference: a case study of question types in an academic library. *Reference Services Review*, 29(3):210–218.

Ding, S., Cong, G., Lin, C., and Zhu, X. (2009). Using Conditional Random Fields to Extract Contexts and Answers of Questions from Online Forums. In *ACL-HLT '08: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics: Human Language Technologies*, pages 710–718, Columbus, Ohio.

el Hadi, W. M. (2004). Human language technology and its role in information access and management. *Cataloging and Classification Quarterly*, 37(1):131–151.

Ellis, L. (1994). *Research methods in the social sciences*. Brown & Benchmark, Madison, WI.

Feng, D., Shaw, E., Kim, J., and Hovy, E. (2006a). An intelligent discussion-bot for answering student queries in threaded discussions. In *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*, pages 171–177, New York, NY, USA. ACM.

Feng, D., Shaw, E., Kim, J., and Hovy, E. (2006b). Learning to detect conversation focus of threaded discussions. In *HLT '06: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 208–215, Morristown, NJ, USA. Association for Computational Linguistics.

Fernandez, R. and Picard, R. (2002). Dialog act classification from prosodic features using support vector machines. In *Proceedings of speech prosody*, pages 291–294.

Foley, M. (2002). Instant messaging reference in an academic library: a case study. *College & Research Libraries*, 63(1):36.

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (2004). Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 669, Morristown, NJ, USA. Association for Computational Linguistics.

Goffman, E. (1981). *Forms of talk*. University of Pennsylvania publications in conduct and communication. University of Pennsylvania Press, Philadelphia, PA, USA.

Grimes, J. E. (1975). *The thread of discourse*. Mouton, Hague, Netherlands.

Groenendijk, J. and Stokhof, M. (1989). Dynamic montague grammar. In *Papers from the Second Symposium on Logic and Language*, pages 3–48. Akademiai Kiadoo.

Grosz, B. J. (1978). Focusing in dialog. Technical Report 166, Artificial Intelligence Center, SRI International, Menlo Park, CA, USA.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman, London, UK.

Harris, M. (1976). History and significance of the emic/etic distinction. *Annual Review of Anthropology*, 5:329–350.

Harris, Z. S. (1952). Discourse analysis. *Language*, 28(1):1–30.

Hirko, B. and Ross, M. (2004). *Virtual reference training: The complete guide to providing anytime, anywhere answers*. The American Library Association.

Hong, L. and Davison, B. D. (2009). A classification-based approach to question answering in discussion boards. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 171–178, New York, NY, USA. ACM.

Howard, E. H. and Jankowski, T. A. (1986). Reference services via electronic mail. *Bulletin of the Medical Library Association*, 74(1):41–44.

Hu, J., Passonneau, R. J., and Rambow, O. (2009). Contrasting the interaction structure of an email and a telephone corpus: a machine learning approach to annotation of dialogue function units. In *SIGDIAL '09: Proceedings of the SIGDIAL 2009 Conference*, pages 357–366, Morristown, NJ, USA. Association for Computational Linguistics.

IBM (2011). Brining smarter computing to big data.

Ingwersen, P. (1982). Search procedures in the library—analysed from the cognitive point of view. *Journal of Documentation*, 38(3):165–191.

Ingwersen, P. (1992). *Information retrieval interaction*, volume 246. Taylor Graham Publishing, London, UK.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of Documentation*, 52(1):3 – 50.

Inoue, K. (2009). Discourse analysis of online chat reference interviews for modeling online information- seeking dialogues. In *In Proc. of ASIST '09, Poster Presentation*, Vancouver, BC, Canada.

Ivanciuc, O. (2007). Applications of support vector machines in chemistry. *Reviews in Computational Chemestry*, 23:291–400.

Janes, J. (2002). Digital Reference: Reference Librarians' Experiences and Attitudes. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 53(7):549–566.

Janes, J., Carter, D., and Memmott, P. (1999). Digital reference services in academic libraries. *Reference and User Services Quarterly*, 39(2):145–150.

Janes, J. and McClure, C. (1999). The Web as a Reference Tool: Comparisons with Traditional Sources. *Public Libraries*.

Jefferson, G. (1972). Side sequences. In Sudnow, D., editor, *Studies in Social Interaction*, pages 294–338. The Free Press, New York, NY, USA.

Jeon, J., Croft, W. B., and Lee, J. H. (2005). Finding semantically similar questions based on their answers. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 617–618, New York, NY, USA. ACM.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142, Berlin. Springer.

Joachims, T. (1999). Making large-scale svm learning practical. In Schölkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA.

Joachims, T. (2008). Svmhmm: Sequence tagging with structural support vector machines. Retrieved from http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html.

Johnstone, B. (2002). *Discourse Analysis*. Blackwell Publishers Inc., Malden, Massachusetts.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and Ess-Dykema, C. V. (1997). Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding*, pages 88–95.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2nd edition.

Kando, N. (1995). Structure of news stories : As relating to the indexing and retrieval. *Journal of Japan Indexers Association*, 19(1):1–17.

Kando, N. (1997). Text-level structure of research papers: Implications for text-based information processing systems. In *Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research*, Aberdeen, Scotland. Springer.

Kaske, N. and Arnold, J. (2002). An unobtrusive evaluation of online real time library reference services. In *American Library Association, Annual Conference*, volume 20, page 2005, Atlanta, GA.

Kasowitz, A., Bennett, B., and Lankes, R. D. (2000). Quality standards for digital reference consortia. *Reference & User Services Quarterly*, 39(4):355.

Katz, W. A. (2002). *Introduction to Reference Work*, volume 1. McGraw-Hill, New York, NY, 8th edition.

Kelly, G. A. (1963). *A theory of personality: The psychology of personal constructs.* W. W. Norton and Company, New York, NY.

Kim, J., Chern, G., Feng, D., Shaw, E., and Hovy, E. (2006). Mining and assessing discussions on the web through speech act analysis. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference*.

Kita, K., Fukui, Y., Nagata, M., and Morimoto, T. (1996). Automatic acquisition of probabilistic dialogue models. In *ICSLP '96: Proceedings of the Fourth International Conference on Spoken Language Processing*, volume 1, pages 196–199. New York: IEEE.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology.* Sage Publications, Thousand Oaks, CA.

Kuhlthau, C., Spink, A., and Cool, C. (1992). Exploration into stages in the information search process in online information retrieval: communication between users and intermediaries. In *ASIS '92: Proceedings of the 55th Annual Meeting of the American Society for Information Science*, pages 67–71, Silver Springs, MD, USA. American Society for Information Science.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *JASIS*, 42(5):361–371.

Kurasaki, K. (2000). Intercoder reliability for validating conclusions drawn from open-ended interview data. *Field methods*, 12(3):179.

Kwasnik, B. H. (1992). A descriptive study of the functional components of browsing. In *Proceedings of the IFIP TC2/WG2.7 Working Conference on Engineering for Human-Computer Interaction*, pages 191–203, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co.

Labov, W. and Waletzky, J. (1967). Narrative analysis. In Helm, J., editor, *Essays on the Verbal and Visual Arts*, pages 12–44. University of Washington Press, Seattle, WA.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 282–289. Citeseer.

Lagace, N. (1999). Establishing online reference services. *The Internet Public Library handbook*, pages 153–183.

Lan, K. C., Ho, K. S., Luk, R. W. P., , and Leong, H. V. (2008). Dialogue act recognition using maximum entropy. *Journal of the American Society for Information Science and Technology*, 59(6):859–874.

Lankes, R. D. (2009). The digital research agenda. In Lankes, R. D., Nicholson, S., and Goodrum, A., editors, *New Concepts in Digital Reference (Synthesis Lectures on Information Concepts, Retrieval & Services)*, chapter 1, pages 1–22. Morgan and Claypool Publishers.

Lankes, R. D. and Kasowitz, A. S. (1998). *The AskA Starter Kit: How To Build and Maintain Digital Reference Services*. Information Resources Publications, Syracuse University.

Lankes, R. D., Silverstein, J., and Nicholson, S. (2006). *Participatory Networks: The Library as Conversation*. Information Institute of Syracuse.

Lau, T. and Horvitz, E. (1999). Patterns of search: Analyzing and modeling web query refinement. In *UM '99: Proceedings of the Seventh International Conference on User Modelling*, pages 119–128, Banff, Canada. Springer Wien.

Lease, M. (2007). Natural language processing for information retrieval: the time is ripe (again). In *PIKM '07: Proceedings of the ACM first Ph.D. workshop in CIKM*, pages 1–8, New York, NY, USA. ACM.

Levinson, S. C. (1983). *Pragmatics.* Cambridge University Press, Cambridge, UK.

Lewis, D. D. and Jones, K. S. (1996). Natural language processing for information retrieval. *Communication ACM*, 39(1):92–101.

Liddy, E. D. (1991). The discourse-level structure of empirical abstracts: an exploratory study. *Information Processing and Management*, 27(1):55–81.

Liddy, E. D. (1998). Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science and Technology*, 24(4).

Lin, J. (2007). Is question answering better than information retrieval? towards a task-based evaluation framework for question series. In *HLT '07: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 212–219, Rochester, New York. Association for Computational Linguistics.

Lipow, A. G. (2002). *The Virtual Reference Librarian's Handlbook.* Neal-Schuman Publishers, Inc.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communication ACM*, 49(4):41–46.

Marom, Y. and Zukerman, I. (2009). An empirical study of corpus-based response automation methods for an e-mail-based help-desk domain. *Comput. Linguist.*, 35:597–635.

McCallum, A., Freitag, D., and Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591–598. Citeseer.

McClure, C., Lankes, R., Gross, M., and Choltco-Devlin, B. (2002). *Statistics, measures and quality standards for assessing digital reference library services: guidelines and procedures.* Information Institute of Syracuse, Syracuse University, Syracuse, NY.

Meadow, C. T. (2007). *Text Information Retrieval Systems*. Academic Press, Inc., Orlando, FL, USA.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, A 209:415–446.

Meteer, M. and Taylor, A. (1995). Dysfluency annotation stylebook for the switch- board corpus. Technical report, Linguistic Data Consortium.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The psychological review*, 63:81–97.

Mitchell, T. M. (2006). The discipline of machine learning. Unpublished CMU-ML-06-108, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Sage Publications.

Nicholson, S. and Lankes, R. D. (2007). The digital reference electronic warehouse project: creating the infrastructure for digital reference research through a multidisciplinary knowledge base. *Reference and User Services Quarterly*, 46(2):45–59.

Nilsen, K. (2004). The library visit study: User experiences at the virtual reference desk. information research. *Information Research*, 9(2).

Oddy, R. N. (1977). Information retrieval through man-machine dialogue. *Journal of Documentation*, 33(1):1–14.

Oddy, R. N., Liddy, E. D., Balakrishnan, B., Bishop, A., Elewononi, J., and Martin, E. (1992). Towards the use of situational information in information retrieval. *Journal of Documentation*, 48(2):123–171.

Pask, G. (1975). *Conversation, cognition and learning*. Elsevier, New York.

Pask, G. (1976). *Conversation Theory Applications in Education and Epistemology*. Elsevier.

Pevzner, L. and Hearst, M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Pike, K. L. (1967). *Language in relation to a unified theory of the structure of human behavior*. Janua Linguarum. Mouton Publishers, The Hague, The Netherlands, 2nd edition.

Pomerantz, J. (2003). *Question taxonomies for digital reference*. PhD thesis, School of Information Studies, Syracuse University, Syracuse, NY.

Propp, V. (1958). *Morphology of the folk-tale*. Indiana University Press, Bloomington.

Punyakanok, V. and Roth, D. (2001). The use of classifiers in sequential inference. *Advances in Neural Information Processing Systems*, 13:995–1001.

Radford, M. (1993). *Relational aspects of reference interactions: A qualitative investigation of the perceptions of users and librarians in the academic library*. PhD thesis, Rutgers The State University of New Jersey.

Radford, M. (2006a). Interpersonal communication in chat reference: Encounters with rude and impatient users. In Lankes, R. D., editor, *The virtual reference desk: Creating a reference future*, pages 41–73. Facet Publishing.

Radford, M., of College, A., and Libraries, R. (1999). *The reference encounter: Interpersonal communication in the academic library*. Association of College and Research Libraries Chicago, IL, USA.

Radford, M. L. (2006b). Encountering virtual users: A qualitative investigation of interpersonal communication in chat reference. *Journal of the American Society for Information Science and Technology*, 57(8):1046–1059.

Radford, M. L. and Connaway, L. S. (2005). Seeking synchronicity: Evaluating virtual reference services from user, non-user, and librarian perspectives. Proposal for a research project, submitted February 1, 2005, to the National Leadership Grants for Libraries program of the Institute of Museum and Library Services (IMLS).

Radford, M. L. and Connaway, L. S. (2007a). Are we getting warmer? query clarification in virtual reference. In *Library Research Round Table ALA Annual Conference (Presentation).*

Radford, M. L. and Connaway, L. S. (2007b). "Screenagers" and live chat reference: Living up to the promise. *Scan*, 26(1):31–39.

Reithinger, N., Engel, R., and Klesen, M. (1996). Predicting dialogue acts for a speech-to-speech translation system. In *ICSLP '96: Proceedings of the International Conference on Spoken Language Processing*, pages 654–657.

Reithinger, N. and Klesen, M. (1997). Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*, pages 2235–2238.

Robertson, S. (1977). The probability ranking principle in ir. *Journal of Documentation*, 33(4):294–304.

Robertson, S. E. (2008). On the history of evaluation ir. *Journal of Information Science*, 34(4):439–456.

Ross, C. and Nilsen, K. (2000). Has the internet changed anything in reference? The library visit study, phase 2. *Reference and user services quarterly*, 40(2):147–155.

Ross, C. S. (1999). Finding without seeking: the information encounter in the context of reading for pleasure. *Information Processing and Management*, 35(6):783 – 799.

Ruesch, J. and Bateson, G. (1951). *Communication: The social matrix of psychology.* Norton, New York, NY.

Rumelhart, D. (1977). Understanding and summarizing brief stories. In LaBerge, D. and Samuels, S., editors, *Basic processes in reading: Perception and comprehension*. Lawrence Earlbaum Associates, Hillsdale, NJ.

Rumelhart, D. (1980). Schemata: the building blocks of cognition. In R. Spiro, B. Bruce, . W. B., editor, *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education*, pages 33–58. Lawrence Earlbaum Associates, Hillsdale, NJ.

Ruppel, M. and Fagan, J. (2002). Instant messaging reference: users' evaluation of library chat. *Reference Services Review*, 30(3):183–197.

Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice hall, Upper Saddle River, NJ, 2nd edition.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Samuel, K., Carberry, S., and Vijay-shanker, K. (1998a). Dialogue act tagging with transformation-based learning. In *ACL-COLING '98: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 1150–1156.

Samuel, K., Carberry, S., and Vijay-Shanker, K. (1998b). An investigation of transformation-based learning in discourse. In *ICML '98: Proceedings of the 15th International Conference on Machine Learning*, pages 497–505, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Saracevic, T. (1996). Modeling interaction in information retrieval (IR): A review and proposal. In *ASIST '96: Proceedings of the Annual Meeting of the American Society for Information Science*, volume 33, pages 3–9, Baltimore, MD.

Saracevic, T., Spink, A., and Wu, M. (1997). Users and Intermediaries in Information Retrieval: What Are They Talking About? In *User modeling: proceedings of the sixth international conference, UM97, Chia Laguna, Sardinia, Italy, June 2-5 1997*, page 43. Springer.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language.* Cambridge University Press, Cambridge, UK.

Shawe-Taylor, J. and Cristianini, N. (2000). *Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, UK.

Shrestha, L. and McKeown, K. (2004). Detection of question-answer pairs in email conversations. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 889, Morristown, NJ, USA. Association for Computational Linguistics.

Silverstein, C. and Henzinger, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forumm*, 33(1):6–12.

Smeaton, A. F. (1999). Using nlp or nlp resources for information retrieval tasks. In Strzalkowski, T., editor, *Natural language information retrieval.* Kluwer Academic Publishers, Dordrecht, NL.

Sparck Jones, K. (1999). What is the role of nlp in text retrieval? In Strzalkowski, T., editor, *Natural Language Information Retrieval*, pages 1–24. Kluwer, Dordrecht.

Spink, A., Goodrum, A., and Robins, D. (1995). Search intermediary elicitations during mediated online searching. In *ASIST '95: Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, volume 32, pages 97–102.

Spink, A., Jansen, B., Wolfram, D., and Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109.

Spink, A. and Saracevic, T. (1998). Human-computer interaction in information retrieval: nature and manifestaions of feedback. *Interacting with Computers*, 10(3):249–267.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373.

Stolcke, A., Shriberg, E., Bates, R., Coccaro, N., Jurafsky, D., Martin, R., Meteer, M., Ries, K., Taylor, P., and Ess-Dykema, C. V. (1998). Dialog act modeling for conversational speech. In *Proceedings of the AAAI-98 Spring Symposium on Applying Machine Learning to Discourse Processing*, number SS-98-01, pages 98–105, Menlo Park, CA, USA. AAAI Press.

Surendran, D. and Levow, G. (2006). Dialog act tagging with support vector machines and hidden markov models. In *Ninth International Conference on Spoken Language Processing*.

Tannen, D. (2007). *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*. Cambridge University Press, 2nd edition.

Tatham, M. (2009). Google received 72 percent of u.s. searches in january 2009. Press Release.

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29:178–194.

Teevan, J., Dumais, S., and Liebling, D. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*.

Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text*. PhD thesis, University of Edinburgh.

Toffler, A. (1984). *The Third Wave*. Bantam Books, New York, NY, USA.

Ussery, B. (2008). Google – average number of words per query have increased! *Beu Blog. Last obtained on December. 13, 2010 from http://www.beussery.com/blog/index.php/2008/02/google-average-number-of-words-per-query-have-increased/.*

Uyar, A. (2009). Google stemming mechanisms. *Journal of Information Science*, 35(5):499–514.

van Dijk, T. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognirion*. Lawrence Earlbaum Associates, Hillsdale, NJ.

Verbree, D., Rienks, R., and Heylen, D. (2006). Dialogue-act tagging using smart feature selection; results on multiple corpora. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 70–73. IEEE.

Walker, M. and Passonneau, R. (2001). Date: a dialogue act tagging scheme for evaluation of spoken dialogue systems. In *HLT '01: Proceedings of the first international conference on Human language technology research*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Watzlawick, P., Beavin, J. H., and Jackson, D. D. (1967). *Pragmatics of human communication: A study of interactional patterns, pathologies, and paradoxes*. Norton, New York, NY.

Weise, F. and Borgendale, M. (1986). EARS: Electronic Access to Reference Service. *Bulletin of the Medical Library Association*, 74(4):300.

White, M. D. (1998). Questions in reference interviews. *Journal of Documentation*, 54(4):443 – 465.

White, M. D., Abels, E. G., and Kaske, N. (2003). Evaluation of chat reference service quality pilot study. *D-Lib Magazine*, 9(2).

White, R. and Roth, R. (2009). Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98.

Wilson, T. D. (1999). Models in information behaviour research. *The Journal of Documentation*, 55(3):249–270.

Wimmer, R. and Dominick, J. (1997). *Mass media research: An introduction*. Wadsworth Pub Co.

Wu, M.-M. (1993). *Information interaction dialogue: A study of patron elicitation in the information retrieval interaction*. PhD thesis, Rutgers The State University of New Jersey.

Wu, M.-M. (2005). Understanding patrons' micro-level information seeking (mlis) in information retrieval situations. *Information Processing and Management*, 41(4):929 – 947.

Wu, M.-M. and Liu, Y.-H. (2003). Intermediary's information seeking, inquiring minds, and elicitation styles. *Journal of the American Society for Information Science and Technology*, 54(12):1117–1133.

Zumalt, J. R. and Pasicznyuk, R. W. (1999). The internet and reference services: A real-world test of internet utility. *Reference & User Services Quarterly*, 38(2):165–172.

Keisuke Inoue

EDUCATION

PhD, Information Science and Technology, School of Information Studies, Syracuse University, NY May 2013

Master of Arts, Linguistics, College of Arts and Science, Syracuse University, NY August 2005

Master of Science, Computer Science, LC Smith College, Syracuse University, NY August 2002

Bachelor of Arts, Law, Waseda University, Tokyo, Japan March 1996

PUBLICATIONS AND PRESENTATIONS

Inoue, K., and McCracken, N. (2010). Automated Keyword Extraction of Learning Materials Using Semantic Relations, Poster presentation at iConference, Urbana-Champaign, NC

Inoue, K. (2009). Discourse Analysis of Online Chat Reference Interviews for Modeling Online Information-Seeking Dialogues, Poster presentation at the Annual Meeting of the American Society for Information Science and Technology, Vancouver, BC, Canada

Inoue, K. (2009). Automated Detection of Subject Area for Question Triage in Digital Reference, Poster presentation at iConference, Chapel Hill, NC

Inoue, K. (2008). Dialogue Act Classification of Online Chat Reference Conversations for Information Retrieval, Presentation at Doctoral Forum of Information Interaction in Context, London, UK

Inoue, K. (2008). A Conversation Repository for Participatory Librarianship, Poster presentation at iConference, Los Angeles, CA

Inoue, K., Snyder, J (2007). Conversation Repository for Participatory Librarianship, Workshop on Discourse Oriented Approach on LIS, Nordic Research School in Library and Information Science, Lund, Sweden

Inoue, K. (2007). The Knowledge Base for a Participatory Library: Virtual Reference as Conversation, Connections, Philadelphia, PA

Howison, J., Inoue, K., and Crowston, K. (2006). Social dynamics of free and open source team communications. In Proceedings of the IFIP 2nd International Conference on Open Source Software, Lake Como, Italy

AWARDS AND HONORS

Beta Phi Mu Eugene Garfield Doctoral Dissertation Fellowship June 2011

OCLC/ALISE Library & Information Science Research Grant January 2011

IMLS Doctoral Student Fellowship Summer 2010

Syracuse University Future Professorial Program Award March 2009