An Investigation of Digital Reference Interviews: Dialogue Act Annotation with the Hidden Markov Support Vector Machine

Bei Yu
School of Information Studies, Syracuse University
byu@syr.edu

Keisuke Inoue
School of Information Studies, Syracuse University
kinoue@syr.edu

Final Report
2011 OCLC/ALISE Library and Information Science Research Grant Project

05 August 2012

Suggested citation:
Yu, Bei and Keisuke Inoue. 2012. "An Investigation of Digital Reference Interviews: Dialogue Act Annotation with the Hidden Markov Support Vector Machine." 2011 OCLC/ALISE research grant report published electronically by OCLC Research. Available online at: http://www.oclc.org/research/grants/reports/2011/yu2011.pdf

1 Introduction

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! QA) characterizes a recent trend in web technologies often referred to as the social web. These trends showcase the importance of supporting linguistic interactions in the information-seeking processes of daily life – something that web search engines lack due to the complexity of this human behavior. In this research project, we investigated linguistic properties of information-seeking CMCs and examined the possibility of automatic identification of properties using the machine learning text classification techniques.

2 Research Questions

The two major goals of the study were: 1) to investigate the discourse of question negotiation in digital reference based on theories drawn from linguistic and information-seeking behavior studies, and 2) to experiment with the automatic detection of the discourse by applying the theories to machine learning technologies. These two goals were addressed in the following four research questions.

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 the automatic recognition of the discourse-level semantics of question negotiation in digital reference?

3 Method of Investigation

The study consisted of two stages: 1) discourse analysis, wherein human annotators analyzed linguistic properties of digital reference transcripts and 2) machine-learning experiments, wherein different algorithms and attributes were examined for automatic annotation of the dialogue acts. The discourse analysis identified communicative functions and domains of information exchanges, as well as socio-emotional functions that appear in the information-seeking communication, based on a discourse analysis framework called dialogue acts. The machine-learning experiments identified appropriate algorithms and attributes for learning dialogue acts of digital reference interviews.

3.1 Data

The data, provided by the Online Computer Library Center (OCLC), was a log of virtual 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 make sure that the information problems presented in the interviews required question negotiations. Each interview used for the analysis consisted, on average, of 26 messages sent between a librarian and user.

3.2 Discourse Analysis

The discourse analysis in this study utilized a dialogue act annotation scheme, which had been developed through an exploratory study conducted by the researchers. 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. The idea of having these four aspects was theoretically motivated by the Dynamic Interpretation Theory (Bunt, 1994), which hypothesizes that dialogues are 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 are associated with the context of the

[1] The data became available to the researcher by courtesy of Dr. Radford, Dr. Connaway, and the OCLC.

dialogue. The annotation scheme that was used for this study is summarized in the Table A.1 in Appendices. The annotation was done in two dimensions: 1) *Function*, which represents the type of effect of a linguistic utterance to the cognitive space of a receiver and 2) *Domain*, which represents the aspect of the cognitive space that the function operates upon. The labels are hierarchically organized so that the level of analysis can be adjusted based on the inter-coder reliability of the annotation.

3.3 Machine Learning Experiment

While various algorithms have been proposed for automatically detecting dialogue acts, most of these algorithms are categorized into two approaches: sequential labeling and text classification. The sequential labeling approach has been used widely in speech and dialogue research, where many early automatic dialogue act annotation experiments were conducted. Recently text classification algorithms have been applied to dialogue act annotations. The Hidden Markov Model (HMM) is the most successful algorithm in sequential labeling (Kita et al., 1996; Reithinger et al., 1996; Stolcke et al. 2000), while the Support Vector Machine (SVM) is the most successful algorithm in text classification (Cohen et al., 2004; Carvalho and Cohen, 2006; Hu et al., 2009). For this study a combination of HMM and SVM, called HM-SVM (Altun, 2003), was examined. HM-SVM replaces the standard HMM training with a discriminative learning procedure based on the maximum/soft margin criterion of SVM, thus it provides the benefits of both algorithms: the ability to learn the sequential labels (HMM) and the ability to learn from various types (and numbers) of features using different kernels (SVM). The experiment also examined the effects of different linguistic attributes (features) for machine learning. Using the standard SVM as the baseline, the experiment was a two-way semi-factorial design, examining the isolated effects and the interactive effects of the two algorithms and various features.

4 Results

4.1 Dialogue Act Analysis

4.1.1 Overview

Three MLIS students collectively annotated 210 online reference interview sessions which consisted of 5,489 messages between librarian and user. The annotators went through a two-week training period where they learned the purpose of the research, the nature and structure of the data, the annotation scheme, and the annotation process. The annotation took place from April to August 2011 (approximately 16 weeks), followed by a final clean-up period from October to November 2012.

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. Each dialogue act consists of two labels: a function and a domain. The function of a dialogue act represents what utterances are intended to do, e.g. provide information or conform to social obligations. Domains represent more detailed descriptions of the dialogue acts by specifying which cognitive state the function operates upon, e.g. what kind of information is exchanged, which aspect of the social relationship is dealt with by the utterance, etc. See table A.1 in the appendices for an overview of the classification scheme.

Each annotator was given roughly the same amount of data per week based on the number of messages (approximately 100 - 150 messages, or five to ten conversations). One annotator (Annotator 2) had to leave the project at the end of May due to a job opportunity and their her annotation volume is correspondingly lower– Annotator 1 annotated 103 conversations, Annotator 2 annotated 62 conversations, and Annotator 3 annotated 97 conversations. On average, each conversation had approximately 26 messages, and each message was broken up 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 did not vary widely, ranging from 1.03 to 1.06.

4.1.1 Evaluation

The evaluation of the annotation was done utilizing standard pair-wise evaluation measures applied to the data that two or more annotators annotated and by analyzing disagreements. The proportions of shared data were maintained at about 20% between two annotators and 10% for all three, following Wimmer and Dominick (2006). The evaluation of text segmentation was done using *WindowDiff*, which was proposed by Pevzner and Hearst (2002).

For each pair of annotators, the average *WindowDiff* was calculated over all the messages that they annotated. The very low values, shown below, indicate excellent agreements between the annotators (the value of WindowDiff varies from 0.0 to 1.0, 0.0 meaning the perfect agreement):

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

Given the hierarchical organization of the annotation scheme (shown in Table A.1), the evaluation of annotation was performed at each hierarchical level of the annotation scheme structure. For each message, agreements (or disagreements) were identified by comparing the dialogue acts that were annotated in the order of occurrence. The dialogue acts that were in the same order of occurrence of another dialogue act were called "matching dialogue acts". Some dialogue acts had matching dialogue acts while others did not. Thus, inter-coder agreements were measured at two levels: 1) all the dialogue acts and 2) all the dialogue acts that had matching dialogue acts. Table 4.1 summarizes the inter-coder agreements.

Level of Analysis	Classes	Agreement*	*Kappa*	*Pi*
Function	7	.87 (.92)	(.88)	(.88)
Partial Domain	19	.80 (.84)	(.82)	(.82)
Partial Domain and Function	25	.79 (.83)	(.81)	(.81)
Full Domain	45	.67 (.70)	(.69)	(.69)
Full Domain and Function	75	.66 (.70)	(.69)	(.69)

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

Table 4.1 Inter-coder Agreements

What constitutes the acceptable level of *Kappa* is open to debate; researchers have proposed a wide range of acceptable levels, from .4 to .9 (Neuendorf, 2002). Many researchers, however, use somewhere between .75 and .8 as a "rule of thumb" (Ellis, 1994). Krippendorff (1980) considers .8 < *K* an indication of good reliability, while .67 < *K* < .8 allows only tentative conclusions. Additionally, Carletta (1996) points out that when the unit of analysis (segmentation) is not provided pre-theoretically, the task of coding may become inherently more difficult in content analysis studies. In this research, the annotation of dialogue act functions and domains at the shallower level were considered as reliable and were used as the primary data for both the qualitative analysis (presented in the next subsection) and machine learning experiments (presented in the next section). The annotation of the domains at the detail level and the combination of the partial domains and functions were used only as the secondary data for the qualitative analysis.

4.1.2 Analysis

The annotated data was analyzed from three angles: frequency distributions (statistics), transitions (conditional probability), and word distributions (linguistics). The outcome of the analysis revealed interesting characteristics of digital reference interviews and provided insights towards selecting features for machine learning. Detailed analysis will be presented in the second author's doctoral dissertation (forthcoming); this section provides an overview.

4.1.2.1 Distribution of Dialogue Acts

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

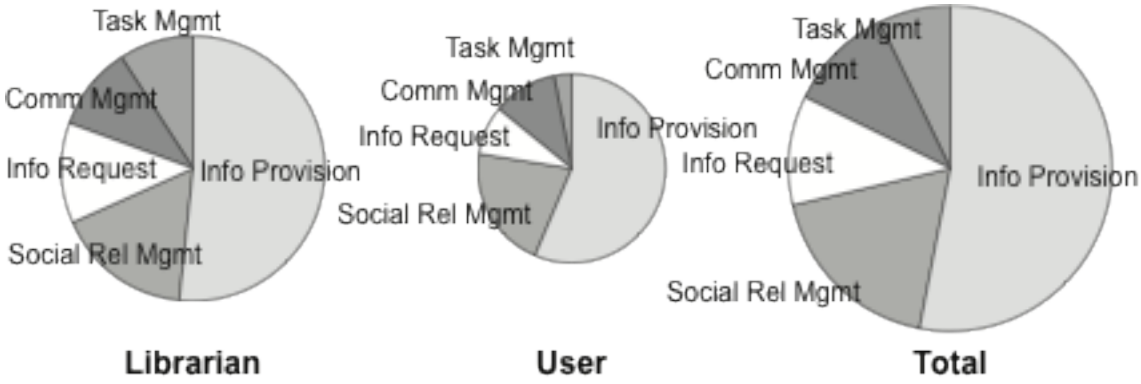

Figure 4.1 Distributions of Dialogue Act Functions

First, librarians contribute to the conversations roughly twice as much as users in overall volume, as well as in every category of functions excepting task management functions where they exceeded users by six times the amount. (Figure 4.1 illustrates the distribution of the Dialogue Act functions by speaker. See Table A.2 for the actual figures.) Thus, the distributions of dialogue act functions by librarians and users are very similar, except for the *Task Management* functions. The asymmetry in the distribution of *Task Management* functions can be explained by the nature of reference sessions; librarians are often the ones that carry out search tasks and explain or suggest tasks to users to users. The symmetry in the distributions of the other functions supports the theories that explain the duality of the communication (Watzlawick et al., 1967; Bunt, 1994).

And secondly, *Information 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, 2006) where the primary goal of the communication is to provide information to satisfy the user's need. The dialogue acts for social relationship were also used frequently (the second most frequently 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 functions. This is consistent with the observations from previous studies (Ruppel and Fagan, 2002; Nilsen, 2004; Radford, 2006), which emphasized the importance of such functions in a successful online reference interview.

The relative position of a dialogue act was determined by the proportion of the sequence number of the containing message to the total number of messages in the conversation. Five positions were defined by dividing message by proportion: *Beginning* (from 0 to .2), *Beginning-mid* (from .2 to .4), *Middle* (from .4 to .6), *Mid-ending* (from .6 to .8), and *Ending* (from .8 to 1.0). The total number of positions was chosen after experimenting with three, five, and ten stages, and five was settled on as a good compromise between precision (i.e. how repeatable the measurement is) and accuracy (i.e. level of detail).

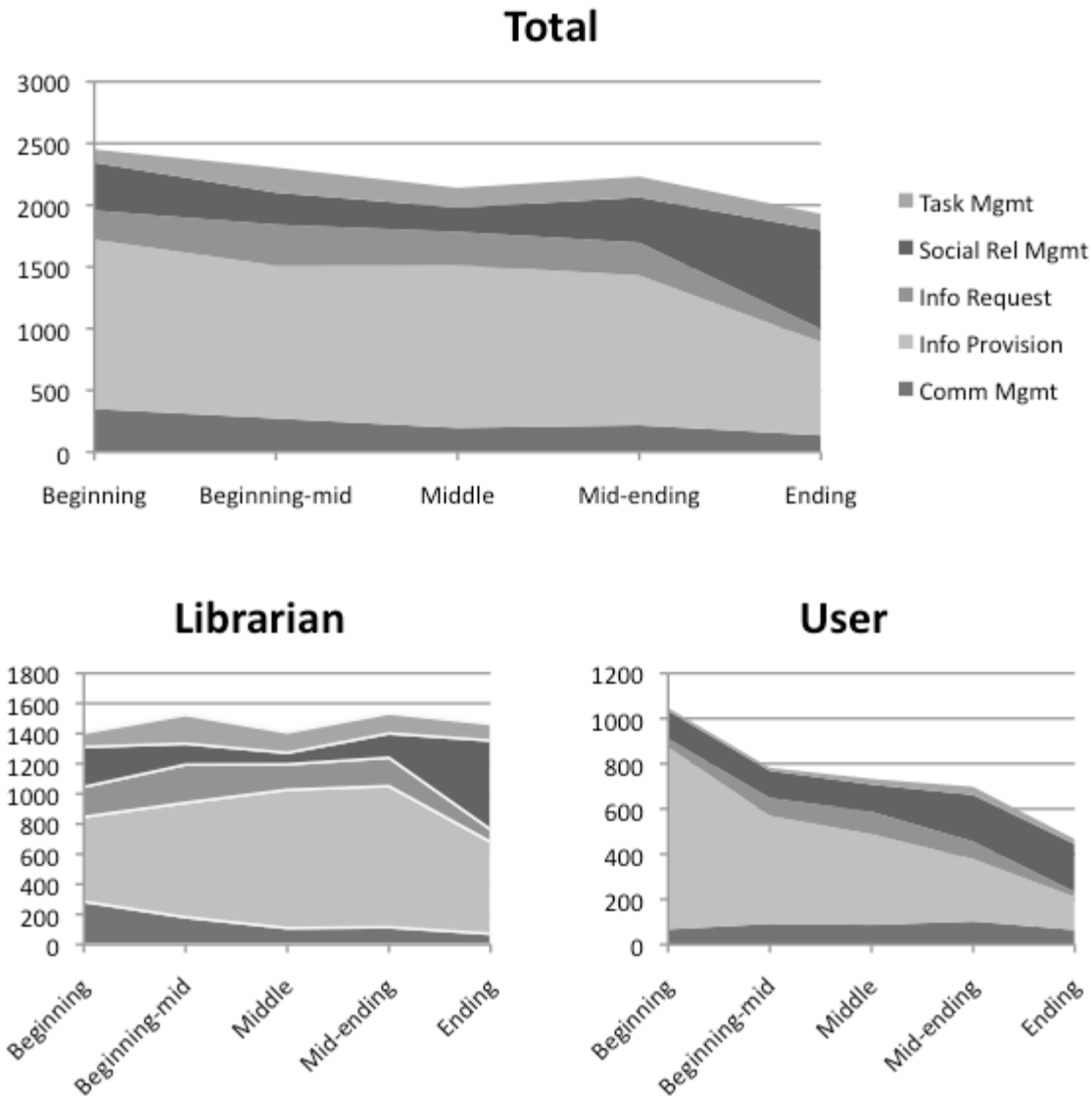

Figure 4.2 Distribution of Functions by Position

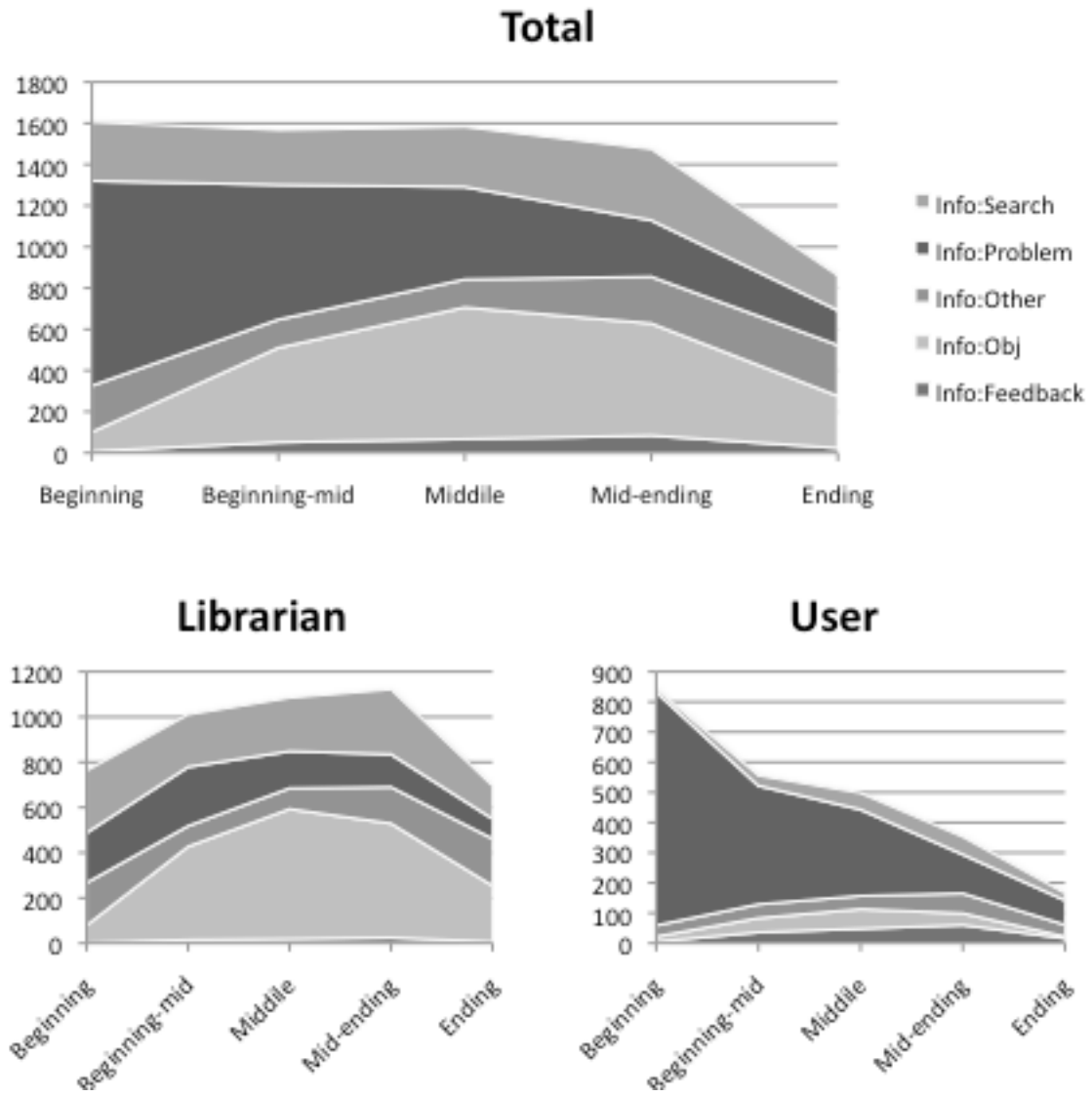

Figure 4.3 Distribution of Information Domains by Position

The distribution of function labels over positions is listed in three tables: Table A.5 for dialogue acts by librarians, Table A.6 for dialogue acts by users, and Table A.4 for total. These tables can be overwhelming so the area charts in Figure 4.2 makes it easier to grasp the general trends in distribution of dialogue act functions over progress of the reference interviews. In addition, the area charts in Figure 4.3 illustrate how the contents of the information exchanges between a librarian and a user changes over the progress of reference interviews. The following observations were made from these figures (and confirmed with Tables A.4, A.5 and A.6).

- The use of *Info Provision* by librarians gradually increases until *Mid-ending* of the conversation, and drops down at *Ending*.
- The use of *Info Provision* by users is dominant at *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.

The observations above collectively suggest two hypotheses. First, reference interviews involve the exchange of pieces of information, repeatedly or iteratively; a conclusion previously suggested by studies such as *berrypicking* by Bates (1989) and micro-level information seeking by Wu (2005). Second, these information exchanges have a pattern over time, namely the increase of provision of information regarding information objects from librarians and decrease of provision of problem description from users. Since other kinds of information exchanges are fairly consistent, the researchers hypothesize that the two vectors may be used to characterize the information-seeking interactions. Further studies are desired to investigate how these vectors are related to characteristics of the interviews such as success, effectiveness and user satisfaction. In this regard, the researchers hope to integrate the outcome of this study as well as these measures to the finings from the research project lead by Connaway and Radford (2011). Connaway and Radford have conducted a series of studies investigating the use of virtual reference service (VRS), using multiple data sources, including the same datasets as the ones this study used and 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 VRS, and that query clarification is the key

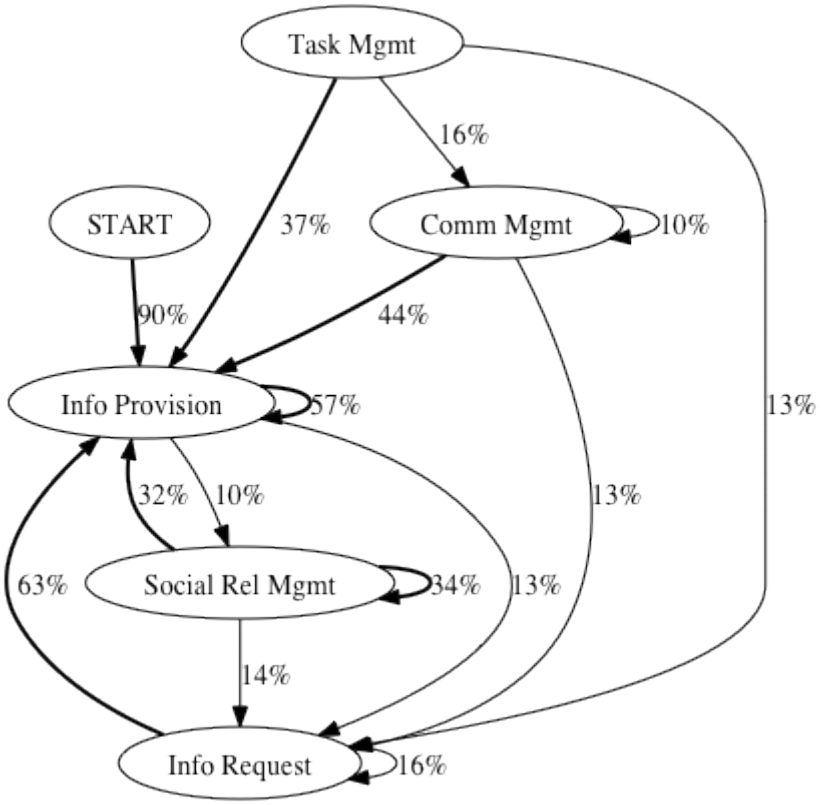

Figure 4.4 Transition of Dialogue Act Functions

for accuracy and effectiveness. If the correlation between the accuracies of VRS and those two vectors is verified, it will be another evidence for these claims.

4.1.2.2 Transition of Dialogue Acts

The transitions of dialogue acts were analyzed utilizing a simple conditional probability, i.e. which type of dialogue act function is likely to follow another. Table A.7 shows the raw frequencies and percentiles of transitions from one dialogue act function to another. Figure 4.4 visualizes the figures. While the figures in the table represent very short and simple transitional sequences they clearly show that the construction of dialog is centered on information provision; the majority of utterances are followed by information provision. The only exception was the *Social Rel Mgmt* function, which was more often followed by another *Social Rel Mgmt* function. This was expected, since the *Social Rel Mgmt* function includes social gestures such as greetings, apologies, gratitude, and valedictions. The transition diagram illustrates the dependency of the distribution of

dialogue act functions to the previous dialogue function, indicating the advantage of machine learning algorithms that utilize sequential states, such as HMM.

4.1.2.3 Document Frequency of Words and Word Sequences

First, document frequencies of words for each function and domain were counted. The words were then sorted based on their frequency. The goal of the analysis was to find out if there are words that uniquely identify a certain function or domain. Table A.8 and A.9 summarize the results. Next the document frequencies of two-word sequences (word bigrams) were counted in the same fashion. Word bi-grams have been previously shown to be effective features for machine learning dialogue acts (Reithinger and Klesen, 1997; Samuel et al., 1998b; Stolcke et al., 2000). As Table A.10 and A.11 show, many of the most frequent bi-grams for each label are indeed unique to that label, indicating they are likely to be helpful in identification and to improve the performance of machine learning versus simple word vectors.

4.2 Machine Learning Experiment

4.2.1 Overview

The goal of the machine learning experiments was two-fold: 1) find the optimal algorithm for recognizing the dialogue acts in the data and 2) find the optimal attributes for recognizing the dialogue acts in the data. Specifically, we tested two algorithms, SVM and SVM-HMM, and seven attributes, the *sequence number*, *speaker*, *message length*, *message position*, *word vector*, and *two-word (bigram) vector*.

The overall structure of the experiment is illustrated in Figure 4.4. The Gold Standard was stored in a database system using MySQL and then processed by Java programs that were developed by the researchers using the Weka software library (Hall et. al 2009) to generate the dataset for machine learning. As for the SVM and HM-SVM implementations, we used Joachims' SVM^{multiclass} for SVM, and SVM^{hmm} for HM-SVM (Joachims 1998, 1999, 2008). All the experiments were subjected to 10-fold cross validation (for each setup, training and testing was repeated ten times). Given the characteristics of the task (high-dimensional, sparse vectors), the linear kernel was used. The Epsilon parameter for the HM-SVM and SVM algorithm, which specifies the required precision to terminate the learning iteration, was set to 0.5, following Joachims's suggestion (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 procedural decisions, selection of the software, and formats of the data, started in December 2011. Weka was initially chosen to be the experiment platform, since it provides immediate access to various other algorithms as well as analytical and evaluation tools. However, Weka did not include HM-SVM for learning algorithms, and thus Joachims's software programs were chosen instead. Since Joachims's SVM^{multiclass} and SVM^{hmm} are both based on a more general version of SVM, SVM^{struct}, the change allowed us to analyze the outcome of the two learning algorithms directly, ensuring continuity of the implementation and parameters for the other parts (e.g. the kernel implementation or the algorithms for the multi-class selection). Two months, from February to March 2012, were spent developing Java programs (to export the data and generate necessary features) and Ruby scripts (providing an environment for the 10-fold cross-validation and analysis). The actual experiments started in May 2012, and ended in June 2012.

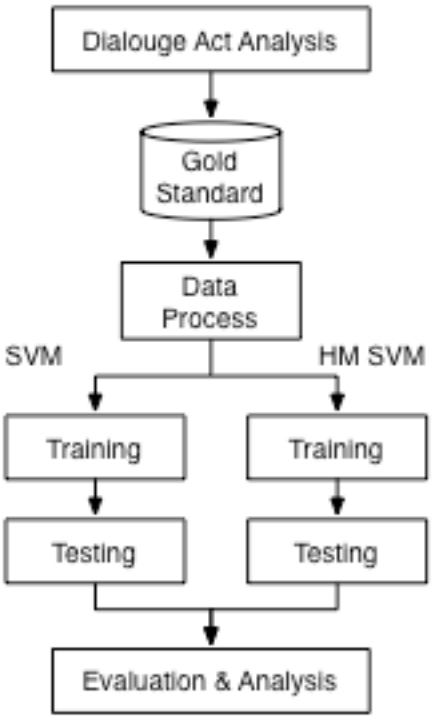

Figure 4.1 Machine Learning Experiment

Given the inter-coder reliability discussed earlier, the experiments were performed at two levels of annotation: 1) Function and 2) Partial Domains. These two levels of annotations received *Kappa > .8* in the dialogue act annotation stage, considered a reliable level of agreement by many researchers. The process depicted in the Figure 4.5 was performed twice.

The following sections describe the outcomes of the machine learning experiments for both levels. 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 sequence 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), and 8) HM-SVM with all of the additional features that had a positive effect (H-XX). The experiments 1) and 2) were to establish that HM-SVM performs better at this task. The experiments 2) through 7) were to verify the individual effect of each additional feature. And lastly, the experiment 8) was to verify the interactive effect of the additional features that improved the result individually. Each setup was named by a prefix that represents 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 (*sequence*: 1, *speaker*: 2, *message length*: 4, *message position*: 8, *word vector*: 16, and *bigram vector*: 32) and the number in each name was the total number of features used for that setup. (So for example, H-17 represents the combination of the HM-SVM algorithm and *sequence* feature (1) and *word vector* feature (16).

4.2.2 Machine Learning Experiments for Dialogue Act Functions

The comparison between S-16 and H-16 shows the substantial advantage HM SVM has over standard SVM. It improved the true positive (TP) rate from 0.6816 to 0.782 (a 0.1166 difference). We believe this is due to the knowledge of the previous dialogue act function giving an advantage in correctly labeling the current dialogue act function, as described earlier. Among the five additional features, three features (*speaker*, *message length,*

and *word bigram vector*) improved the results while the other two (*sequence number* and *message position*) slightly hurt the outcome. We have also shown that the *speaker* feature has a good correlation with the distribution of dialogue act functions. It is common-sense that the *message length* is related to the distribution of dialogue acts. The *bigram vector* feature made by far the most improvement over the baseline system as an individual feature (0.0431, compared to 0.0112 and 0.0093 for the other two), which was also expected given that two-word sequences often characterize utterances. And lastly, the combination of the three features provided the best improvement (0.0510). While this improvement was not as good as the simple sum of the individual improvements of the three features, this due to the three features not being completely independent from each other (e.g. librarians tend to send longer messages and longer messages tend to have higher bi-gram feature values). Overall, the results show that HM SVM successfully learns the annotation of dialogue act functions with appropriate features. The best result the experiment produced was .8784 for precision and .8492 for recall. Given that the inter-coder agreement was .87 (or .92, when the number of dialogue acts matched) we consider this is the best result we could hope for. The results are summarized in Table 4.2.

Setup	TP Rate	FP Rate	Precision	Recall	F-Measure
S-16	0.6816	0.3184	0.9035	0.6816	0.7517
H-16	**0.7982**	**0.2018**	**0.8522**	**0.7982**	**0.8112**
H-17	0.7893	0.2107	0.8508	0.7893	0.8047
H-18	**0.8094**	**0.1906**	**0.8636**	**0.8094**	**0.8225**
H-20	**0.8075**	**0.1925**	**0.8552**	**0.8075**	**0.8198**
H-24	0.7975	0.2025	0.8494	0.7975	0.8106
H-48	**0.8413**	**0.1587**	**0.8687**	**0.8413**	**0.8469**
H-52	**0.8492**	**0.1508**	**0.8784**	**0.8492**	**0.8547**

Table 4.2 Results Summary (Function)

4.2.3 Machine Learning Experiments for Dialogue Act Domains

The experiment for annotating dialogue act domains was done in the same fashion as the one for dialogue act functions. Because of the higher number of labels (nineteen, compared to five) and the lower inter-coder agreement rate (.80, compared to .87), it was expected that domains would be harder to learn to annotate than functions. The results of the experiments for functions are summarized in Table 4.3.

Setup	TP Rate	FP Rate	Precision	Recall	F-Measure
S-16	0.4430	0.5570	0.6776	0.4430	0.4725
H-16	**0.6909**	**0.3091**	**0.7622**	**0.6909**	**0.7144**
S-17	0.6808	0.3192	0.7441	0.6808	0.7018
S-18	**0.7046**	**0.2954**	**0.7839**	**0.7046**	**0.7266**
S-20	0.6826	0.3174	0.7548	0.6826	0.7055
S-24	**0.6946**	**0.3054**	**0.7543**	**0.6946**	**0.7171**
S-48	**0.7178**	**0.2822**	**0.7759**	**0.7178**	**0.7367**
S-58	**0.7400**	**0.2600**	**0.7837**	**0.7400**	**0.7528**

Table 4.3 Results Summary (Domain)

As the table shows the performance for domains were, overall, lower than the performance for functions. The comparison between S-16 and H-16 shows, once again, how HM SVM is far more effective than standard SVM for the task. HM-SVM improved the TP rate from 0.4430 to 0.6906 (0.2479). Among the five additional features three features (*speaker*, *message position,* and *word bigram*) improved the results while the other two (*message length* and *sequence number*) slightly hurt the outcome. As in the analysis section the relative position of a message has a correlation with the dialogue act domain of the message, so this result was expected. The *bigram vector* feature (H-48), once again, made the most improvement over the baseline system (H-16) as an individual feature (0.0269). Interestingly the combination of the three features (H-58) provided more improvement (0.0491) than the sum of all the features' improvements (0.0443); there was an interactive effect. This actually reflects the analysis in the previous section: while the distribution of the dialogue act domains depend on the relative position of a message, the dependency (correlation) seems to be stronger when the speaker is identified (Figure 4.3).

Overall, the experiment results show that HM SVM successfully learns the annotation of dialogue act domains with appropriate features. Although the performance figures are lower than the figures for the functions we consider results satisfactory given that it was a harder problem. The overall best performance (precision: 0.7837, recall: 0.7400, and f-measure: 07528) is comparable to, if not better than, previous similar studies (Surendran and Levow (2006); Lan et al. (2008); Hu et al. (2009)).

5 Conclusion and Further Research

This study investigated the discourse property of digital reference transactions and experimented with automatic identification of dialogue acts using machine learning techniques. The first stage of the investigation revealed how the digital environment has changed the nature of information-seeking interactions by analyzing the dialogue acts of digital reference transactions in multiple aspects. Some researchers have suggested that the development of web information technologies might have changed people's attitudes towards information seeking (e.g. Radford and Connaway (2007)). The analysis presented in this study shows how librarians are responding to these attitudes providing important implications for the design of information services such as the traditional library reference service or the emerging social web systems, as well as training information professionals. The second part of the study, the machine learning experiment, provided the dialogue act model proof of concept 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 could be learned by following certain procedures (algorithms). The experiment employed semi-factorial combinations of different algorithms and features, showing that dialogue acts were machine-learnable. The experiments also demonstrated practical applications of the dialogue act 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.

The presented work will be part of the second author's doctoral dissertation. We plan to submit the presented work to conferences of the relevant fields: library and information science, information retrieval, computational linguistics, and other human language technologies.

As an extension of the presented research, we plan to apply the presented methodology to a broader range of data, such as Twitter conversations. Micro blogging services, such as Twitter, have gained popularity rapidly as a mode of online communication and are known to have created their own norms of communication such as abbreviations and formality (Finin et. al, 2010). Thus, while the increasing data imply a greater opportunity for research, there are challenges as well. As the next step from the presented work, we would like to investigate the information-seeking behavior of Twitter conversations by using the same methodology as presented in this report: analyzing the dialogue acts and applying the machine learning techniques. We believe such study will further validate the findings of the presented work and establish (or refine) them in the broader context of the social web.

6 Acknowledgment

The researchers would like to thank OCLC and ALISE for this opportunity. We would especially like to thank Dr. Lynn Silipigni Connaway for her support and efforts to make this research happen. We are looking forward to continuing our effort on this important thread of research and collaborating with OCLC in the future.

Reference:

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.

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

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

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

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.

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.

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

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

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. Annotating named entities in twitter data with crowdsourcing. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, CSLDAMT ’10, pages 80–88, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

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).

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

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

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.

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 Sch¨olkopf, 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.

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.

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.

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

Radford, M. L. (2006). 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. (2007). "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.

Ruppel, M. and Fagan, J. (2002). Instant messaging reference: users' evaluation of library chat. Reference Services Review, 30(3):183–197.

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.

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.

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.

Wimmer, R.D. and Dominick, J.R. (2006) Mass media research: An introduction, Wadsworth Pub Co

Wu, M. (2005). Understanding patrons' micro-level information seeking (MLIS) in information retrieval situations. Information Processing and Management, 41(4):929 – 947.

Appendices

Dimension	Code		Description
Function	Information Transfer	Request	Requesting information from the recipient.
		Provision	Providing information to the recipient.
	Task Transfer	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.
	Dialogue Management		Managing physical aspects of communication, such as the channel, place, etc.
	Relationship Management		Managing socio-emotional aspects of communication.
Domain	Informational	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 Annotation Scheme

Final report of a 2011 OCLC/ALISE Library and Information Science Research Grant project.
Available online at:
http://www.oclc.org/research/grants/reports/2011/yu2011.pdf

Label	Librarian	User	Total
Info Provision	3788 (51%)	2104 (55%)	5892 (53%)
Social Rel Mgmt	1231 (17%)	785 (21%)	2016 (18%)
Info Request	892 (12%)	314 (8%)	1206 (11%)
Comm 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

Table A.2 Distribution of Dialogue Act Functions by Speaker

Label	Librarian	User	Total
Info:Problem	878 (12%)	1662 (45%)	2540 (23%)
Info:Object	1814 (25%)	174 (5%)	1988 (18%)
Info:Search	1177 (16%)	183 (5%)	1360 (12%)
Info:Other	740 (10%)	232 (6%)	972 (9%)
Info: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%)
Comm:Feedback	243 (3%)	379 (10%)	622 (6%)
Comm:Pausing	431 (6%)	10 (0%)	441 (4%)
Comm:Channel	86 (1%)	26 (1%)	112 (1%)
Total	7321	3729	11050

Table A.3 Distribution of Dialogue Act 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.4 Distribution of Dialogue Act Functions by 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.5 Distribution of Dialogue Act Functions by Position (Librarian)

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.6 Distribution of Dialogue Act Functions by 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 beginning of an interview session.

Table A.7 Transitions of Dialogue Act Functions

Function		Ten Most 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.8 Most Frequent Words by Function

Domain		Five Most 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%)

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

Table A.9a Most Frequent Words by Domain

Domain		Five Most Common Terms
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.9b Most Frequent Words by Domain (Cont.)

Function		Ten Most 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 bigrams.

Table A.10 Most Frequent Word Bigrams by Function

Domain		Five Most 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%)

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

Table A.11a Most Frequent Word Bigrams by Domain

Domain		Five Bi-grams
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 bigrams.

Table A.11b Most Frequent Word Bigrams by Domain (cont.)