Dialogue Act Modeling for Non-Visual Web Access

Vikas Ashok
Dept of Computer Science
Stony Brook University
Stony Brook , New York

Yevgen Borodin
Charmtech Labs LLC
CEWIT SBU R & D Park
Stony Brook , New York

Svetlana Stoyanchev
AT&T Labs Research
New York City, New York
(While at Columbia University)

I V Ramakrishnan
Charmtech Labs LLC
CEWIT SBU R & D Park
Stony Brook , New York

vganjiguntea@cs.sunysb.edu, borodin@charmtechlabs.com,
sstoyanchev@cs.columbia.edu, ram@charmtechlabs.com

Abstract

Speech-enabled dialogue systems have the potential to enhance the ease with which blind individuals can interact with the Web beyond what is possible with screen readers - the currently available assistive technology which narrates the textual content on the screen and provides shortcuts to navigate the content. In this paper, we present a dialogue act model towards developing a speech enabled browsing system. The model is based on the corpus data that was collected in a wizard-of-oz study with 24 blind individuals who were assigned a gamut of browsing tasks. The development of the model included extensive experiments with assorted feature sets and classifiers; the outcomes of the experiments and the analysis of the results are presented.

1 Introduction

The Web is the "go-to" computing infrastructure for participating in our fast-paced digital society. It has the potential to provide an even greater benefit to blind people who once required human assistance with many of their activities. According to the American Federation for the Blind, there are 21.5 million Americans who have vision loss, of whom 1.5 million are computer users (AFB, 2013).

Blind users employ screen readers as the assistive technology to interact with digital content (e.g.., JAWS (Freedom-Scientific, 2014) and VoiceOver (Apple-Inc., 2013)). Screen readers serially narrate the content of the screen using text-to-speech engines and enable users to navigate in the content using keyboard shortcuts and touch-screen gestures.

Navigating content-rich web pages and conducting online transactions spanning multiple pages requires using shortcuts and this can get quite cumbersome and tedious. Specifically, in online shopping a user typically browses through product categories, searches for products, adds products to cart, logs into his/her account, and finally makes a payment. All these steps require screen-reader users listen through a lot of content, fill forms, and find links and buttons that have to be selected to get through these steps. If users do not want to go through all content on the page, they have to remember and use a number of different shortcuts. Beginner users often use the "Down" key to go through the page line by line, listening to all content on the way (Borodin et al., 2010).

Now suppose that blind users were to tell the web browser what they wanted to accomplish and let the browsing application automatically determine what has to be clicked, fill out forms, help find products, answer questions, breeze through checkout, and wherever possible, relieve the user from doing all the mundane and tedious low-level operations such as clicking, typing, etc. The ability to carry out a dialogue with the web browser at a higher level has the potential to overcome the limitations of shortcut-based screen reading and thus offers a richer and more productive user experience for blind people.

The first step toward building a dialogue-based system is the understanding of what users could say and dialogue act modeling. Although dialogue act modeling is a well-researched topic (with details provided in related work - Section 2), it has remained unexplored in the context of web accessibility for blind people. The commercial speech-based applications have been around for a while and new ones continue to emerge at a rapid pace; however, these are mainly stand-alone (e.g.., Apple's Siri) domain specific systems that are not connected to web browsers, which precludes dialogue-based interaction with the Web. Current spoken input modules integrated with web

Proceedings of the SIGDIAL 2014 Conference, pages 123–132,
Philadelphia, U.S.A., 18-20 June 2014.

browsers are limited to certain specific functionalities such as search (e.g.., Google's voice search) or are used as a measure of last resort (e.g.., Siri searching for terms online).

In this paper, we made a principal step towards building a dialogue-based assistive web browsing system for blind people; specifically, we built a dialogue act model for non-visual access to the Web. The contributions of this paper include: 1) a unique dialogue corpus for non-visual web access, collected during the wizard-of-oz user study conducted with 24 blind participants (Section 3); 2) the design of a suitable dialogue act scheme (Section 3); 3) experimentation with classifiers capable of identifying the dialogue acts associated with utterances based on combinations of lexical/syntactic, contextual, and task-related feature sets (Section 4); 4) investigation of the importance of each feature set with respect to classification performance to assess whether simple lexical/syntactic features are sufficient for obtaining an acceptable performance (Section 5).

2 Related Work

While previous research addressed spoken dialogue interfaces for a domain-specific websites, such as news or movie search (Ferreras and Cardeñoso-Payo, 2005; Wang et al., 2014), dialogue interface to generic web sites is a novel task. Spoken dialogue systems (SDS) can be classified by the type of initiative: system, user, or mixed initiative (Lee et al., 2010). In a system-initiative SDS, a system guides a user through a series of information gathering and information presenting prompts. In a user-initiative system, a user can initiate and steer the interaction. Mixed-initiative systems allow both system and user-initiated actions.

Dialogue systems also differ in the types of dialogue manager: finite state based, form based, or agent based (Lee et al., 2010), (Chotimongkol, 2008). Finite state and form filling systems are usually system-initiative. These systems have a fixed set of dialogue states and finite set of possible user commands that map to system actions. In contrast, a speech-enabled browsing system proposed in this work is an agent-based system. The set of actions of this system correspond to user actions during web browsing. The domain of possible user commands at each point of the dialogue depends on the current web page that is viewed by a user. The dialogue state in a voice browsing system is compiled at run-time as the user can visit any web page.

While a users dialogue acts in a form-based or finite state system depends primarily on a dialogue state, in an agent-based system with user-initiative, the space of users dialogue acts at each dialogue state is open. To determine dialogue manager action, it is essential for the system to identify users intent or dialogue act. In this work, we address dialogue act modelling for open-domain voice web browsing as a proof of concept for the system.

Dialogue act (DA) annotation schemes for spoken dialogue systems follow theories on speech acts originally developed by Searle (1975). A number of DA annotation schemes have been developed previously (Core and Allen, 1997), (Carletta et al., 1997). Several of dialogue tagging schemes strive to provide domain-independence (Core and Allen, 1997), (Bunt, 2011).

Bunt (2011) developed a NIST standardized domain-independent annotation scheme which incorporates elements from the previously developed annotation schemes. It is a hierarchical multi-dimensional annotation scheme. Each functional segment (part of an utterance corresponding to a DA) can have a general purpose function, such as Inform, Propositional Question, Yes/No Question, and a dimension-specific function in any number of 10 defined dimensions, such as Task, Feedback, or Time management.

In the analysis of human-computer dialogues, it is common to adopt DA annotation schemes to suit specific domains. Generic domain-independent schemes are geared towards the analysis of natural human-human dialogue and provide rich annotation structure that can cover complexity of natural dialogue. Domain-specific dialogues use a subset of the generic dialogue structure. For example, Ohtake et al. (2009) developed a DA scheme for tourist-guide domain motivated by a generic annotation scheme (Ohtake et al., 2010), and Bangalore and Stent (2009) created a dialogue scheme for a catalogue product ordering dialogue system. In our work we design DA scheme for Web-Browsing domain motivated by the DAMSL (Core and Allen, 1997) schema for task-oriented dialogue.

We used a Wizard-of-Oz (WOZ) approach to collect an initial dataset of spoken voice com-

Task	τ_u	τ_d
Shopping	121	16
Email	92	16
Flight	180	16
Hotel	179	16
Job	76	16
Admission	144	16
Overall	792	96

Table 1: Corpus details. τ_u - number of utterances, τ_d - number of dialogs.

mands by both blind and sighted users. WOZ is commonly used before building a dialogue system (Chotimongkol, 2008), (Ohtake et al., 2009), (Eskenazi et al., 1999).

In previous work on dialogue modelling, Stolcke et al. (2000) used HMM approach to predict dialogue acts in a switchboard human-human dialogue corpus achieving 65% accuracy. Rangarajan Sridhar et al. (2009) applied a maximum entropy classifier on the Switchbord corpus. Using a combination of lexical, syntactic, and prosodic features, the authors achieve accuracy of 72% on that corpus. Following the work of Rangarajan Sridhar et al. (2009), we use supervised classification approach to determine dialogue act on the annotated corpus of human-wizard web-browsing dialogues.

3 Corpus and Annotation

In this section, we describe the corpus and the associated dialogue act scheme. The corpus was collected using a WOZ user study with 24 blind participants. Exactly 50% of the participants indicated that they were very comfortable with screen readers, while the remaining 50% said they were not comfortable with computers. We will refer to them as "experts" and "beginners" respectively.

The study required each participant to complete a set of typical web browsing tasks (shopping, sending an email, booking a flight, reserving a hotel room, searching for a job and applying for university admission) using unrestricted speech commands ranging from simple commands such as "click the search button", to complex commands such as "buy this product". Unknown to the participants, these commands were executed by a wizard and appropriate responses were narrated using a screen reader. The dialogs were effective; almost every participant was able to complete each assigned task by engaging in a dialogue with the wizarded interface.

As shown in Table 1, the corpus consists of a total of 96 dialogs collected during the execution of 6 tasks and captures approximately 22 hours of speech with a total of 792 user utterances and 774 system utterances. There is exactly 1 dialogue per task for any given participant. Each user turn consists of a single command that is usually a simple sentence or phrase. Each system turn is either narration of webpage content or information request for the purpose of either form filling or disambiguation. Therefore, each dialogue turn was treated as a single utterance and every utterance was identified with a single associated dialogue act.

The corpus was manually annotated with dialogue act labels and the labeling scheme was verified by measuring the inter-annotator agreement. The rest of this section describes the annotation scheme.

3.1 Dialogue Act Annotation

The dialogue act annotation scheme was inspired by the DAMSL scheme (Core and Allen, 1997) for task oriented dialogue. The proposed scheme was also influenced by extended DAMSL tagset (Stolcke et al., 2000) and the DIT++ annotation scheme (Bunt, 2011). We customized the annotation scheme to suit the non-visual web access domain, thereby making it more relevant to our corpus and tasks.

Table 2 lists the dialogue acts for both user and system utterances. The user dialogue act tagset consists of labels representing task related requests (Command-Intention, Command-Task, Command-Multiple, Command-Navigation), inquiries (Question-Task, Help-Task) and information input (Information-Task), whereas the system DA tagset contains labels representing information requests (Prompt), answers to user inquiries (Question-Answer, Help-Response) and other system responses (Short-Response, Long-Response, etc.) to user commands.

Inter-rater agreement values for different tasks in the corpus are presented in Table 3. The κ values for all tasks are above 0.80, which according to Fleiss' guidelines (Fleiss, 1973), indicates excellent inter-rater reliability on the DA annotation. Therefore, the DA tagset is generic enough to be applicable for a wide varity of tasks that can be performed on the web. Note that the dialogue act scheme was specially designed for non-visual web

User dialogue Acts		
Dialogue Act	**Description**	**Frequency**
Command-Intention	Indication of user's intention or end goal, e.g. *I wish to buy a Bluetooth speaker*	0.117
Command-Task	Basic action commands like *click*, *select*, *enter*, etc.	0.072
Command-Multiple	Complex commands requiring an execution plan comprising a sequence of basic commands, e.g. *buy this product*, *book this room*, etc.	0.162
Command-Navigation	Commands directing the movement of cursor like *go to*, *stop*, *next* etc.	0.136
Information-Task	Information required for completing a task, e.g. *departure date/return date* information for flight booking task, *first name*, *phone number*, etc.	0.442
Question-Task	Task specific questions like *What is the cheapest flight?*, *What is the basic salary?*, etc.	0.041
Self-Talk	Utterances not directed towards the system, e.g. *hmmm, what should I do next?*	0.002
Help-Task	Request for help when the user wishes to speak with the experimenter, e.g. *Help, what does that mean?*	0.024
System dialogue Acts		
dialogue Act	**Description**	**Frequency**
Prompt	Request for information from user to complete a task, e.g. *First Name, text box blank*	0.460
Short-Response	A short response to a user command, e.g. *description of product*, *brief details of flight*, *acknowledgements*, etc.	0.198
Long-Response	A lengthy response to a user command, e.g. *Narration of entire page*, *list of search results*, etc.	0.120
Keyboard-Response	Response to user keyboard actions	0.072
Article-Response	Narration of an article	0.034
Question-Answer	Response to a user question regarding task (non-help)	0.044
No-Response	No response for some navigation commands like *Stop*	0.041
Help-Response	Response to a help request from the user	0.026

Table 2: dialogue acts for non-visual Web access

access. Insofar as sighted people are concerned, a more elaborate scheme would be required since their utterances are dominated by visual cues, a fact that was confirmed by a parallel user study with sighted participants on the same set of web tasks that were used in the wizard-of-oz study.

4 Features

This section describes the different feature sets that we experimented with for our classification tasks. The vector representation for training the DA classifiers integrates several types of features (Table 4): unigrams ($\mathcal{U}$) and syntactic features ($\mathcal{S}$), context related features ($\mathcal{C}$), task related features ($\mathcal{T}$), presence of words anywhere in an utterance($\mathcal{P}$) and presence of words at the beginning of an utterance($\mathcal{B}$). The last two feature sets are similar to the ones used in Boyer et al. (2010).

Task	κ
Shopping	0.865
Email	0.829
Flight	0.894
Hotel	0.848
Job	0.824
Admission	0.800

Table 3: Inter-rater agreement measured in terms of Cohen's κ for all tasks in the corpus.

The feature sets $\mathcal{C}$, $\mathcal{P}$, $\mathcal{B}$ and $\mathcal{S}$ are specific to the domain of non-visual web access and were hand-crafted based on the following three factors: knowledge of the browsing behavior of blind users reported in previous studies, e.g. (Borodin et al., 2010); manual analysis of the corpus; mitigate the effect of noise that is usually present in standard lexical/syntactic feature sets such as n-grams and parse tree rules. Each of the features in $\mathcal{C}$, $\mathcal{P}$, $\mathcal{B}$ and $\mathcal{S}$ were crafted to have a close correspondence to some dialogue act. For example, p_{nav} is closely tied to the *Command-Navigation* dialogue act.

4.1 Unigrams

Unigrams ($\mathcal{U}$ in Table 4) are one of the commonly used lexical features for training dialogue act classifiers (e.g. (Boyer et al., 2010), (Stolcke et al., 2000), (Rangarajan Sridhar et al., 2009)). Encoding unigrams as features is based on the observation that some words appear more frequently in certain dialogue acts compared to other dialogue acts. For example, approximately 73% of "*want*" occur in the *Command-Intention* DA, 100% of "*skip*" occur in the *Command-Navigation* DA and approximately 92% of "*select*" occur in the *Command-Task* DA. Word-DA corrections can also be automatically identified using SVM classifers trained on unigram features. Table 5

Overall Feature Set		
UNIGRAMS ($\mathcal{U}$)		
Feature	**Description**	**Binary**
u	Unigrams	N
PRESENCE OF WORDS IN COMMANDS ($\mathcal{P}$)		
p_{iyou}	The utterance contains either *I* or *you*	Y
p_{help}	The utterance contains the word *help*	Y
p_{helpq}	The utterance contains words usually associated with help requests. E.g., *how*, *am I*, etc.	Y
p_{prev}	The immediately preceding system DA is *Prompt* and the utterance contains words also present in this immediately preceding system utterance	Y
p_{intent}	The utterance contains words , *need*, *desire*, *prefer*, *like* and their synonyms	Y
$p_{browser}$	The utterance contains words also present in the web browser tab title. E.g., *email*, *job*	Y
p_{html}	The utterance contains references to HTML elements. E.g., *form*, *box*, *link*, *page*, etc.	Y
p_{basic}	The utterance contains a verb representing basic operations on a web page. E.g., *click*, *edit*.	Y
p_{nbasic}	The utterance contains a verb not related to basic web page operations; a verb usually associated with task or domain related actions. E.g. *send*, *open*, *compose*, etc.	Y
p_{nav}	The utterance contains words related to cursor movement. E.g., *go to*, *continue*, *next*, etc.	Y
$p_{question}$	The utterance contains words usually associated with questions. E.g., *what*, *when*, *why*	Y
SYNTACTIC STRUCTURE OF COMMANDS ($\mathcal{S}$)		
s_{np}	The utterance is a noun phrase with atleast two words	Y
s_{noun}	The utterance consists of a single noun	Y
s_{basic}	The utterance consists of a single verb representing basic web page operations. E.g., *click*, *edit*, *erase*, *select*, etc.	Y
s_{nbasic}	The utterance consists of a single verb representing task or domain related actions. e.g. *send*, *open*, *compose*, *order*, etc.	Y
CONTEXT RELATED FEATURES ($\mathcal{C}$)		
c_{first}	The utterance is the first command to be issued when a new website is loaded in the browser	Y
$c_{previous}$	dialogue act of the immediately preceding system utterance	N
POSITION OF WORDS IN COMMANDS ($\mathcal{B}$)		
b_{nav}	The utterance begins with word(s) related to cursor movement. e.g. *go to*, *continue*, etc.	Y
$b_{question}$	The utterance begins with a word that is usually associated with a question. E.g., *what*, *when*, *where*, *why*, etc.	Y
b_i	The utterance begins with the personal pronoun *I*.	Y
b_{helpq}	The utterance begins with word(s) usually associated with help requests. E.g., *how*, *am I*	Y
TASK RELATED FEATURES ($\mathcal{T}$)		
t_{name}	Name of the task associated with the utterance	N

Table 4: Feature set for user dialogue act classification. The complete list of words associated with each feature in $\mathcal{P}$ and $\mathcal{B}$ is provided in Appendix A.

presents few such correlations. Note that some of the words in Table 5 are task-specific (noise); a consequence of using a small dataset.

4.2 Presence of Words in Commands

In constract to unigram features that take into account all possible word-DA correlations, the presence-of-word features ($\mathcal{P}$ in Table 4) are limited to certain specific words that have strong correlations with the DA types. For each feature $p \in \mathcal{P}$, if the presence of certain specific words associated with p occur in an utterance, then p is set to *true*. The set of words for every p that corresponds to some dialogue act d was contructed by determining the discriminatory words for d using simple statistical analysis of the corpus (e.g. relative frequencies of words) as well as by an examination of the weights of different words learnt by the SVM classifier trained on a development dataset using unigram features alone. e.g.., the words *continue* and *skip* occur much more frequently in Command-Navigation than in other dialogue acts (see Table 5) and hence are included in p_{nav}. Note however that not all discriminatory words in Table 5 were used. Only generic words, independent of any specific task, were selected (see Appendix A for details).

4.3 Syntactic Structure of Commands

The binary syntactic features ($\mathcal{S}$ in Table 4) were automatically extracted using the Stanford parser (Klein and Manning, 2003). As in word-DA correlations, some of the syntactic structure-DA correlations were also identified by a manual in-

Dialogue Act	Discriminatory Words
Command-Intention	want, compose, book, for, look, email, find, an, accounting, Stanford, a, airplane, message, I, music, get, ticket, positions, need, bluetooth, jobs, new
Command-Task	repeat, choose, delete, select, link, edit, enter, erase, clear, fill, in, click, third, at, body, box, again, blue, that
Command-Multiple	play, read, senior, send, reviews, Harlem, artists, study, submit, details, law, description, Kitaro, mornings, availability, apply, construction, pay, reservations, proceed, it, this, available
Comand-Navigation	skip, next, previous, go, page, finish, stop, item, continue, back, line, before, box, first, second, to, top, home, part, would
Information-Task	JFK, customer, no, August, July, USA, October, Kahalui, October30th, anytime, coach, today, nonstop, movies, York
Question-Task	price, time, fare, layover, times, is, what's, anything, cheaper, best, flight, airline, complete, onestop, departure, cards, price, much, cost, weekly.
Help-Task	help, do, mean, does, say, can, supposed, something, how, use, voice, have, apply, reservation, by, address, give, get

Table 5: Top discriminative unigrams based on weights from SVM classifier.

vestigation of the corpus. For example, 82.1% of single noun-only utterrances (s_{noun}) have the DA *Information-Task*, 76.2% of "*basic*" verb-only utterances (s_{basic}) have the DA *Command-Task* and 83.3% of "*non-basic*" verb-only utterances (s_{nbasic}) have the DA *Command-Multiple*.

4.4 Context Related Features

The local context ($\mathcal{C}$ in Table 4) provides valuable cues to identify the dialogue act associated with a user utterance. It was observed during the study that user utterance is influenced to a large extent by the immediately preceding system utterance. For example, 89.95% of all user utterances immediately following the system *Prompt* were observed to be *Information-Task*. In addition, most of the time (probability 87.5%), the first utterance issued for a task was *Command-Intention*.

4.5 Position-of-Word in Commands

Design of feature set $\mathcal{B}$ in Table 4 was inspired by an analysis of the corpus which revealed that certain dialogue acts are characterized by the presence of certain words at the beginning of the corresponding utterances. For example, 93.4% of all *Command-Navigation* utterances begin with a cursor-movement related word (e.g. next, previous, etc. see Appendix A for the complete list).

4.6 Task Related Features

Since it is possible for different tasks to exhibit different feature vector patterns for the same dialogue act, incorporating task name ($\mathcal{T}$ in Table 4) as an additional feature may therefore improve classification performance by exploiting these variations (if any) between tasks.

Group	Composition
$\mathcal{G}1$	$\mathcal{U}$
$\mathcal{G}2$	$\mathcal{P} \cup \mathcal{B} \cup \mathcal{S}$
$\mathcal{G}3$	$\mathcal{C} \cup \mathcal{B} \cup \mathcal{S}$
$\mathcal{G}4$	$\mathcal{C} \cup \mathcal{P} \cup \mathcal{S}$
$\mathcal{G}5$	$\mathcal{C} \cup \mathcal{P} \cup \mathcal{B}$
$\mathcal{G}6$	$\mathcal{C} \cup \mathcal{P} \cup \mathcal{B} \cup \mathcal{S}$
$\mathcal{G}7$	$\mathcal{C} \cup \mathcal{P} \cup \mathcal{B} \cup \mathcal{S} \cup \mathcal{T}$
$\mathcal{G}8$	$\mathcal{C} \cup \mathcal{P} \cup \mathcal{B} \cup \mathcal{S} \cup \mathcal{U}$

Table 6: Feature groups.

5 Classification Results

All classification tasks were performed using the WEKA toolkit (Hall et al., 2009). The classification experiments were done using Support Vector Machine (frequently used for benchmarking), J48 Decision Tree (appropriate for a small size mostly binary feature set) and Random Forest classifiers. The model parameters for all classifiers were optimized for maximum performance.

In addition, experiments were also performed to assess the utility of each feature set (Table 4). Specifically, the performance of classifiers with different combinations (Groups 1-8 in Table 6) of feature sets was evaluated to assess the importance of each individual feature set. We primarily focussed on domain-specific feature sets ($\mathcal{P}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{S}$). Observe that group $\mathcal{G}6$ differs from any of $\mathcal{G}2 - \mathcal{G}5$ by exactly one feature set. This lets us to assess the individual utility of $\mathcal{P}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{S}$. In addition, we also extended $\mathcal{G}6$ by including $\mathcal{U}$ ($\mathcal{G}7$) and $\mathcal{T}$ ($\mathcal{G}8$) to determine if there was any noticeable improvement in performance. $\mathcal{G}1$ with only unigram features serves as a baseline. All reported results (Table 7) are based on 5-fold cross validation: 632 instances for training and 158 instances for testing. Table 7 presents the classification results for different feature groups. The DA *Self-Talk* was excluded from classification due to insufficient number (2) of data points.

5.1 Classification Performance

Overall Performance: As seen in Table 7, the tree-based classifiers (J48 and RF) performed better than SVM in a majority of the feature groups (6 out of 8). The random forest classifier yielded the best performance (91% Precision, 90% Recall) for feature group $\mathcal{G}6$, whereas the $\mathcal{G}3$-SVM combination had the lowest performance (69% Precision, 67% Recall). However, all groups includ-

DA	MODEL	Performance of Feature Groups															
		$\mathcal{G}1$		$\mathcal{G}2$		$\mathcal{G}3$		$\mathcal{G}4$		$\mathcal{G}5$		$\mathcal{G}6$		$\mathcal{G}7$		$\mathcal{G}8$	
		P	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
CI	SVM	.83	.80	.84	.95	.71	.95	**.91**	**.96**	.82	.90	**.91**	.95	.89	**.96**	.89	.94
	J48	.74	.74	.83	.90	.80	.93	.84	.95	.81	.93	.83	.95	.85	.93	**.91**	.95
	RF	.76	.74	.81	.90	.85	.94	.88	.90	.80	.87	.84	.93	.88	.89	.87	.95
CT	SVM	.87	.73	.86	.81	.93	.30	.89	.87	.84	.81	.89	.83	.89	.81	.92	.88
	J48	.80	.64	.80	.70	**1.0**	.28	.88	.79	.80	.70	.85	.75	.83	.87	.86	.67
	RF	.72	.58	.84	.89	.81	.26	.88	.89	.85	.85	.79	**.93**	.77	.78	.88	.80
CM	SVM	.73	.65	.77	.58	.36	.30	.78	.64	.78	.59	.78	.64	.80	.62	.79	.78
	J48	.74	.36	.78	.79	.68	.87	**.83**	.59	.81	.78	.76	.83	.81	.80	.76	.87
	RF	.79	.56	.80	.81	.68	.83	.80	.59	.82	.79	.81	.83	.80	.82	.76	**.89**
CN	SVM	.89	.84	.93	.87	**.96**	.82	.67	**.96**	.94	.87	**.96**	.89	.94	.87	.90	.92
	J48	.89	.65	.95	.95	**.96**	.92	.65	.93	.95	.95	.95	.92	.92	.93	.87	.90
	RF	.82	.86	.94	.94	.95	.92	.66	.95	.95	.95	.95	.95	.94	.93	.91	.88
IT	SVM	.70	.89	.82	.93	.70	.81	.81	.79	.82	.93	.82	.93	.82	.94	.85	.90
	J48	.54	.93	.96	.97	.94	.97	.80	.82	.96	.97	.97	.96	.96	.97	.94	.94
	RF	.65	.93	**.98**	**.98**	.95	.97	.81	.82	.97	**.98**	**.98**	.97	**.98**	**.98**	.97	.92
QT	SVM	.66	.46	.87	.27	.90	.30	.80	.30	.62	.31	.80	.31	.70	.33	.85	.49
	J48	.44	.36	.62	.33	.80	.23	.90	.30	.53	.34	.62	.31	.56	.47	**.93**	.32
	RF	.63	.36	.65	.31	.61	.39	.78	.27	.54	.35	.83	.39	.68	**.51**	.87	.33
HT	SVM	.77	.71	.73	.65	.80	.45	.79	.63	.63	.67	.78	.63	.72	.64	.92	.76
	J48	.86	.79	.80	.57	.80	.33	.81	.60	.70	.50	.81	.55	.55	.52	**.93**	**.91**
	RF	.85	.70	.79	.65	.78	.33	.75	.60	.74	.67	.90	.48	.67	.67	.90	.80
Overall	SVM	.77	.76	.83	.82	.69	.67	.80	.79	.82	.82	.84	.83	.84	.83	.85	.85
	J48	.70	.66	.88	.88	.87	.85	.80	.78	.88	.88	.89	.88	.88	.89	.87	.86
	RF	.74	.73	.90	**.90**	.86	.85	.80	.79	.89	.89	**.91**	**.90**	.90	.89	.88	.87

Table 7: Classification Results. The overall performance is the weighted average over all dialogue acts. Notation: J48-Decision Tree, RF-Random Forest, SVM-Support Vector Machine, P-Precision, R-Recall, CI-Command-Intention, CT-Command-Task, CM-Command-Multiple, CN-Command-Navigation, IT-Information-Task, QT-Question-Task, HT-Help-Task. The best performances for each DA are highlighted in bold.

ing $\mathcal{G}3$ did better than $\mathcal{G}1$ with tree-based classifiers. $\mathcal{G}1$ was consistently outperformed by the other groups.

Performance on dialogue acts: In $6/8$ feature groups, the performance of SVM with respect to IT dialogue act was significantly worse than that of tree-based classifiers. However, SVM produced consistently good results ($> 80\%$ in most cases) for the CI and CT dialogue acts. All classifiers performed very well in case of CN dialogue act ($> 80\%$ for $7/8$ groups). However, none of the classifiers performed well in case of QT.

5.2 Importance of feature sets

From Table 7, it can be inferred that contextual features ($\mathcal{C}$) do not contribute to improving overall classification performance. In particular, for each classifier, the difference in overall performance between groups $\mathcal{G}2$ (excluding $\mathcal{C}$) and $\mathcal{G}6$ (including $\mathcal{C}$) is very small (worst case: 1% difference in both P and R). However, inclusion of $\mathcal{C}$ significantly improved the classification performance of RF for QT and CI dialogue acts (18% improvement in P, 8% improvement in R for QT, 3% improvement in both P and R for CI). Even in case of J48, where group $\mathcal{G}6$ yields the best performance,

Dialogue Act	Discriminatory Rules
Command-Intention	• $c_{first} \wedge \neg b_{nav} \wedge \neg p_{html} \wedge \neg s_{noun}$ • $c_{first} \wedge \neg b_{nav} \wedge p_{html} \wedge p_{iyou}$ • $\neg c_{first} \wedge \neg b_{nav} \wedge p_{intent} \wedge \neg p_{nav} \wedge \neg p_{question}$
Command-Task	• $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge \neg p_{helpq} \wedge p_{basic} \wedge \neg p_{nbasic}$ • $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge \neg p_{helpq} \wedge p_{basic} \wedge p_{nbasic} \wedge p_{html}$
Command-Multiple	• $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge \neg p_{helpq} \wedge \neg p_{basic} \wedge \neg p_{nbasic} \wedge c_{previous} = [h\|k\|l\|n] \wedge \neg p_{html} \wedge \neg p_{question}$ • $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge \neg p_{helpq} \wedge \neg p_{basic} \wedge p_{nbasic} \wedge c_{previous} = [\hat{}p]$
Comand-Navigation	• $c_{first} \wedge b_{nav}$ • $c_{first} \wedge \neg b_{nav} \wedge p_{html} \wedge \neg p_{iyou}$ • $\neg c_{first} \wedge b_{nav} \wedge \neg s_{np}$ • $\neg c_{first} \wedge b_{nav} \wedge s_{np} \wedge c_{previous} = [s\|a]$
Information-Task	• $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge \neg p_{helpq} \wedge \neg p_{basic} \wedge \neg p_{nbasic} \wedge c_{previous} = [p]$ • $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge \neg p_{helpq} \wedge \neg p_{basic} \wedge p_{nbasic} \wedge c_{previous} = [p] \wedge \neg p_{iyou}$
Question-Task	• $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge \neg p_{helpq} \wedge \neg p_{basic} \wedge \neg p_{nbasic} \wedge c_{previous} = [h\|k\|l\|n] \wedge \neg p_{html} \wedge p_{question}$ • $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge \neg p_{helpq} \wedge \neg p_{basic} \wedge \neg p_{nbasic} \wedge c_{previous} = [q\|s\|a] \wedge \neg p_{nav} \wedge \neg p_{html} \wedge \neg s_{noun}$
Help-Task	• $\neg c_{first} \wedge \neg b_{nav} \wedge \neg p_{intent} \wedge p_{helpq} \wedge p_{iyou} \wedge \neg b_i$

Table 8: A select sample of J48 rules ($conf \geq 0.75$ and descending order of support) for group $\mathcal{G}6$. Notation: $\neg c_{first}$ stands for $c_{first} = false$ and c_{first} stands for $c_{first} = true$.

Utterance	Actual DA	Predicted DA	Comments
"*Continue to booking it*"	Command-Multiple	Command-Navigation	This utterance was issued while performing the *book a hotel room* task. This command essentially is the same as "book it". The presence of a navigation related verb *continue* at the beginning caused the classifiers to incorrectly classify it as *Command-Navigation*.
"*I am looking to check in on July 23rd*"	Information-Task	Command-Intention	This utterance was in response to a system prompt for check-in date while performing the *book a hotel room* task. The presence of first person nominative pronoun "I" caused the classifiers to categorize it as *Command-Intention*.
"What does that mean?"	Help-Task	Question-Task	This utterance was directed towards the experimenter and therefore it was annotated as *Help-Task*. However, the absence of the keyword *help* and the presence of a Wh-word *what* at the beginning of the command caused the classifiers to incorrectly classify this command as *Question-Task*.
"*Best available price?*" "*Ok, return time?*" "*Price?*" "*Layover?*"	Question-Task	Command-Multiple Information	The absence of Question related words like *Wh-words*, *is*, etc. at the beginning coupled with the fact that these commands are *noun phrases* caused the classifiers to incorrectly classify them as either *Command-Multiple* or *Information*.

Table 9: A few incorrectly classified utterances.

contextual features were found to be a component of some of the high-confidence, high-support J48 rules (Table 8) for CI and QT. Similar claims can also be made for syntactic features($\mathcal{S}$), where although there is not much difference in overall performance between groups $\mathcal{G}5$ and $\mathcal{G}6$ (Worst Case: 2% drop in P, 1% drop in R), improvements were observed in case of RF for QT and CI dialogue acts (29% improvement in P, 4% improvement in R for QT, 4% improvement in P, 6% improvement in R for CI).

Excluding either word-existential features ($\mathcal{P}$) or word-position related features ($\mathcal{B}$), however, caused a significant drop in overall performance (Worst case: 15% drop in P, 16% drop in R without $\mathcal{P}$, 11% drop in both P and R without $\mathcal{B}$). Table 8 further highlights the importance of feature set $\mathcal{P}$, since over 50% of the high performing J48 rules (Table 8) have at least one feature of type $\mathcal{P}$ with *true* as their truth values.

It can be seen in Table 7 that adding either unigrams or task-name to the existing feature set of $\mathcal{G}6$ does not affect the overall performance. However, the use of unigram features improved results of all the classifiers for the HT DA. No such DA specific improvements were seen with task-name as an added feature to $\mathcal{G}6$. This suggests that the feature values of $\mathcal{G}6$ for all DAs are *task-independent*.

5.3 Prediction Errors

It is clear from Table 7 that the prediction accuracies of CM, QT and HT are not nearly as good as those of other dialogue acts. Table 9 provides some insights into this issue via illustrative examples from the corpus.

Notice that the errors in case of CI, CM and HT are mostly related to choice of words used in the utterances, whereas mistakes in the prediction of QT are mainly due to inadequate information or the incompleteness of the utterances. Therefore, it is recommended that the speech enabled web dialogue systems enforce a constraint requiring users to express their complete thoughts in each of their corresponding utterances.

6 Conclusion

Experiments with the dialogue act model described in the paper indicate that with a small set of simple lexical/syntactic features it is possible to achieve a high overall dialogue act recognition accuracy (over 90% precision and recall) using simple and well-known tree-based classifiers such as decision trees and random forests. It is hence possible to build speech-enabled dialogue-based assistive web browsing systems with low computational overhead that, inturn, can result in low latency response times - a critical requirement from a usability perspective for blind users. Finally, a dialogue model for non-visual web access, such as the one described in this paper, can be the key driver of goal-oriented web browsing - a next generation assistive technology that will empower blind users to stay focused on high-level browsing tasks, while the system does all of the low-level operations such as clicking on links, filling forms, etc., necessary to accomplish the tasks.

Acknowledgements

Research reported in this publication was supported by the National Eye Institute of the National Institutes of Health under award number 1R43EY21962-1A1. We would like to thank Lighthouse Guild International and Dr. William Seiple in particular for helping conduct user studies.

References

AFB. 2013. Facts and figures on american adults with vision loss. http://www.afb.org/info/blindness-statistics/adults/facts-and-figures/235, January.

Apple-Inc. 2013. Voiceover for os x. http://www.apple.com/accessibility/osx/voiceover/.

Srinivas Bangalore and Amanda J Stent. 2009. Incremental parsing models for dialog task structure. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102. Association for Computational Linguistics.

Yevgen Borodin, Jeffrey P Bigham, Glenn Dausch, and IV Ramakrishnan. 2010. More than meets the eye: a survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, page 13. ACM.

Kristy Elizabeth Boyer, Eun Young Ha, Robert Phillips, Michael D Wallis, Mladen A Vouk, and James C Lester. 2010. Dialogue act modeling in a complex task-oriented domain. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–305. Association for Computational Linguistics.

Harry Bunt. 2011. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2):222–245.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

Ananlada Chotimongkol. 2008. *Learning the structure of task-oriented conversations from the corpus of in-domain dialogs*. Ph.D. thesis, SRI International.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, pages 28–35. Boston, MA.

Maxine Eskenazi, Alexander I Rudnicky, Karin Gregory, Paul C Constantinides, Robert Brennan, Christina L Bennett, and Jwan Allen. 1999. Data collection and processing in the carnegie mellon communicator. In *EUROSPEECH*.

César González Ferreras and Valentín Cardeñoso-Payo. 2005. Development and evaluation of a spoken dialog system to access a newspaper web site. In *INTERSPEECH*, pages 857–860.

J.L. Fleiss. 1973. *Statistical methods for rates and proportions Rates and proportions*. Wiley.

Freedom-Scientific. 2014. Screen reading software from freedom scientific. http://www.freedomscientific.com/products/fs/jaws-product-page.asp.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Cheongjae Lee, Sangkeun Jung, Kyungduk Kim, Donghyeon Lee, and Gary Geunbae Lee. 2010. Recent approaches to dialog management for spoken dialog systems. *JCSE*, 4(1):1–22.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka, and Satoshi Nakamura. 2009. Annotating dialogue acts to construct dialogue systems for consulting. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 32–39. Association for Computational Linguistics.

Kiyonori Ohtake, Teruhisa Misu, Chiori Hori, Hideki Kashioka, and Satoshi Nakamura. 2010. Dialogue acts annotation for nict kyoto tour dialogue corpus to construct statistical dialogue systems. In *LREC*.

Yury Puzis, Yevgen Borodin, Rami Puzis, and IV Ramakrishnan. 2013. Predictive web automation assistant for people with vision impairments. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1031–1040. International World Wide Web Conferences Steering Committee.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth Narayanan. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, 23(4):407–422.

John R Searle. 1975. Indirect speech acts. *Syntax and semantics*, 3:59–82.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Lu Wang, Larry Heck, and Dilek Hakkani-Tur. 2014. Leveraging semantic web search and browse sessions for multi-turn spoken dialog systems.

A List of Words Predictive of Dialogue Acts

Table 10 lists all the words associated with presence-of-word ($\mathcal{P}$) and position-of-word ($\mathcal{B}$) related features (Table 4) used in this work. Notice that all words specified in Table 10 are task-independent. This ensures that the proposed feature set is generic enough to be applicable for a wide variety of tasks on the web. The proposed list of words can be easily extended by adding synonyms, which can be obtained automatically from publicly available sources like WordNet (Miller, 1995).

Features	Predictive Words
p_{iyou}	*I, you*
p_{help}	*help*
p_{helpq}, b_{helpq}	*how*, *can*, *do*, *am I*
p_{prev}	dynamically determined at runtime
p_{intent}	*want*, *like*, *would*, *need*, *prefer*
$p_{browser}$	dynamically determined at runtime
p_{html}	*body*, *page*, *form*, *box*, *field*, *search*, *link*, *button*, *list*, *dropdown*
p_{basic}	*clear*, *select*, *fill*, *delete*, *click*, *edit*, *erase*, *submit*, *repeat*, *choose*, *enter*, *check*
p_{nbasic}	any verb not in the p_{basic} list above
p_{nav}, b_{nav}	*skip*, *go to*, *next*, *first*, *last*, *back*, *continue*, *previous*, *stop*, *go back*, *finish*, *home page*
$p_{question}$, $b_{question}$	*what*, *where*, *why*, *when*, *how*

Table 10: Complete list of predictive words for features in $\mathcal{P}$ and $\mathcal{B}$ of Table 4.

As explained earlier, the words in Table 10 were selected by performing simple statistical analysis of corpus and also by examining the word-weights produced by the SVM classifier trained on unigram features alone. In other words, some of the words in Table 10 were borrowed from Table 5 that lists discriminatory unigrams for different dialogue acts. Note that the task-dependent words (e.g. "Stanford", "airplane", etc.) in Table 5 were ignored while constructing Table 10.

B Sample Dialogue in the Corpus

Table 11 presents an example of a dialogue that was collected during the execution of the *Shopping* task by a participant in the Wizard-Of-Oz study. For deeper understanding, the wizard actions for every user utterance are also listed.

Subject:	"I want to buy CD"
Wizard:	[enter 'CDs' in the search box, clicks 'Go', and narrate search results one by one]
Screen reader:	Verbatim 97458 700 MB 80 Minute . . . Disc CD-R by Verbatim link . . .
Subject:	"Search for new age items"
Wizard:	[searches for 'new age items', clicks on results]
Screen reader:	Age of Wushu - Free Amazon . . . Game Connect link by Snail Games USA . . .
Subject:	"stop"
Wizard:	[presses pause shortcut]
Subject:	"new age music CDs"
Wizard:	[searches for 'new music CDs']
Screen reader:	The ultimate Most Relaxing New Age . . . .
Subject:	"new age music CDs by Kitaro"
Wizard:	[Searches for 'new music cds by Kitaro']
Screen reader:	Most Relaxing New Age Music link by Kitaro, List 8 items, $10.87 link, order in the next 3 hours . . . Tenku link by Kitaro . . .
Subject:	"sample recordings"
Wizard:	[follows the current link]
Screen reader:	Page loading
Wizard:	[plays the first sample track]
Screen reader:	[audio sample]
Subject:	"next"
Wizard:	[plays the next sample track]
Screen reader:	[audio sample]
Subject:	"next"
Wizard:	[plays the next sample track]
Screen reader:	[audio sample]
Subject:	"go back"
Wizard:	[presses back button]
Screen reader:	page loading, [repeats the visited link]
Subject:	"next CD"
Wizard:	[clicks the title of the next item in search result]
Screen Reader:	Ancient link by Kitaro . . . $14.98 link . . .
Subject:	"listen to audio"
Wizard:	[follows link]
Screen Reader:	Page loading
Wizard:	[plays the next sample track]
Screen reader:	[audio sample]
Subject:	"next"
Wizard:	[plays the next sample track]
Screen reader:	[audio sample]
Subject:	"buy this cd"
Wizard:	[clicks 'Add to cart' button, then clicks 'Proceed to Checkout' button]
Screen reader:	[reads out all captions]

Table 11: An example dialogue from corpus along with associated wizard actions.