Combining Verbal and Nonverbal Features to Overcome the 'Information Gap' in Task-Oriented Dialogue

Eun Young Ha, Joseph F. Grafsgaard, Christopher M. Mitchell,
Kristy Elizabeth Boyer, and James C. Lester
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
{eha, jfgrafsg, cmmitch2, keboyer, lester}@ncsu.edu

Abstract

Dialogue act modeling in task-oriented dialogue poses significant challenges. It is particularly challenging for corpora consisting of two interleaved communication streams: a dialogue stream and a task stream. In such corpora, information can be conveyed implicitly by the task stream, yielding a dialogue stream with seemingly missing information. A promising approach leverages rich resources from both the dialog and the task streams, combining verbal and non-verbal features. This paper presents work on dialogue act modeling that leverages body posture, which may be indicative of particular dialogue acts. Combining three information sources (dialogue exchanges, task context, and users' posture), three types of machine learning frameworks were compared. The results indicate that some models better preserve the structure of task-oriented dialogue than others, and that automatically recognized postural features may help to disambiguate user dialogue moves.

1 Introduction

Dialogue act classification is concerned with understanding users' communicative intentions as reflected in their utterances. It is an important first step toward building automated dialogue systems. To date, the majority of work on dialogue act modeling has addressed spoken dialogue (Samuel et al., 1998; Stolcke et al., 2000; Surendran and Levow, 2006; Bangalore et al., 2008; Sridhar et al., 2009; Di Eugenio et al., 2010). However, with the increasing popularity of computer-mediated means of conversation, such as instant messaging and social networking services, automated analysis of textual dialogue holds much appeal. Dialogue act modeling for textual conversations has many practical application areas, which include web-based intelligent tutoring systems (Boyer et al., 2010a), chat-based online customer service (Kim et al., 2010), and social media analysis (Joty et al., 2011).

Human interaction involves not only verbal communication but also nonverbal communication. Research on nonverbal communication (Knapp and Hall, 2006; Mehrabian, 2007; Russell et al., 2003) has identified a range of nonverbal cues, such as posture, gestures, eye gaze, and facial and vocal expressions. However, the utility of these nonverbal cues has not been fully explored within the context of dialogue act classification research. Previous research has leveraged prosodic cues (Sridhar et al., 2009; Stolcke et al., 2000) and facial expressions (Boyer et al., 2011) for automatic dialogue act classification, but other types of nonverbal cues remain unexplored. As a first step toward a dialogue system that learns its behavior from a human corpus, this paper proposes a novel approach to dialogue act classification that leverages information about users' posture. Posture has been found to be a significant indicator of a broad range of emotions (D'Mello and Graesser, 2010; Kapoor et al., 2007; Woolf et al., 2009). Based on the premise that emotion plays an

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 247–256,
Seoul, South Korea, 5-6 July 2012.

important role in dialogue, this work hypothesizes that adding posture features will improve the performance of automatic dialogue act models.

The domain considered in this paper is task-oriented textual dialogue collected in a human tutoring study. In contrast to conventional task-oriented dialogue corpora (e.g., Carletta et al., 1997; Jurafsky et al., 1998; Ivanovic, 2008) in which conversational exchanges are carried out within a single channel of dialogue between the dialogue participants, the corpus used in this work utilizes two separate and interleaved streams of communication. One stream is the textual conversation between a student and a tutor (*dialogue stream*). The other is the student's problem-solving activity (*task stream*). As will be described in Section 3, the interface used in the corpus collection was designed to allow the tutor to monitor the student's problem-solving activities. Thus, the student's problem-solving activities and the tutor's monitoring of those activities functioned as an implicit communication channel. This characteristic of the corpus poses significant challenges for dialogue act modeling. First, because the dialogue stream and the task stream are interleaved, the dialogue stream alone may not be coherent. Second, since information can be exchanged implicitly via the task stream, the dialogue likely contains substantial *information gaps*[1].

Addressing these challenges, the dialogue act models described in this paper combine three sources of information: the verbal information from the dialogue stream, the task-related context from the task stream, and information about users' posture. This paper makes several contributions to the dialogue research community. First, it is the first effort to explore posture as a nonverbal cue for dialogue act classification. Second, the proposed approach is fully automatic and ready for real-world application. Third, this paper explicitly defines the notion of *information gap* in task-oriented dialogue consisting of multiple communication channels, which has only begun to be explored in the context of dialogue act classification (Boyer et al., 2010a). Finally, this paper examines adaptability of previous dialogue act classification approaches in conventional task-oriented domains by comparing three classifiers previously applied to dialogue act modeling for task-oriented dialogue.

[1] In this paper, *information gap* is defined as the information that is missing from the explicit verbal exchanges between the dialogue participants but conveyed by the implicit task stream.

2 Related Work

A rich body of research has addressed data-driven approaches for dialogue act modeling. Russell et al. (2003) applied a transformation-based learning approach for dialogue act tagging for spoken dialogue, using speaker direction, punctuation, marks, and cue phrases. Stolcke et al. (2000) modeled the structure of dialogue as an HMM, treating the dialogue acts as the observations emitted from the hidden states of the learned HMM. More recently, Bangalore et al. (2008) proposed a unified approach to task-oriented dialogue, in which both the user dialogue act classification and the system dialogue act selection were informed by a shared maximum entropy dialogue act classifier. Sridhar et al. (2009) also used a maximum entropy model, exploring the utility of different representations of prosodic features. Di Eugenio et al. (2010) used a memory-based classifier, in combination with a modified latent semantic analysis (LSA) technique by augmenting the original word-document matrix in LSA with rich linguistic features.

While most work on dialogue act modeling has focused on spoken dialogue, a recent line of investigation has explored the analysis of textual conversation, such as asynchronous online chat conversation (Wu et al., 2005; Forsyth, 2007; Reitter et al., 2010; Joty et al., 2011) and synchronous online chat conversation (Ivanovic, 2008; Kim et al., 2010; Boyer et al., 2010a). Wu et al. (2005) proposed a transformation-based learning approach for an asynchronous chat posting domain, utilizing regular expression-based selection rules. For a similar domain, Forsyth (2007) applied neural networks and Naïve Bayes classification technique using lexical cues. Ritter et al. (2010) and Joty et al. (2011) applied unsupervised learning approaches to dialogue act modeling for Twitter conversations, in which dialogue acts were automatically discovered by clustering raw utterances. Work by Ivanovic (2008) and Kim et al. (2010) analyzed one-to-one synchronous online chat dialogue in a task-oriented

customer service domain. Ivanovic (2008) applied maximum entropy, naïve Bayes, and support vector machines using word *n*-gram features. Kim et al. (2010) compared the CRF, HMM-SVM, and Naïve Bayes classifiers using word *n*-grams and features extracted from the dialogue structure, in which CRF achieved the highest performance. Boyer et al. (2010a) investigated dialogue act modeling for task-oriented tutorial dialogue, applying a logistic regression approach using lexical, syntactic, dialogue structure, and task structure features.

Some previous dialogue act modeling work (Boyer et al., 2011; Sridhar et al., 2009; Stolcke et al., 2000) leveraged nonverbal information such as prosodic cues (Sridhar et al., 2009; Stolcke et al., 2000) and facial expressions (Boyer et al., 2011). Stolcke et al. (2000) combined various prosodic features such as pitch, duration, and energy. Sridhar et al. (2009) represented the sequence of prosodic features as *n*-grams. Boyer et al. (2011) leveraged confusion-related facial expressions for tutorial dialogue.

Like Boyer et al. (2010a), this work addresses dialogue act classification for task-oriented textual conversation in a web-based tutoring domain. In contrast to Boyer et al. (2010a), whose approach directly leveraged manually annotated features, making it challenging to apply the proposed model to a real-world system, the present work is fully automatic and ready for real-world application. A novel feature of this work is its utilization of nonverbal cues carried by users' posture. This is the first dialogue act classification work that leverages posture information.

3 Data

The corpus used in this paper consists of textual exchanges between a student and a tutor in a web-based remote-tutoring interface for introductory programming in Java. The corpus was collected from a series of six tutoring lessons, covering progressive topics in computer science over the course of four weeks. The tutoring interface consisted of four windows: a *task* window displaying the current programming task; a *code* window in which the student writes Java code; an *output* window for displaying the result of compiling and running the code; and a *chat* window for instant exchange of textual dialogue between the student and tutor. With this tutoring interface, the student and the tutor were able to exchange textual dialogue and share a synchronized view of the task. Apart from sending dialogue messages, the only action the tutor could perform to affect the student's interface was advancing to the next programming task.

3.1 Data Collection

The data collection conducted in Fall 2011 paired 42 students with one of four tutors for six forty-minute tutoring sessions on introductory computer science topics. The students were chosen from a first-year engineering course and were pre-screened to filter out those with significant programming experience. The tutors were graduate students with previous tutoring or teaching experience in Java programming. Students were compensated for their participation with partial course credit. The students worked with the same tutor for the entire study. Each lesson consisted of between four and thirteen distinct subtasks.

During each tutoring session, the dialogue text exchanged between the student and the tutor was logged to a database. Additional runtime data including content of the student's Java code, the result (e.g., success or failure) of compiling and running the student's code, and the IDs of the subtask were logged. All logged data were time-stamped at a millisecond precision. Students' body posture was recorded at a rate of 8 frames per second with a Kinect depth camera, which emits infrared rays to measure distance for each pixel in a depth image frame. The camera was positioned above the student's computer monitor, ensuring the student's upper body is centered in the recorded image. Tutors were not recorded.

3.2 Dialogue Act Annotation

For the work described in this paper, a subset of the collected data was manually annotated, which include the first of the six tutoring lessons from 21 students. This corpus contains 2564 utterances (1777 tutor, 787 student). The average number of utterances per tutoring session was 122 (min = 74; max = 201). The average number of tutor utterances per session was 84.6 (min = 51; max = 137) and the average number of student utterances per session was 37.4 (min = 22; max = 64).

Extending a previous annotation scheme used for similar task-oriented tutorial dialogue (Boyer et al., 2010b), the scheme used in this work consists of 13 dialogue act tags (Appendix). The dialogue turns that contained more than one dialogue function were segmented into multiple utterances before being assigned a dialogue act tag. The annotation scheme did not constrain any of the dialogue act tags as applying either to students' or tutors' utterances only; however, the resulting distribution of the tags in the annotated corpus show certain dialogue act tags were more relevant to either students' or tutors' utterances. Figure 1 depicts an excerpt from the corpus with the manually applied dialogue act annotations.

Tutor: hang on :) [S]
Tutor: When we show you example code, it is not the code you need to write. [S]
Tutor: Look at the task again. [H]

Student writes programming code

Tutor: YUP [PF]
Tutor: Perfect [PF]
Tutor: OK. Go ahead and test. [DIR]
Student: And I don't need anything in the parentheses? [Q]
Tutor: Line 9 is correct. You do NOT need anything inside the parentheses. [A]
Student: Ok [ACK]

Student compiles and runs code successfully

Tutor: Good. [PF]
Tutor: Moving on. [S]

Tutor advances to the next task.

Student writes programming code

Tutor: Syntactically correct. But there is a logic error [LF]
Tutor: When will the output statement display your request to the player? [Q]
Student: AFTER they put in their name [A]
Tutor: Exactly [PF]

Figure 1. Corpus Excerpt with Dialogue Act Annotation

Three human annotators were trained to apply the scheme. The training consisted of an iterative process involving collaborative and independent tagging, followed by refinements of the tagging protocol. At the initial phase of training, the annotators tagged the corpus collaboratively. In later phases annotators tagged independently. To compute agreement between different annotators, 24% (5 of the 21 sessions) of the corpus were doubly annotated by two annotators. All possible pairs of the annotators participated in double annotation. The aggregate agreement was .80 in Cohen's Kappa (Cohen, 1960).

3.3 Posture Estimation

Posture has been found to be a significant indicator of a broad range of emotions such as anxiety, boredom, confusion, engaged concentration (or flow), frustration, and joy (D'Mello and Graesser, 2010; Kapoor et al., 2007; Woolf et al., 2009). Early investigations into posture utilized pressure-sensitive chairs which provided indirect measures of upper-body posture (D'Mello and Graesser, 2010; Kapoor et al., 2007; Woolf et al., 2009). Newer, computer vision-based techniques provide more detailed postural data (Sanghvi et al., 2011). The present work uses a posture estimation algorithm developed to automatically detect the head, mid torso, and lower torso through depth image recordings of seated individuals (Grafsgaard et al., 2012). With this estimation algorithm, posture is represented as a triple of *head depth* (distance between camera and head)*, mid torso depth,* and *lower torso depth.*

A dataset of depth camera recordings from the first of the six tutoring lessons consists of 512,977 depth image frames collected across 18.5 hours of computer-mediated human-human tutoring among 33 participants.[2] For each depth image frame, the posture algorithm scanned through the three middle regions that corresponded to head, mid-torso, and lower-torso of the recorded person, and selected a single representative depth pixel from each region. The boundaries for each region were heuristically determined relying on the placement of the students' chairs in the middle of the depth recording view at a common distance. Given these constraints, the model was manually verified by two independent human judges to have 95.1% accuracy across 1,109 depth image snapshots corresponding to one-minute intervals across the dataset. The algorithm output for each depth image was labeled as erroneous if either judge found that any of the posture tracking points did not coincide with its target region. Example output of the algorithm is shown in Figure 2.

[2] The other 9 sessions were not successfully recorded because of technical errors.

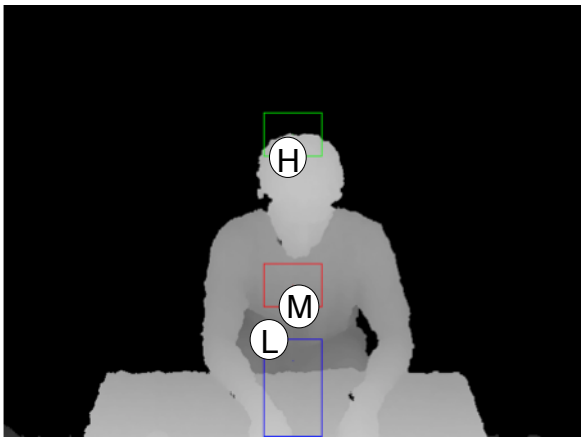

Figure 2. Automatically detected posture points (H = *headDepth*, M = *midTorsoDepth*, L = *lowerTorsoDepth*)

4 Features

For web-based one-to-one dialogue systems, it is important to achieve efficient runtime performance. To maximize real-world feasibility of the learned dialogue act classifiers, this work only considers the features that can be automatically extracted at runtime. In addition, the use of linguistic analysis software, such as a part-of-speech tagger and a syntactic parser, is intentionally restrained. One might argue that rich linguistic analysis may provide additional information to dialogue act classifiers, potentially improving the performance of learned models. However, there is a trade-off between additional information obtained by rich linguistic analysis and processing time. In addition, previous work (Boyer et al., 2010a) found part-of-speech and syntax features did not provide obvious benefit for dialogue act classification in a domain similar to the one considered in this work. The dialogue act classifiers described in this paper integrate four classes of features automatically extracted from three sources of information: the textual dialogue utterances, task-related runtime information logged into the database, and the images of the students recorded by depth cameras. Each feature class is explained in the following subsections.

4.1 Lexical Features

Based on previous dialogue act classification research (Bangalore et al., 2008; Boyer et al., 2010a; Kim et al., 2010), this work utilizes word *n*-grams as features for dialogue act classification. In the experiment reported in Section 5, unigrams and bigrams were used. Adding higher order *n*-grams did not improve model accuracies. In our corpus (Section 3), the nature of the student dialogues is informal and utterances contain many typos. To remove undesirable noise in the data such as typos and rare words, *n*-grams were filtered out according to their frequency in the training data (i.e., *n*-grams that appear less than a predefined cutoff threshold in the training data are not included as features). The value of the cutoff threshold was empirically determined by testing the values between 0 and 10 on a development data set that consisted of 20% of randomly selected dialogue sessions. The value of 3 was selected as it yielded the highest classification accuracy.

4.2 Dialogue Context Features

While lexical features characterize the intrinsic nature of individual utterances, the context of the utterance within a larger dialogue structure provides additional information about a given utterance in relation with other utterances. This work considers the following dialogue context features:

- ***Utterance Position:*** Specifies the relative position of an utterance at a given turn. The value of this feature indicates whether the utterance is the first one in a given turn, the second or later one in a given turn, or the given turn consists of a single utterance.
- ***Length:*** Specifies the number of a given utterance in terms of individual word tokens.
- ***Previous Author:*** Indicates whether the author of the previous utterance was *student* or *tutor*.
- ***Previous Tutor Dialogue Act:*** Specifies dialogue act of the most recent tutor utterance. The value of this feature is directly extracted from the manual annotation in the corpus, because in the broader context of our work, tutor dialogue moves will be determined by an external dialogue management module.

4.3 Task Context Features

In our data, students' problem-solving activities (e.g., reading the problem description, writing computer programming code, and compiling and running the code) functioned as an implicit communication channel between students and tutors (Section 1). Because of the existence of this

implicit communication channel, the dialogue exchanges between students and tutors likely contain substantial information gaps. To overcome such information gaps, it is important to identify effective task context features. The present work leverages the following task context features, which can be automatically extracted during runtime:

- ***Previous Task Action:*** Specifies the type of the most recent problem-solving action performed by the student. The value could be *message* (writing a textual message to the tutor) *code* (writing code in the code window), or *compile_run* (compiling or running the code).
- ***Task Begin Flag:*** A binary feature that indicates whether a given utterance is the first one since the current problem task was posted.
- ***Task Activity Flag:*** Another binary feature indicating that a given utterance was preceded by a student's task activity.
- ***Last Compile/Run Status:*** Specifies the status (e.g., *begin*, *stop*, *success*, *error*, *input sent*) of the most recent compile/run action performed by the students.

In addition to the listed task context features, the utility of time information was also explored, such as the amount of time taken for previous coding activity and the elapsed time since the beginning of the current task. However, these features did not positively impact the performance of the learned models and were thus excluded.

4.4 Posture Features

After preprocessing recorded image frames with the estimation algorithm (Section 3.3), students' postures were represented as tuples of three different integer values, each respectively representing *head depth, mid torso depth, and lower torso depth*. To extract posture features, the time window of n seconds directly preceding a given utterance was compared with the previous time window of the same size in terms of *min*, *max*, *median*, *average*, and *variance* of each depth value. The indicators of whether each of these values has *increased*, *decreased* or *remained the same* were considered as potential posture features. To avoid introducing errors to the model by insignificant changes in posture, an error tolerance τ was allowed (i.e., the two compared postures were considered the same unless the amount of the change in the posture was greater than τ).

Optimal values for n and τ were empirically determined, selecting the values that maximized classification accuracy on the development data set. For n, the values between 0 and 60 were compared at an interval of 10. The value of 50 was selected for head depth and 60 for both mid torso depth and lower torso depth. Similarly, the value of τ was determined by comparing the values between 0 and 200 with an increment of 10. The selected value was 100.

All the potential posture features were examined in an informal experiment, in which each of the potential posture features were added to the combination of the lexical, the dialogue context, and the task context features. The posture features that improved the classification accuracy after adding them were included in the present dialogue act models. The selected posture features are *min of head depth* and *max, median, and average of lower torso depth*. None of the *mid torso depth* features were selected.

5 Experiment

The goal of this experiment is twofold: (1) to evaluate the effectiveness of the feature classes and (2) to compare the performance of three classifiers: maximum entropy (ME), naïve Bayes (NB), and conditional random field (CRF). These classifiers are chosen because they have been shown effective for dialogue act modeling in traditional task-oriented textual dialogue, in which conversational exchanges were carried out by a single channel of dialogue (Ivanovic, 2008; Kim et al., 2010). Previous result by Kim et al. (2010) suggests a structured model such as CRF yields more accurate dialogue act model compared to unstructured models (e.g., Naïve Bayes), because of its ability to model the sequential patterns in target classification labels. This experiment examines whether a similar finding is observed for our domain, which exhibits substantial information gaps due to the existence of an implicit communication channel, the task stream.

5.1 Dialogue Act Modeling

All classifiers were built using the MALLET package (McCallum, 2002). This experiment used the manually annotated portion of the data

described in Section 3. The original dialogue scheme (Section 3.2) was slightly modified by introducing an additional dialogue act, *GR*, in order to distinguish conventional expressions, such as *greetings* and *thanks*, from other information-delivering utterances. For this modified scheme, annotator agreement was 0.81 in Cohen's Kappa on the doubly annotated portion of the corpus. 6 among the 21 dialogue sessions in the annotated data do not have accompanying images due to technical problems with the depth camera, thus these sessions were excluded from this experiment. Table 1 shows the distribution of the student dialogue act tags in the resulting corpus of 15 dialogues used in this experiment. The most frequent tag was *A* (*answer*), followed by *ACK* (*acknowledgement*) and *Q* (*question*). The features were extracted by aligning three sources of information (the textual dialogue corpus, the task-related runtime log data, and the recorded images) by timestamp. Word boundaries in the dialogue corpus were recognized by the surrounding white spaces and punctuations.

The dialogue context features (D) leveraged in this paper includes *previous tutor dialogue act*. This feature takes the manually annotated value in the corpus, because this work assumes the existence of an external dialogue manager. However, since the external dialogue manager is not likely to achieve 100% accuracy in predicting human tutor dialogue acts, it would be informative to estimate a reasonable range of the accuracies of the student dialogue act model, taking into account the errors introduced by the dialogue manager. For this reason, two versions of the dialogue context features were considered in this experiment: one that leverages the full set of dialogue context features (D) and the other that excludes previous tutor dialogue act (D-). These respectively provide the maximum and the minimum expected accuracy of the student dialogue act model, when used with a dialogue manager.

Student Dialogue Act	Distribution
A (answer)	192 (34.7%)
ACK (acknowledgement)	124 (22.4%)
Q (question)	92 (16.6%)
S (statement)	76 (13.7%)
GR (greeting and thanks)	52 (9.4%)
C (clarification)	6 (1.0%)
RF (request for feedback)	5 (.9%)
RC (request confirmation)	2 (.4%)
O (other)	5 (.9%)
Total	**554**

Table 1. Student dialogue acts in the experiment data

The models were trained and tested using five-fold cross validation, in which the 15 dialogue sessions were partitioned into 5 non-overlapping sets of the same size (i.e., 3 sessions per partition). Each set was used for testing exactly once.

5.2 Results

Table 2 reports the average classification accuracies from the five-fold cross validation. The majority baseline accuracy for our data is .347, when the classifier always chooses the most frequent dialog act (*A*). The first group of rows in Table 3 report the accuracies of individual feature classes. All of the individual features performed better than the baseline. The improvement from the baseline was significant except for D- with CRF. The most powerful feature class was dialogue context class when the full set was used. The second group in Table 3 shows the effects of incrementally combining the feature classes. Adding dialogue act features to the lexical features (L + D) brought significant improvement in the classification accuracy for ME and CRF. Adding posture features (L + D + T + P) also improved the accuracy of ME by a statistically significant margin. The last group shows similar results for ME when the previous tutor dialogue act was excluded from the dialogue context, except that the improvement achieved by adding the posture features (L + D- + T + P) was not significant.

Features		ME	NB	CRF
Individual	Lexical (L)	.696**	.703**	.599**
	Dialogue (D)	**.711****	**.715****	**.696****
	Dialogue- (D-)	.477**	.473**	.405
	Task (T)	.405**	.396*	.386*
	Posture (P)	.382*	.385*	.399*
Max	L + D	.772§§	.724	.692§§
	L + D + T	.777	.729	.694
	L + D + T + P	**.789**‡	.714	.682
Min	L + D-	.724§§	.681	.606
	L + D- + T	.733	.671	.627
	L + D- + T + P	**.750**	.676	.644

Table 2. Classification accuracies ($^{*}p < .05$, $^{**}p < .01$ compared to baseline; $^{\S\S}p < .01$ compared to L; and $^{\ddagger}p < .05$ compared to L + D + T, with paired-samples *t*-test)

The highest accuracy was achieved by ME when using all four classes of the features, with maximum (L + D + T + P) .789 and minimum (L + D- + T + P) .750. For both the maximum and the minimum conditions, the differences among the classifiers were significant ($p < .01$, one-way repeated measure ANOVA), with post-hoc Tukey HSD tests revealing ME was significantly better than both NB ($p < .05$) and CRF ($p < .01$). There was no significant difference between NB and CRF.

6 Discussion

The experiment described in Section 5 compared the utility of lexical, dialogue context, task context, and posture features for dialogue act classification. The results indicate the effectiveness of these features. Particularly, adding the dialogue context and the posture features improved the accuracy of the maximum entropy model. Although the margin of improvement achieved by adding posture features was relatively small, the improvement was statistically significant ($p < .05$) for the maximum condition (L + D + T + P), which suggests that the users' posture during computer-mediated textual dialogue conveys important communicative messages.

The experiment also compared three classifiers: maximum entropy, naïve Bayes, and CRF. Interestingly, CRF was the worst-performing model for our data, contradicting the previous finding by Kim et al. (2010), in which CRF (a structured classifier) performed significantly better than Naïve Bayes (a non-structured classifier). This contradictive result suggests that, in our domain, the presence of an implicit communication channel resulted in substantial information gaps in the dialogue and it poses new challenges that were not encountered by conventional task-oriented domains consisting of a single communication channel.

The maximum entropy classifier achieved the best overall performance, reaching accuracy of .789. This is an encouraging result compared to previous work in a similar domain. Boyer et al. (2010a) reported an accuracy of .628 for dialogue act classification in a similar domain. However, a direct comparison is not applicable since different data were used in their work.

7 Conclusions and Future Work

Dialogue act modeling for a task-oriented domain in which the dialogue stream is interleaved with the task stream poses significant challenges. With the goal of effective dialogue act modeling, this work leverages information about users' posture as non-verbal features. An experiment found that posture is a significant indicator of dialogue acts, in addition to lexical features, dialogue context, and task context. The experiment also compared three statistical classifiers: maximum entropy, naive Bayes, and CRF. The best performing model was maximum entropy. Using all features, the maximum entropy achieved .789 in accuracy.

Several directions for future work are promising. First, given the encouraging finding that nonverbal information plays a significant role as a communicative means for task-oriented dialogue, various types of non-verbal information can be investigated, such as gesture and facial expressions. Second, incorporating richer task features, such as in our case, deep analysis of student code, may contribute to more accurate dialogue act modeling. Third, it is important to generalize the findings to a larger data set, including across other task-oriented domains. Finally, the community is embracing a move toward annotation-lean approaches such as unsupervised or semi-supervised learning, which hold great promise for dialogue modeling.

Acknowledgments

This research was supported by the National Science Foundation under Grant DRL-1007962. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bangalore, S., Di Fabbrizio, G., & Stent, A. (2008). Learning the structure of task-driven human-human dialogs. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(7), 1249-1259.

Boyer, K. E., Grafsgaard, J. F., Ha, E. Y., Phillips, R., & Lester, J. C. (2011). An affect-enriched dialogue act classification model for task-oriented dialogue. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human*

Language Technologies (pp. 1190-1199). Portland, OR.

Boyer, K. E., Ha, E. Y., Phillips, R., Wallis, M. D., Vouk, M. A., & Lester, J. C. (2010a). Dialogue Act Modeling in a Complex Task-Oriented Domain. *Proceedings of the 11th Annual SIGDIAL Meeting on Discourse and Dialogue* (pp. 297-305). Tokyo, Japan.

Boyer, K. E., Phillips, R., Ingram, A., Ha, E. Y., Wallis, M., Vouk, M., & Lester, J. (2010b). Characterizing the effectiveness of tutorial dialogue with hidden markov models. *Proceedings of the 10th international conference on Intelligent Tutoring Systems* (pp. 55-64). Pittsburgh, PA.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, *23*, 13–31.

Cavicchio, F. (2009). The modulation of cooperation and emotion in dialogue: The REC corpus. *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop* (pp. 81 - 87). Suntec, Singapore.

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

Di Eugenio, B., Xie, Z., & Serafin, R. (2010). Dialogue act classification, instance-based learning, and higher order dialogue structure. *Dialogue and Discourse*, *1*(2), 81 - 104.

D'Mello, S., & Graesser, A. (2010). Mining Bodily Patterns of Affective Experience during Learning. *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 31-40). Pittsburgh, PA.

Forsyth, E. N. (2007). *Improving Automated Lexical and Discourse Analysis of Online Chat Dialog*. Master's thesis. Naval Postgraduate School.

Grafsgaard, J. F., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2012). Analyzing Posture and Affect in Task-Oriented Tutoring. *Proceedings of the 25th Florida Artificial Intelligence Research Society Conference* (pp. 438-443). Marco Island, FL.

Ivanovic, E. (2008). *Automatic instant messaging dialogue using statistical models and dialogue acts*. Master's thesis. The University of Melbourne.

Joty, S. R., Carenini, G., & Lin, C.-Y. (2011). Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (pp. 1807-1813). Barcelona, Catalonia, Spain.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., et al. (1998). *Switchboard discourse language modeling project report*. Baltimore, MD.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies*, *65*(8), 724-736.

Kim, S. N., Cavedon, L., & Baldwin, T. (2010). Classifying dialogue acts in one-on-one live chats. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 862-871). Cambridge, MA.

Knapp, M. L., & Hall, J. A. (2006). *Nonverbal Communication in Human Interaction* (6th ed.). Belmont, CA: Wadsworth/Thomson Learning.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. Available from http://mallet.cs.umass.edu

Mehrabian, A. (2007). *Nonverbal Communication*. New Brunswick, NJ: Aldine Transaction.

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised modeling of twitter conversations. *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter* (pp. 172 - 180). Los Angeles, CA.

Russell, J. A., Bachorowski, J. A., & Fernandez-dols, J. M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology*, *54*, 329-349.

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. *Proceedings of the 6th international conference on Human-robot interaction* (pp. 305-312). Lausanne, Switzerland.

Sridhar, R., Bangalore, S., & Narayanan, S. (2009). Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech and Language*, *23*(4), 407 - 422.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., et al. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339-373.

Surendran, D., & Levow, G.-A. (2006). Dialog act tagging with support vector machines and hidden

Markov models. *Proceedings of Interspeech* (pp. 1950 - 1953). Pittsburgh, PA.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. W. (2009). Affect-aware tutors recognising and responding to student affect. *International Journal of Learning Technology*, *4*(3/4), 129-164.

Wu, T., Khan, F. M., Fisher, T. A., Shuler, L. A., & Pottenger, W. M. (2005). Posting Act Tagging Using Transformation-Based Learning. In T. Y. Lin, S. Ohsuga, C.-J. Liau, X. Hu, & S. Tsumoto (Eds.), *Foundations of Data Mining and knowledge Discovery* (pp. 319 - 331). Springer.

Appendix. Dialogue Act Annotation Scheme and Inter-rater Agreement

Tag	Description	Frequency	Agreement (*k*)
H	**Hint:** The tutor gives advice to help the student proceed with the task	Tutor: 133 Student: 0	.50
DIR	**Directive:** The tutor explicitly tells the student the next step to take	Tutor: 121 Student: 0	.63
ACK	**Acknowledgement:** Either the tutor or the student acknowledges previous utterance; conversational grounding	Tutor: 41 Student: 175	.73
RC	**Request for Confirmation:** Either the tutor or the student requests confirmation from the other participant (e.g., *"Make sense?"*)	Tutor: 11 Student: 2	Insufficient data
RF	**Request for Feedback:** The student requests an assessment of performance or work from the tutor	Tutor: 0 Student: 5	1.0
PF	**Positive Feedback:** The tutor provides a positive assessment of the student's performance	Tutor: 327 Student: 0	.90
LF	**Lukewarm Feedback:** The tutor provides an assessment that has both positive and negative elements	Tutor: 13 Student: 0	.80
NF	**Negative Feedback:** The tutor provides a negative assessment of the student's performance	Tutor: 1 Student: 0	.40
Q	**Question:** A question regarding the task that is not a direct request for confirmation or feedback	Tutor: 327 Student: 120	.95
A	**Answer:** An answer to an utterance marked Q	Tutor: 96 Student: 295	.94
C	**Correction:** Correction of a typo in a previous utterance	Tutor: 10 Student: 6	.54
S	**Statement:** A statement regarding the task that does not fit into any of the above categories	Tutor: 681 Student: 174	.71
O	**Other:** Other utterances, usually containing only affective content	Tutor: 6 Student: 10	.69