Multimodality and Dialogue Act Classification in the RoboHelper Project

Lin Chen
Department of Computer Science
University of Illinois at Chicago
851 S Morgan ST, Chicago, IL 60607
lchen43@uic.edu

Barbara Di Eugenio
Department of Computer Science
University of Illinois at Chicago
851 S Morgan ST, Chicago, IL 60607
bdieugen@uic.edu

Abstract

We describe the annotation of a multimodal corpus that includes pointing gestures and haptic actions (force exchanges). Haptic actions are rarely analyzed as full-fledged components of dialogue, but our data shows haptic actions are used to advance the state of the interaction. We report our experiments on recognizing Dialogue Acts in both offline and online modes. Our results show that multimodal features and the dialogue game aid in DA classification.

1 Introduction

When people collaborate on physical or virtual tasks that involve manipulation of objects, dialogues become rich in gestures of different kinds; the actions themselves that collaborators engage in also perform a communicative function. Collaborators gesture while speaking, e.g. saying "Try there?" while pointing to a faraway location; they perform actions to reply to their partner's utterances, e.g. opening a cabinet to comply with "please check cabinet number two". Conversely, they use utterances to reply to their partner's gestures and actions, e.g. saying "not there, try the other one" after their partner opens a cabinet. Gestures and actions are an important part of such dialogues; while the role of pointing gestures has been explored, the role that haptic actions (force exchanges) play in an interaction has not.

In this paper, we present our corpus of multimodal dialogues in a home care setting: a helper is helping an elderly person perform activities of daily living (ADLs) such as preparing dinner. We investigate how to apply Dialogue Act (DA) classification to these multimodal dialogues. Many challenges arise. First, an utterance may not directly follow a spoken utterance, but a gesture or a haptic action. Likewise, the next move is not necessarily an utterance, it can be a gesture (pointing or haptics) only, or a multimodal utterance. Third, when people use gestures and actions together with utterances, the utterances become shorter, hence the textual context that has been used to advantage in many previous models is impoverished. Our contributions concern: exploring the dialogue functions of what we call *Haptic-Ostensive (H-O)* actions (Foster et al., 2008), namely haptics actions that often perform a referential function; experimenting with both offline and online DA classification, whereas most previous work only focuses on offline classification (Stolcke et al., 2000; Hastie et al., 2002; Di Eugenio et al., 2010a); highlighting the role played by multimodal features and dialogue structure (in the form of dialogue games) as concerns DA classification.

Our work is part of the RoboHelper project (Di Eugenio et al., 2010b) whose ultimate goal is to deploy robotic assistants for the elderly so that they can safely remain living in their home. The models we derive from our experiments are the building blocks of a multimodal information-state based dialogue manager, whose architecture is shown in Figure 1. The dialogue manager performs reference resolution, specifically resolving third person pronouns and deictics in utterances; classifies utterances to DAs; infers the dialogue games for utterances; updates the dialogue state, and finally decides what the next step is in the interaction. We have discussed our approach to multimodal reference resolution in (Chen et al., 2011; Chen and Di Eugenio, 2012). In this paper, we focus on the Dialogue Act classification component. We will also touch on Dialogue Game inference. Our collaborators are developing the speech processing, vision and haptic recognition components (Franzini and Ben-Arie, 2012; Ma and Ben-Arie, 2012; Javaid and Žefran, 2012), that, when integrated with the dialogue manager we are building,

Proceedings of the SIGDIAL 2013 Conference, pages 183–192,
Metz, France, 22-24 August 2013.

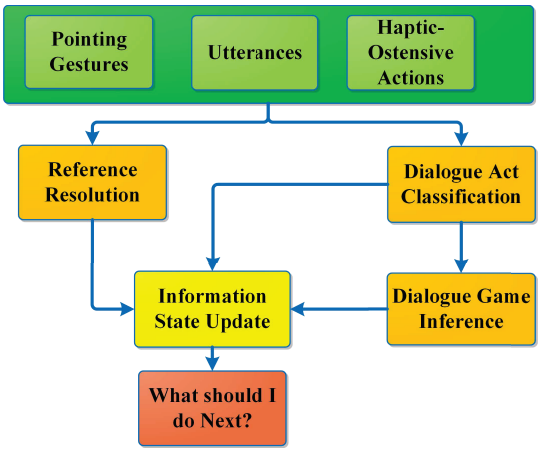

Figure 1: System Architecture

will make the interface situated in and able to deal with a real environment.

After discussing related work in Section 2, we present our multimodal corpus and the multidimensional annotation scheme we devised in Section 3. In Section 4 we discuss all the features we used to build machine learning models to classify DAs. Sections 5 is devoted to our experiments and the results we obtained. We conclude and discuss future work in Section 6.

2 Related Work

Due to its importance in dialogue research, DA classification has been the focus of a large body of research (Stolcke et al., 2000; Sridhar et al., 2009; Di Eugenio et al., 2010a; Boyer et al., 2011). Some of this work has been made possible by several available corpora tagged with DAs, including HCRC Map Task (Anderson et al., 1991), CallHome (Levin et al., 1998), Switchboard (Graff et al., 1998), ICSI Meeting Recorder (MRDA) (Shriberg et al., 2004), and the AMI multimodal corpus (Carletta, 2007).

Researchers have applied various approaches to this task. Initially only simple textual features were used, e.g. n-grams were used to model the constraints for DA sequences in an HMM model (Stolcke et al., 2000). Zimmermann et al. (2006) investigated the joint segmentation and classification of DAs using prosodic features. Sridhar et al. (2009) showed that prosodic cues can improve DA classification for a Maximum Entropy based model. Di Eugenio et al. (2010a) extended Latent Semantic Analysis with linguistic features, including dialogue game information. Boyer et al. (2011) integrates facial expressions to significantly improve the recognition of several DAs, whereas Ha et al. (2012) shows that automatically recognized postural features may help to disambiguate DAs.

It should be pointed out that most of this work focuses on offline DA classification – namely, DA classification is performed on the corpus using the gold-standard classification for the previous DA(s). Since some sort of history of previous DAs is used by all systems, using online classification for the previous DAs will unavoidably impact performance (Sridhar et al., 2009; Kim et al., 2012). Additionally, for models such as HMMs and CRF that approach the problem as sequence labeling, online processing means that only a partial sequence is available.

3 The ELDERLY-AT-HOME Corpus

This work is based on the ELDERLY-AT-HOME corpus, a multimodal corpus in the domain of elderly care (Chen and Di Eugenio, 2012). The corpus contains 20 human-human dialogues. In each dialogue, a helper (HEL) and an elderly person (ELD) perform *Activities of Daily Living* (ADL) (Krapp, 2002), such as getting up from chairs, finding pots, cooking pasta. The setting is a fully equipped studio apartment used for teaching and research in a partner university (see Figure 2). The corpus contains 482 minutes of recorded videos, which comprise 301 minutes of what we call *effective video*, obtained by eliminating irrelevant content such as explanations of the tasks and interruptions by the person who accompanied the elderly subject (who is not playing the part of the helper). This 301 minutes contain 4782 spoken turns. The corpus includes video and audio data in .avi and .wav format, haptics data collected via instrumented gloves in .csv format, and the transcribed utterances in xml format.

The *Find* subcorpus of our corpus comprises only *Find* tasks, where subjects look for and retrieve various kitchen objects such as pots, silverware, pasta, etc. from various locations in the apartment. We define a *Find* task as a continuous time span during which the two subjects are collaborating on finding objects. *Find* tasks naturally arise while performing an ADL such as preparing dinner. Figure 3 shows a *Find* task example.

Figure 2: Data Collection Experiment

1	ELD	And there is a spoon down there, in the second drawer? *[Point(ELD,Drawer1)]*
2	HEL	Down there?*[Point(HEL,Drawer1)]*
3	ELD	Yes.
4	HEL	This? *[Touch(HEL,Drawer1)]*
5	ELD	Uh-huh.
6	HEL	*[Open(HEL,Drawer1)]*
7	ELD	A spoon.
8	HEL	Is this the spoon? *[Takeout(HEL,spoon1)]*
9	ELD	No, the second drawer.
10	HEL	*[Close(HEL,Drawer1),Open(HEL,Drawer2)]*
11	ELD	Yes, there it is.
12	HEL	This one?*[Takeout(HEL,spoon2)]*
13	ELD	Yes, uh-huh.
14	HEL	OK.

Figure 3: Find Task Example

3.1 Annotation

We devised a multidimensional annotation scheme since we are interested in investigating the role played in the interaction by modalities different from speech. Our annotation scheme comprises three main components: the multimodal event annotation, which includes annotating for pointing gestures, haptic-ostensive actions, their features, and their relationships to utterances; the dialogue act annotation; and the referential expression annotations already described in (Chen et al., 2011; Chen and Di Eugenio, 2012).

3.1.1 Multimodal Event Annotation

To study the roles played by different sorts of multimodal actions, and how they contribute to the flow of the dialogue, pointing gestures, Haptic-Ostensive (H-O) actions, and the relations among them have been annotated on the *Find* subcorpus. The *Find* subcorpus contains 137 *Find* tasks, collected from the dialogues of 19 pairs of subjects from the larger corpus. [1] The multimodal annotation tool Anvil (Kipp, 2001) was used to transcribe all the utterances, and to annotate for all categories described in this paper. Each annotation category is an annotation group in Anvil. For each subject, one track is defined for each annotation group, for a total of 4 tracks per subject in Anvil.

Pointing gestures are used naturally when people refer to a far away object. We define a pointing gesture as a hand gesture without physical contact with the target. Our definition of pointing gesture does not include head or other body part movements used to indicate targets. Our corpus includes very few occurrences of those; additionally, our collaborators in the RoboHelper project focus on recognizing hand gestures. We have identified two types of pointing gestures. The first is, pointing gestures with an identifiable target, which is usually indicated by a short time stable hand pointing. The other type is without a fixed target. It usually happens when the subject points to several targets in a short time, or the subject just points to a large space area.

For a pointing gesture, we mark two attributes: the time span and the target. The time span of a pointing gesture starts when the subject initiates the hand movement, ends when the subject starts to draw the hand back. We have devised a Referring Index System (Chen and Di Eugenio, 2012) to mark the different types of targets: single identifiable target, multiple identifiable targets and unidentifiable target.

During *Find* tasks, subjects need to physically interact with the objects, e.g. they need to open cabinets to get plates, to put a pot on the stove etc. Those physical contact actions often perform a referring function as well, either adding new entities to the discourse model, or referring to an already established referent. For example, in Figure 3, the action [Touch(Hel,Drawer1)] that accompanies Utt_4 disambiguates *This* by referring to Drawer1, tantamount to a pointing gesture; conversely, the action [Takeout(HEL,spoon1)] associated with Utt_8 establishes a referent for spoon1. Following (Foster et al., 2008), we label Haptic-Ostensive (H-O) those actions that involve physical contact with an object, and that can at the same time perform a referring function. Note that target objects here exclude the partner's body parts, as when HEL helps ELD get up from a chair.

No existing work that we know of identifies

[1] One pair of subjects was excluded, because ELD appeared confused. Our goal was to recruit elderly subjects with intact cognitive functions, but this subject was an exception.

types of H-O actions. Hence, we had to define our own categories, based on the following two principles: (1) The H-O types must be grounded in our data, namely, the definitions are empirically based: these H-O actions are frequently observed in the corpus. (2) They are within the scope of what our collaborators can recognize from the haptic signals. The five H-O action types we defined are:

- **Touch**: when the subject only touches the targets, no immediate further actions are performed
- **MANIP-HOLD**: when the subject takes out or picks up an object and holds it stably for a short period of time
- **MANIP-NO-HOLD**: when the subject takes out or picks up an object, but without explicitly showing it to the other subject
- **Open**: starts when the subject has physical contact with the handle of the fridge, a cabinet or a drawer, and starts to pull; ends when the physical contact is off
- **Close**: when the subject has physical contact with the handle of the fridge, a cabinet or a drawer, and starts to push; ends when the physical contact is off

For H-O action annotation, three attributes are marked: time span, target and action type. The "Target" attribute is similar to the "Target" attribute in pointing gesture annotation. Since H-O actions are more accurate than pointing gestures (Foster et al., 2008), the targets are all identifiable.

Table 1 provides distributions of the length in seconds for different types of events in the *Find* corpus. Table 2 shows the counts of different events divided by type of participant. From these two tables, it is apparent that:

- Pointing gestures and H-O actions were frequently used: their total corresponds to 61% of the number of utterances
- Utterances are short: only 1.7", and 4.2 words on average
- ELD performed 66% of pointing gestures, and HEL 97.5% of H-O actions

Multimodal Event Relation Annotation. Pointing gestures and H-O actions can accompany an utterance, e.g. see move 2 in Figure 3: HEL asks "Down there" while pointing to a drawer; or can be used independently, e.g. see move 6 in Figure 3: HEL does not utter any words, but opens the drawer after ELD confirms that is the right drawer with "Uh-huh". In the latter case, HEL used an action to respond to ELD. Pointing gestures and H-O actions are followed by utterances as well, e.g. move 11 in Figure 3: after HEL opens a drawer, ELD says "Yes, there it is".

Utterances	Pointing	H-O Actions	Total
2555"	571"	1088"	4377"

Table 1: Find Subcorpus: Length in seconds

	ELD	HEL	Total
Utterances	756	760	1516
Words	3612	2981	6593
Pointing	219	113	332
H-O Actions	15	582	597

Table 2: Find Subcorpus: Counts

To understand how pointing gestures and H-O actions participate in the dialogues and how they interact with utterances, we further annotated the relationship between utterances, pointing gestures and H-O actions. Just using timespans is not sufficient. It is not necessarily the case that utterance U is associated with gesture / H-O action G if their timespans overlap. This type of annotation is purely local: the fact that turns 2-5 in Figure 3 confirm which drawer to open, would be captured at the dialogue game level.

First, we assign to each utterance, pointing gesture and H-O action a unique event index, so that we can refer to these events with their indices. For pointing gestures and H-O actions, we define two more attributes: "associates" and "follows". If a pointing gesture or H-O action is associated with an utterance, the "associates" value will be the index of that utterance; by default, the "associates" value is empty. If a pointing gesture or H-O action independently follows an utterance, the "follows" value will be that utterance's index. E.g., for move 6 in Figure 3, we mark the H-O action "Open" with "follows [5]".

For utterances, we only mark the "follows" attribute. If an utterance directly follows a pointing gesture or H-O action, we use the index of the pointing gesture or H-O action as the "follows" value. By default, the "follows" attribute of an utterance is empty. It means that an utterance fol-

lows its immediate previous utterance.

We define a *move* as any combination of related utterances, pointing gestures and H-O actions, performed by the same subject. On the basis of the event relation annotations, we can compute the dialogue's move flow using the following algorithm.

1. Order all the utterances in a *Find* task session by the utterance start time
2. Until all the utterances are processed, for each unprocessed utterance u_i:
   (a) If u_i *follows* a pointing gesture or H-O action, that pointing gesture or H-O action forms a new *move* m_k; add m_k to the sequence before u_i
   (b) Find all the pointing gestures and H-O actions labelled as *associates* of u_i. These events form the *move* m_i together with u_i
   (c) Recursively find the events which follow the last generated *move*, together with all their associated events to form another *move*

This algorithm computes 1791 *moves*, as shown in Table 3. More than 90% of pointing gestures are used with utterances. Only 377 out of 596 H-O actions are included in the *moves*, mostly because the H-O action "Close" frequently follows an "Open" action (these cases are not detected by the algorithm, because they don't advance the dialogue).

	ELD	HEL	Total
Utterances	545	507	1052
Pointing	9	11	20
H-O	5	213	218
Utterance&Pointing	209	100	309
Utterance&H-O	2	153	155
Total	770	984	1754

Table 3: Moves Statistics in Find Corpus

3.1.2 Dialogue Act Annotation

Since the *Find* corpus is task-oriented in nature, we built on the dialogue act inventory of HCRC MapTask, a well-known task oriented corpus (Anderson et al., 1991). The MapTask tag set contains 11 moves:[2] *instruct, explain, check, align, query-w, query-yn; acknowledge, reply-y, reply-n, reply-w, clarify*. However, this inventory of DAs does not cover utterances that are used to respond to gestures and actions, such as Utt.$_{11}$ in Figure 3. The semantics of the *reply-*{*y/n/w*} tags does not cover these situations. Hence, we devised three more tags, which apply **only** to statements that follow a move composed exclusively of a gesture or an action (in the sense of "follow" just discussed):

- **state-y**: a statement which conveys "yes", such as Utt.$_{11}$ in Figure 3.
- **state-n**: a statement which conveys "no", e.g. if Utt.$_{11}$ had been *Wait, try the third drawer.*
- **state**: still a statement , but not conveying acceptance or rejection, e.g. *So we got the soup.*

Hence, the DAs in {*state-y*, *state-n*, *state*} are used to tag responses to actions, and the DAs in {*reply-y*, *reply-n*, *reply-w*} are used to tag responses to utterances. Table 4 shows the distribution of DAs by subject.

Dialogue Act	ELD	HEL	Total	Ratio
Instruct	295	19	314	20.7%
Acknowledge	22	186	208	13.7%
Reply-y	179	3	182	12.0%
Check	1	155	156	10.3%
Query-yn	23	133	156	10.3%
Query-w	3	144	147	9.7%
Reply-w	132	4	136	9.0%
State-y	40	36	76	5.0%
State-n	16	50	66	4.4%
Reply-n	27	9	36	2.4%
State	7	15	22	1.5%
Explain	10	4	14	0.9%
Align	1	2	3	0.3%
Total	756	760	1516	100%

Table 4: Dialogue Act Counts in Find Corpus

Intercoder Agreement. In order to verify the reliability of our annotations, we double coded 15% of the data for pointing gestures, H-O actions and DAs. These are the dialogues from 3 pairs of subjects, and contain 22 *Find* tasks. Because the pointing gestures and H-O actions are time span based, when we calculate agreement, we use an overlap based approach. If the two annotations from the two coders overlap by more than 50% of the event length, and the other attributes are the same, we count this as a match. We used κ to measure the reliability of the annotation (Cohen, 1960). We obtained reasonable values: for pointing gestures, κ=0.751, for H-O actions, κ=0.703, and for DAs, κ=0.789.

[2] A twelfth move, *Ready*, does not appear in our corpus.

4 Experimental Setup

We ran experiments classifying the DA tag for the current utterance. We employ supervised learning approaches, specifically: Conditional Random Field (CRF) (Lafferty et al., 2001), Maximum Entropy (MaxEnt), Naive Bayes (NB), and Decision Tree (DT). These algorithms are widely used for DA classification (Sridhar et al., 2009; Ivanovic, 2008; Ha et al., 2012; Kim et al., 2012). We used Mallet (McCallum, 2002) to build CRF models. MaxEnt models were built using the MaxEnt [3] package from the Apache OpenNLP package. Naive Bayes and Decision Tree models were built with the Weka (Hall et al., 2009) package (for decision trees, we used the J48 implementation). All the results we will show below were obtained using 10 fold cross validation.

4.1 Features

Among our goals were not only to obtain effective classifiers, but also to investigate which kind of features are most effective for our tasks. As a consequence, beyond textual features and dialogue history features, we experimented with multimodal features extracted from other modalities, utterance features, and automatically inferred dialogue game features.

Textual features (TX) are the most widely used features for DA classification (Stolcke et al., 2000; Bangalore et al., 2008; Sridhar et al., 2009; Di Eugenio et al., 2010a; Kim et al., 2010; Boyer et al., 2011; Ha et al., 2012; Kim et al., 2012). The textual features we use include lexical, syntactic, and heuristic features.

- Lexical features: Unigrams of the words and part-of-speech tags in the current utterance. The words used in the features are processed using the morphology tool from the Stanford parser (De Marneffe and Manning, 2008).
- Syntactic features: The top node and its first two child nodes from the sentence parse tree. If an utterance contains multiple sentences, we use the last sentence. Sentences are parsed using the Stanford parser.
- Number of sentences and number of words in the utterance. We use Apache OpenNLP library [4] to detect sentences and tokenize them.
- Heuristic features: whether an utterance contains WH words (e.g. *what, where*), whether an utterance contains yes/no words (e.g. *yes, no, yeah, nope*).

Utterance features (UT) are extracted from the current utterance's meta information. Previous research showed that utterance meta information such as the utterance speaker can help classify DAs (Ivanovic, 2008; Kim et al., 2010).

- The actor of the utterance
- The time length of the utterance
- The distance of the current utterance from the beginning of the dialogue

The **pointing gesture feature (PT)** indicates whether the actor of the current utterance u_i is making a pointing gesture G, i.e., whether G is associated with u_i, and hence, part of move m_i.

Haptic-Ostensive features (H-O) indicate whether the actor of the current utterance u_i is performing any H-O action G i.e., whether G is associated with u_i, and hence, part of move m_i; and the type of that action, if yes.

Location features (LO) include the locations of the two actors, whether they are in the same location, whether the actor of the current utterance changes the location during the utterance. Since we do not have precise measurement of subjects' locations, we annotate approximate locations by dividing the apartment into four large areas: kitchen, table, lounge and bed.

The **dialogue game feature (DG)** models hierarchical dialogue structure. Some previous research on DA classification has shown that hierarchical dialogue structure encoded via the notion of conversational games (Carlson, 1983) significantly improves DA classification (Hastie et al., 2002; Sridhar et al., 2009; Di Eugenio et al., 2010a). In MapTask, a game is defined as a sequence of moves starting with an initiation (instruct, explain, check, align, query-yn, query-w) and encompassing all utterances up until the purpose of the game has been fulfilled, or abandoned. In the *Find* corpus, dialogue games have not been annotated. In order to use the DG feature, we use a just-in-time approach to infer dialogue games. For each dialogue, we maintain a stack for dialogue games. When an utterance is classified as an initiating DA tag, we assume the dialogue has

[3] http://maxent.sourceforge.net
[4] http://opennlp.apache.org/

entered a new dialogue game, and push the DA label as the dialog game to the top of the stack. The DG feature value is the top element of the stack. The dialogue game feature is always inferred at run time during classification process, just before an utterance is being processed. Hence, when we classify the DA for the current utterance u_i, the DG value that we use is the closest preceding initiating DA.

Dialogue history features (DH) model what happened before the current utterance (Sridhar et al., 2009; Di Eugenio et al., 2010a). We encode:

- The previous move's actor
- Whether the previous move has the same actor as the current move
- The type of the previous move; if it is an utterance, its DA tag; if it is an H-O action, the type of H-O action

5 DA Classification Experiments

We ran the DA classification experiments with three goals. First, we wanted to assess the effectiveness of different types of features, especially, the effectiveness of gesture, H-O action, location and dialogue game features. Second, we wanted to compare the performances of different machine learning algorithms on such a multimodal dialogue dataset. Third, we wanted to investigate the performances of different algorithms in the online and offline experiment settings. The DA classification task could be treated as a sequence labeling problem (Stolcke et al., 2000). However, different from other sequence labeling problems such as part-of-speech tagging, a dialogue system cannot wait until the whole dialogue ends to classify the current DA. A dialogue system needs online DA classification models to classify the DAs when a new utterance is processed by the system. There are two differences between online and offline DA classification modes. First, when we generate the dialogue history and dialogue game features, we use the previously classified DA tag results for online mode, while we use the gold-standard DA tags for offline mode. Second, MaxEnt (using beam search) and CRF evaluate and classify all the utterances in a dialogue at the same time in offline mode; however in online mode, MaxEnt and CRF can only work on the partial sequence up to the utterance to classify. Whereas this may sound obvious, it explains why the performance of these classifiers may be even more negatively affected in online mode with respect to their offline performance, as compared to other classifiers. We will see that indeed this will happen for CRF, but not for MaxEnt.

To evaluate feature effectiveness, we group the features into seven groups: textual features (TX), utterance features (UT), pointing gesture feature (PT), H-O action features (H-O), location features (LO), dialogue game feature (DG), dialogue history features (DH). Then we generate all the combinations of feature groups to run experiments. For each classification algorithm, we ran 10-fold cross-validation experiments, for each feature group combination, in both online and offline mode. It would be impossible to report all our results. Similarly to (Ha et al., 2012), we report our results with single feature groups and incremental feature group combinations, as shown in Table 5. Whereas all combinations were tried, the omitted results do not shed any additional light on the problem. The majority baseline, which always assigns the most frequent tag to every utterance, has an accuracy of 20.3%.

The CRF offline model performs best, which confirms the results of (Kim et al., 2010; Kim et al., 2012). This is due to the strong correlation between dialogue history features (DH) and the states of the CRF. In online mode, when there is noise in the previous DA tags, the CRF's performance drops significantly ($p \leq .005$, using χ^2). A significant drop in performance from offline to online mode also happens to NB ($p \leq .005$) and DT ($p < .025$). MaxEnt performs very stably, the best online model performs only .015 worse than the best offline model. The best MaxEnt offline model beats the other algorithms' best models except CRF, while the MaxEnt online model outperforms all the other algorithms' online models. Our results thus demonstrate that MaxEnt works best for online DA classification on our data.

As concerns features, for online models, textual features (TX) are the most predictive as a feature type used by itself. When we add pointing gesture (PT), H-O features (H-O) and location features (LO) together to textual features, we notice a significant performance improvement for most models (except CRF models). For MaxEnt, which gives the best results for online models, none of the gesture, H-O action and location features alone significantly improve the results, but all three to-

Features	CRF		MaxEnt		NB		DT	
	Offline	Online	Offline	Online	Offline	Online	Offline	Online
1. TX (Textual)	.654	.641	.630	.630	.449	.453	.450	.450
2. UT (Utterance)	.506	.376	.353	.353	.417	.417	.392	.392
3. PT (Pointing)	.225	.155	.210	.210	.212	.212	.212	.212
4. H-O (Haptic-Ostensive)	.187	.147	.237	.237	.243	.243	.212	.212
5. LO (Location)	.259	.176	.264	.264	.259	.259	.265	.265
6. DG (Dialogue Game)	.737	.136	.305	.189	.212	.212	.212	.212
7. DH (Dialogue History)	**.895**	.119	.480	.302	.478	.284	.471	.294
8. TX+PT	.654	.651	.639	.639	.453	.453	.450	.450
9. TX+PT+H-O	.670	**.649**	.637	.637	.456	.456	.449	.449
10. TX+PT+H-O+LO	.648	.645	.657*	.657*	.523*	.523*	.536*	.536*
11. TX+PT+H-O+LO+UT	.668	.612	.685	.685	.563	.563	.568	.568
12. TX+PT+H-O+LO+UT+DG	.770**	.528	.722**	.709**	.566	.591**	.576	.607**
13. TX+PT+H-O+LO+UT+DG+DH	.847†	.475	**.757**†	**.742**†	**.635**†	**.606**	**.671**†	**.627**

Table 5: Dialogue Act Classification Accuracy: * indicates significant improvement after adding PT+H-O+LO to TX (cf. lines 1 and 10); ** indicates significant improvement after adding DG to TX+PT+H-O+LO+UT (cf. lines 11 and 12); †indicates significant improvement after adding DH to TX+PT+H-O+LO+UT+DG (cf. lines 12 and 13); bold font indicates the feature group set giving best performance for each column.

gether do. This confirms the finding of (Ha et al., 2012) that non-verbal features help DA classification. To assess which feature is the most important among those three non-verbal features, we examined the experiment results with a leave-one-out strategy, that is for each classifier in offline and online modes, we leave one of the gesture, H-O and location features out from the full experiment feature set (TX+PT+H-O+LO+UT+DG+DH). No significant difference was discovered.

When the dialogue game features (DG) are added to the models, performance increases significantly for CRF offline model ($p<.005$), MaxEnt offline ($p<.005$) and online ($p<.05$) models, NB online model ($p<.05$) and DT online model ($p<.005$). It confirms previous findings, including by our group (Di Eugenio et al., 2010a), that dialogue game features (DG) play a very important role in DA classification, even via the simple approximation we used. When the dialogue history features (DH) are added to the models, performance increased significantly for all the offline models and the MaxEnt online model, with $p<.005$. This confirms previous findings that dialogue history helps with DA classification.

6 Conclusions and Future Work

In this paper we described our multimodal corpus which is annotated with multimodal information (pointing gestures and H-O actions) and dialogue acts. Our corpus analysis shows that people actively use pointing gestures and H-O actions alongside utterances in dialogues. The function of H-O actions in dialogue had hardly been studied before. Our experiments show that MaxEnt performs best for the online DA classification task. Multimodal and dialogue game features both improve DA classification.

Short-term future work includes manual annotation for dialogue games, in the hope that more accurate dialogue game features may further improve DA classification. Longer term future work includes prediction of the specific next move – the specific DA and/or the specific gesture, pointing or H-O action. We have now developed some of the building blocks of an information-state based multimodal dialogue manager. The major aspects we still need to address are defining the information-state for the *Find* task, and developing rules to update the information-state with multimodal information, the classified DAs, and the co-reference resolution models we already built (Chen et al., 2011; Chen and Di Eugenio, 2012). Once the information-state component is in place, we can expect better and more detailed predictions.

Acknowledgments

This work is supported by award IIS 0905593 from the National Science Foundation. Thanks to the other members of the RoboHelper project, for their many contributions, especially to the data collection effort.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, and Henry S. 1991. The HCRC Map Task corpus. *Language and Speech*, 34(4):351.

Srinivas Bangalore, Giuseppe Di Fabbrizio, and Amanda Stent. 2008. Learning the structure of task-driven human–human dialogs. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1249–1259.

K.E. Boyer, J.F. Grafsgaard, E.Y. Ha, R. Phillips, and J.C. Lester. 2011. An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1190–1199. Association for Computational Linguistics.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Lauri Carlson. 1983. *Dialogue games: An approach to discourse analysis*. D. Reidel Publishing Company.

Lin Chen and Barbara Di Eugenio. 2012. Co-reference via pointing and haptics in multi-modal dialogues. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* The Association for Computational Linguistics.

Lin Chen, Anruo Wang, and Barbara Di Eugenio. 2011. Improving pronominal and deictic coreference resolution with multi-modal features. In *Proceedings of SIGdial 2011, the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 307–311, Portland, Oregon, June. Association for Computational Linguistics.

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

Marie-Catherine De Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

Barbara Di Eugenio, Zhuli Xie, and Riccardo Serafin. 2010a. Dialogue act classification, higher order dialogue structure, and instance-based learning. *Dialogue & Discourse*, 1(2):1–24.

Barbara Di Eugenio, Miloš Žefran, Jezekiel Ben-Arie, Mark Foreman, Lin Chen, Simone Franzini, Shankaranand Jagadeesan, Maria Javaid, and Kai Ma. 2010b. Towards Effective Communication with Robotic Assistants for the Elderly: Integrating Speech, Vision and Haptics. In *Dialog with Robots, AAAI 2010 Fall Symposium*, Arlington, VA, USA, November.

M.E. Foster, E.G. Bard, M. Guhe, R.L. Hill, J. Oberlander, and A. Knoll. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pages 295–302. ACM.

Simone Franzini and Jezekiel Ben-Arie. 2012. Speech recognition by indexing and sequencing. *International Journal of Computer Information Systems and Industrial Management Applications*, 4:358–365.

David Graff, Alexandra Canavan, and George Zipperlen. 1998. Switchboard-2 Phase I.

Eun Young Ha, Joseph F. Grafsgaard, Christopher Mitchell, Kristy Elizabeth Boyer, and James C. Lester. 2012. Combining verbal and nonverbal features to overcome the "information gap" in task-oriented dialogue. In *Proceedings of SIGdial 2012, the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 247–256, Seoul, South Korea, July. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Helen Wright Hastie, Massimo Poesio, and Stephen Isard. 2002. Automatically predicting dialogue structure using prosodic features. *Speech Communication*, 36(1–2):63–79.

Edward Ivanovic. 2008. Automatic instant messaging dialogue using statistical models and dialogue acts. Master's thesis, University of Melbourne.

Maria Javaid and Miloš Žefran. 2012. Interpreting communication through physical interaction during collaborative manipulation. Draft, October.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of EMNLP 2010, the Conference on Empirical Methods in Natural Language Processing*, pages 862–871. Association for Computational Linguistics.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2012. Classifying dialogue acts in multi-party live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 463–472, Bali, Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

Michael Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 1367–1370.

Kristine M. Krapp. 2002. *The Gale Encyclopedia of Nursing & Allied Health*. Gale Group, Inc. Chapter Activities of Daily Living Evaluation.

John D. Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc.

L. Levin, A. Thymé-Gobbel, A. Lavie, K. Ries, and K. Zechner. 1998. A discourse coding scheme for conversational Spanish. In *Fifth International Conference on Spoken Language Processing*.

K. Ma and J. Ben-Arie. 2012. Multi-view multiclass object detection via exemplar compounding. In *IEEE-IAPR 21st International Conference on Pattern Recognition (ICPR 2012)*, Tsukuba, Japan, November.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

E. Shriberg, R. Dhillon, S.V. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Cambridge, MA, April 30-May 1.

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

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C.V. Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Matthias Zimmermann, Andreas Stolcke, and Elizabeth Shriberg. 2006. Joint segmentation and classification of dialog acts in multiparty meetings. In *ICASSP 2006, the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1. IEEE.