Classifying Dialogue Acts in Multi-party Live Chats

Su Nam Kim
School of IT
Monash University
Clayton, VIC, Australia
su.kim@monash.edu.au

Lawrence Cavedon
School of CS and IT
RMIT University
Melbourne, VIC, Australia
lawrence.cavedon@rmit.edu.au

Timothy Baldwin
Dept. of Computing and
Information Systems
The University of Melbourne
VIC, Australia
tb@ldwin.net

Abstract

We consider the task of classifying chat contributions by dialogue act in a multi-party setting. This extends the problem significantly over the 1-1 chat scenario due to the semi-asynchronous and "entangled" nature of the contributions by chat participants. We experiment with a number of machine learning approaches, using different categories of features: lexical, contextual, structural, keyword and dialogue interaction information. For evaluation, we developed gold-standard data using online forums from the USA Library of Congress. We found that, for multi-party dialogues, features based on 1-gram and keywords produced best performance, while features exploiting structure and interaction did not perform as well as previously reported results over 1-to-1 chats.

1 Introduction

Dialogue Acts (or **DAs**) are discourse units (or utterances) that represent the semantics of contributions to a dialogue at the level of illocutionary force. Dialogue acts have been studied in various types of conversations — spoken/written dialogue contributions (Stolcke et al., 2000; Wu et al., 2002; Kim et al., 2010a), sentence-level (Lampert et al., 2008), paragraph-level (Cong et al., 2008), or complete messages consisting of several paragraphs (Cohen et al., 2004). Authors have argued that automatic dialogue act identification could help in a range of applications, such as meeting summarisation (Murray et al., 2006), email summarisation, conversational agents, speech recognition (Stolcke et al., 2000), or human social intention detection (Jurafsky et al., 2009). They can also be useful in information-sharing chats in online forums (Kim et al., 2010b; Wang et al., 2011).

Recently, **live chat** has received growing attention since chat services and similar applications have gained popularity as a communication method. However, the majority of previous work on dialogue act classification for dialogue has been carried out over *spoken* dialogue. Although spoken and written dialogue have similarities, they have distinct features which make it difficult to reuse existing methods for live chats. For example, spoken dialogue introduces difficulties due to errors inherent in speech recognition output, but allows acoustic and prosodic features to be leveraged (e.g. Stolcke et al. (2000)). Conversely, live chats introduce other types of complications, including ill-formed data and *entanglement* (especially for multi-party conversations) due to the semi-asynchronous nature of the interaction (e.g. (Werry, 1996)). As a result, studying live chats is a necessary step toward building accurate live chat systems.

To date, relatively little work has targeted dialogue act classification over live chat data. Wu et al. (2002) and Forsyth (2007) investigated multi-party casual chats, while Ivanovic (2008) and Kim et al. (2010a) focused on 1-on-1 chats in customer service centre settings. However, these previous approaches are not directly applicable to other types of live chats, such as forum-style chats that allow multiple participants. Additionally, many live chat applications, such as online forums and online meetings, presume an environment that allows multiple

26th Pacific Asia Conference on Language,Information and Computation pages 463–472

participants to discuss specific topics. While Forsyth (2007) investigates chat involving multiple participants, the conversations are casual and not topic-focused. The semantics and structure of dialogues depend on the nature and structure of the conversations, thus requiring different dialogue act categories and classification approaches.

In this paper, we target the classification of dialogue acts in multi-user forums carried out through live chats. 1-on-1 live chats are popular for consumer service support or individual meetings. However, this does not allow multiple users to participate in the chats. On the other hand, as more meetings are taking place via live chat, we believe that studying live chats in multi-user environments is a necessary step towards building such systems. In addition, we have developed a live chat dataset from library forum chats, involving multiple simultaneous users. The dataset contains live chats extracted from online forums conducted at the US Library of Congress.

To develop automatic methods for dialogue act classification in live chats, we explored four types of features: *context*, *structure*, *keyword*, and *dialogue interaction*. In addition, we compare the systems in terms of the number of participants as well as the types of chats (i.e., casual vs. forum chats). In evaluation, we investigate the utility of each feature category over different types of live chat over two multi-user datasets: (i) online forums from the US Library of Congress, and (ii) Forsyth's NPS (Naval Postgraduate School) casual chats, and.

2 Task Setup

We experiment with two different types of live chats: (i) forum chats involving specific discussion topics; and (ii) casual chats (i.e., (Forsyth, 2007)'s NPS casual chat data). Since there was no existing available data for the first type, we developed the data for evaluation ourselves. The remainder of this section describes the data and dialogue act categories in detail.

2.1 Dataset 1: Live Forum Chats

We collected online forum chats with multiple participants from the US Library of Congress. The live chats contain 33 online discussions that the Library's Educational Outreach team hosted for teachers between 2002 and 2006.

To define dialogue acts suitable for this data, we investigated existing sets of dialogue acts from both spoken dialogues and live chats. Many have been based on the Dialogue Act Markup in Several Layers (DAMSL) scheme (Allen and Core, 1997), initially applied to the TRAIN corpus of transcribed spoken task-oriented dialogues. In live chats, Wu et al. (2002) and Forsyth (2007) defined 15 dialogue acts for casual online conversations based on previous sets (Samuel et al., 1998; Jurafsky et al., 1998; Stolcke et al., 2000) and characteristics of conversations. Ivanovic (2008) proposed 12 dialogue acts applying DAMSL to customer service chats.

We found that forum chats are not dissimilar to customer service chats in terms of the nature of conversations (e.g. question, request, thanking, etc.), and so decided to adopt the DA set defined by Ivanovic (2008). To the 12 dialogue acts from Ivanovic (2008), we added two further dialogue acts — BACKGROUND and OTHER. BACKGROUND is designed to cover contributions containing information about the participants themselves, which often occurs before discussions. OTHER covers chat contributions that do not belong to any other dialogue acts. We also compared our defined set of DAs to that for NPS casual chats in Forsyth (2007). Although both datasets contain multiple participants, they differ in the nature of their content; thus, we found problems applying the DA set from Forsyth (2007) directly to the library chat forums. However, we observed that there is overlap between the two sets of dialogue acts (e.g. (OPENING vs. GREET), (EXPRESSION vs. EMOTION), YN/WH-QUESTION for both, etc.). The final list of dialogue acts we applied to the library dataset is shown in Table 1.

In preprocessing, we first removed system log messages.[1] Second, we replaced expressions such as emoticons and exclamations (e.g. *:-), wow*), email addresses (e.g. (learningpage@loc.gov), URLs (e.g. *http://memory.loc.gov*), locations (e.g. *Texas*), and institute names (e.g. *University of Houston*) with the tokens *EMOTION, EMAIL, URL, LOCATION, INSTITUTE*, respectively. We also replaced user

[1] System log messages indicate the status of participants, such as *join* and *depart*.

Dialog Act	Example	%	Dialog Act	Example	%
STATEMENT	we have a website for photo gallery.	47.76	WH-QUESTION	What is this?	3.26
RESPONSE-ACK	yes, great, i agree,..	11.73	OPENING	Hi, Greeting!	3.03
EXPRESSION	:-), wow, oh!	7.71	YES-ANSWER	yes, sure,	1.67
THANKING	thanks, thank you for ..	6.54	CLOSING	bye, good night,..	1.55
YN-QUESTION	is there a website for .. ?	5.84	DOWNPLAY	no problem, you're welcome!	0.49
REQUEST	click this, go to xx..	4.97	OTHER	or, but	0.40
BACKGROUND	i am user2, i teach 4th grad	4.76	NO-ANSWER	no, nope	0.28

Table 1: Dialogue Act Tagset for the USA Library of Congress forum Chats: definitions and examples

names with the unique token *USER_ID* for privacy. Third, we applied a sentence tokenizer in order to separate the data into tentative discourse units, then further manually segmented/confirmed the units. This culminated in 5,276 utterances over 15 library forum chats, each containing at least 200 discourse units (between 238 and 666 discourses per live chat). The proportion of utterances for each dialogue act type is listed in Table 1.

To develop a gold-standard, we hired two annotators (including one of the authors) both of whom have significant experience in annotating similar datasets. Before conducting the actual annotation task, we conducted a pilot task over library forum chats which were not selected in our final dataset, and confirmed the feasibility of the dialogue acts. The inter-annotator agreement was 81.4% with kappa value 0.74, indicating reliable agreement.

2.2 Dataset 2: Casual Live Chats

We used the NPS casual chats developed by Forsyth (2007) as our second dataset. Table 2 shows the dialogue act tags, examples, and the distribution of dialogue acts in the dataset. The dataset contains 10,567 utterances spanning 15 conversations. It also includes POS tags which are modified to make it more specific to the categories based on Penn Treebank tags. For privacy, the actual user names have been replaced with anonymous IDs. Forsythe reports that one person labeled and verified the gold-standard labels and automatically tagged POS tags; thus, Forsythe does not report any inter-annotator agreement statistics on the NPS dataset. It is also hard to re-annotate the NPS dataset to ascertain inter-annotator agreement statistics, due to a lack of published guidelines.

3 Features

3.1 Bag-of-words Features

Contextual information has been used for dialogue act classification with both spoken and written dialogues (e.g. (Samuel et al., 1998; Bangalore et al., 2006; Ivanovic, 2008)). For live chats, Ivanovic (2008) and Kim et al. (2010a) used unigrams and/or variations of n-grams as basic features. Kim et al. (2010a) suggest that higher-order n-grams (i.e., 2-grams and both 1,2-grams) do not perform significantly better than using unigrams only, and that using lemmas performs better than using raw words. Based on these previous results, we tested raw and lemmatized unigrams only, with TF·IDF and Boolean values as our base features. In addition, we noticed that despite the data appearing cleaner than that of Ivanovic (2008), there are still typos and out-of-vocabulary words in the data. To handle these, we tested word-stems as an attempt to reduce errors from those words. In sum, we tested 12 combinations, using (raw, lemmatized, stemmed 1-grams), (with and without POS), (with Boolean vs. TF·IDF values). Details of lemmatization and stemming are presented in Section 4.1.

3.2 Structural Information

Kim et al. (2010a) has demonstrated the effectiveness of structural information for classifying dialogue acts over 1-on-1 live chats. Most live chat sessions we used are significantly longer and contain multiple participants, thereby reducing the alignment of related dialogue-contributions. However, we observed that there is still some degree of structural regularity, e.g. GREETING at the beginning and ending, and the presence of BACKGROUND

Dialog Act	Example	%	Dialog Act	Example	%
STATEMENT	well i thought you and I will end up together :-(	30.14	EMPHASIS	I do believe he is right.	1.80
SYSTEM	JOIN	24.91	CONTINUER	an thought I’d share	1.59
GREET	hiya 10-19-40sUser43 hug	12.90	REJECT	u r not on meds	1.50
EMOTION	lol	10.47	YES ANSWER	why yes I do 10-19-40sUser24, lol	1.02
YES/NO Q.	cant we all just get along	5.20	NO ANSWER	no I had a roomate who did though	0.68
WH-QUESTION	11-08-20sUser70 why do you feel that way?	5.04	CLARIFY	i meant to write the word may	0.36
ACCEPT	yeah it does, they all do	2.20	OTHER	sdfjsdfjlf	0.33
BYE	night ya’all	1.85			

Table 2: Dialogue Act Tagset for the NPS Casual Chats: definitions and examples

after GREETING. Our second observation is that some dialogue acts are associated with shorter dialogues, e.g. *hi, bye* for GREETING, or *excellent* for EXPRESSION. Third, we observed that some users tend to ask questions while others tend to answer them. Similar to representatives at customer service centers, hosts of the forums tend to request actions or to pose questions. Based on our observations, we tested the following four features:

- *Distance*: The distance from the first utterance to the target utterance. We test both absolute distance ($Distance_{absolute}$) and percentage distance relative to the total conversation ($Distance_{relative}$);

- *TermCount*: The number of terms in the target utterance;

- *UserID*: User ID (1–180 for library forum chats, 1 –1,377 for NPS casual chats);

- *User=Host?*: If User of target utterance is the host (1) or not (0). Note that this feature is applied to library forum chats only.

Note that in online systems, we do not know the total length of conversations, and thus the feature $Distance_{relative}$ (relative position of the target utterance) is tested only for comparison purposes.

3.3 Keyword Information

Forsyth (2007) used manually-crafted keywords for classification and reported high accuracies even with this simple technique. Stolcke et al. (2000) also reported that lexical features were strong indicators of dialogue act in spoken dialogue. We similarly observed that some words are strongly associated with specific dialogue acts. However, since the nature and dialogue acts of different datasets are themselves different, specific keywords are needed for our library forum chats. As such, we first selected candidate terms for keywords by using the frequent terms per DA and manually extracted keywords which are associated with specific dialogue acts. In essence, keywords are not equivalent to the full set of n-grams, but rather a targeted subset of n-grams (of varying length) associated with specific dialogue acts. The following list shows examples of keywords for dialogue acts over library forum chats. Note that since STATEMENT includes unfocused chats, we do not propose keywords for this dialogue act.

- BACKGROUND: *I live, location, institute*

- OPENING: *hi, hello, greeting, welcome*

- CLOSING: *see you, bye, goodnight*

- THANKING: *thank you, thanks*

- DOWNPLAY: *no problem, you’re welcome*

- EXPRESSION: *emotion*

- YES-ANSWER: *yes* with a question mark in the preceding 5 sentences

- NO-ANSWER: *no* (without *problem*)

- REQUEST: *please, click, let’s*

- RESPONSE-ACK: *!, great, yes, ok* in Utterances of length ≤ 3

- WH-QUESTION: question mark with *how, what, when, where, who, why*

- YN-QUESTION: question mark without *how, what, when, where, who, why*

For the NPS casual chats, we used all keyword features (f0–f26) defined in Forsyth (2007). However, we observed that some of his features are not available at the time of the target utterance unless we have access to the completed conversation (e.g. *f3. Number of posts in the future that contained a Yes or No word*) — i.e., for online systems, these features are not usable. As a result, we tested two different sets of features: using all features; and using only those based on information available at the target utterance. It is also not clear how to extract the exact same keywords as used in Forsyth (2007), and as such, we expect our results to differ slightly from those in the original paper.

In addition, to overcome the data dependency of the keyword feature, we proposed new features using the distribution of terms over dialogue acts. That is, we computed the term frequency (TF) of each term over the 14 dialogue acts in the training data, and accumulated TF from all terms in the target utterance into a 14×1 vector to represent the feature for the target utterance. For example, suppose that for the target utterance *Welcome back*, *welcome* occurs 100 and 20 times with dialogue-acts OPENING and DOWNPLAY, respectively, and *back* occurs 10 and 5 times with OPENING and STATEMENT, respectively. Then the TF vectors for the terms are "100 0 0 0 0 0 0 0 0 0 0 0 20 0" and "10 0 0 0 0 5 0 0 0 0 0 0 0 0". By adding all TFs from both terms, we finally obtain "110 0 0 0 0 5 0 0 0 0 0 0 20 0" as the final feature for the target utterance. We also tested three different TF values listed below. Further, we tested two different options for choosing terms in an utterance: using all terms vs. using selected terms for which the majority label has TF of at least 50%. Returning to our example from above, for *back*, the proportion of TF with OPENING and STATEMENT is 0.333 and 0.677, respectively — thus, none of the labels have 50% total TF for *back*, and we would hence discard this term for the "selected terms" option.

- $\textit{InfoDistribution}_{Raw/Raw.5}$: raw counts;

- $\textit{InfoDistribution}_{Percent/Percent.5}$: percentage counts;

- $\textit{InfoDistribution}_{Label/Label.5}$: a dialogue act with maximum TF.

3.4 Interaction among Utterances

Finally, we investigated the interaction between features proposed in Bangalore et al. (2006) and Kim et al. (2010a). Bangalore et al. (2006) used sentences to provide dialogue act information of previous utterances over spoken dialogues; Kim et al. (2010a) used predicted dialogue acts directly. A major point of difference for us is that our data contains multiple participants; thus, the interactions among utterances tend to be more indirect. Moreover, due to difficulties in utterance disentanglement similarly shown in Elsner and Charniak (2008)), we expect reduced effectiveness over our data of such information (although some degree of interaction exists). However, to partly help with disentanglement, we noticed that some users mentioned the user name(s) of the users they are responding to in their posts, which allows us to identify the utterances they link to. Based on these observations, we tested the five interaction features listed below:

- *Prev1, Prev2, Prev3*: dialogue act(s) or sentence(s) from $1 \sim 3$ previous utterances;

- *User*: a dialogue act or sentence from 1 previous utterance in which the user is the same as the author in the target utterance;

- *TextUser*: A dialogue act or sentence from 1 previous utterance which is authored by the user mentioned in the target utterance. For example, for *USER63, thanks for the information.*, we would identify *USER63* as the user and use his/her latest utterance as a feature.

4 Evaluation

4.1 Experimental Setup

To develop our system, we first performed POS tagging using Lingua::EN::Tagger, lemmatization using *morph* (Minnen et al., 2001), and stemming using the English Porter stemmer.[2]

For our learners, we used the Naïve Bayes (NB) implementation in the WEKA machine learning toolkit (Witten and Frank, 2005), a support vector machine (SVM),[3] and the CRF implementation in Mallet (McCallum, 2002).[4] We ran 15-fold cross validation, using our 15 dialogues. All results are reported in terms of micro-averaged F-score, unless otherwise noted. As a baseline, instead of using the majority vote (**47.76** and **24.91** for library forum chats and NPS casual chats, respectively), we used a system built using bag-of-words features only (see Table 3), one for each machine learner.

4.2 Result 1: Bag-of-Words

Table 3 shows the performances of our different dialogue act classification systems using variations of 1-grams. It shows that stemmed unigrams without POS tags performed best with BoW features over library forum chats, while stemmed unigrams with POS tags generally achieved the highest performance over NPS casual chats. Note that we only show performance using *Boolean* values, since those using TF·IDF were lower. We also tested 2-grams and mixed 1/2-grams, and found they each reduced performance. Overall, we found that stemming improved performance. We noticed that for library forum chats, due to ill-formed data, POS tagging did not perform well. On the other hand, POS tags in NPS casual chats were improved by the automatic method (see Forsyth (2007) for how to correctly perform POS tagging over casual chats). As a result, performance using POS tags is different over the two different sets. However, by considering the performance over NPS casual chats, we conclude that high-quality POS tags would help to improve classification performance. Between the three learners, the CRF performance is superior to the others; this aligns with previous research (e.g. (Kim et al., 2010a)) most likely because the conversations are structured, despite the entanglement issue.

4.3 Result 2: Structural Information

Table 4 shows the performance using structural features. As base systems, we used *stemmed unigrams* for library forum chats and *stemmed unigrams with POS tags* for NPS casual chats, since all three learners generally performed best using stemmed unigrams. Overall, we found that structural features did not work well in multi-party live chats, in contrast to the results of Kim et al. (2010a) over 1-on-1 live chats. We presume this is because the data contains multiple participants, blurring the structural information. The semi-asynchronous nature of the interaction also poses serious issues for disentanglement, thus adding more difficulty in identifying the association between dialogue acts and structural information. However, term counts and user ids improved the performance slightly over NPS casual chats. Also, the *User=Host?* feature worked best using the CRF for library forum chats. We hypothesize that this is because the hosts tend to have stronger association with specific dialogue acts (e.g. REQUEST) in this setting. A user's contribution to the conversation would also be associated with some specific dialogue acts (e.g. some tend to ask while others tend to answer). As discussed in Section 3.2, we compared absolute and relative distances and found no difference.

4.4 Result 3: Keyword Information

Table 5 shows the performance using keyword features. As above, the base systems use *stemmed unigrams* for library forum chats and *stemmed unigrams with POS tags* for NPS casual chats, since overall, structural information did not improve performance. With library forum chats, we found that adding keywords to contextual features improved performance over all three learners, since some terms occur only in specific dialogue acts. However, with the NPS casual chats, keyword features did not perform well, in contrast to the findings of Forsyth (2007). We hypothesis that the lower performance is due to the specific keywords used in this work, as compared to those used in Forsyth (2007). We also found that using selected features that are available at the time of the target utterance performed better. This suggests that classifying dialogue acts can be performed as an online task. Further, using the dis-

[2] Available at `http://tartarus.org/~martin/PorterStemmer/`

[3] `http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html`

[4] `http://mallet.cs.umass.edu`

	Library Forum						NPS					
	without POS			with POS			without POS			with POS		
	NB	SVM	CRF	NB	SVM	CRF	NB	SVM	CRF	NB	SVM	CRF
Raw†	76.88	76.18	**79.38**	72.75	73.06	74.79	75.91	74.89	78.53	**72.97**	**73.54**	**75.46**
Lemma	**76.90**	74.68	79.25	**73.09**	58.03	**76.67**	76.02	73.82	78.58	**74.02**	68.22	**78.17**
Stem	**77.58**	**76.48**	79.26	**73.23**	59.51	**76.66**	76.59	75.17	78.13	**74.28**	68.61	**78.39**

Table 3: **Performance with BoWs**: performances exceeding the baseline are bold-faced. The baseline system is marked with †.

	Library Forum			NPS		
Feature	NB	SVM	CRF	NB	SVM	CRF
Baseline	77.58	76.48	79.26	74.28	68.61	78.39
+Distance$_{Relative}$	77.22	74.72	78.09	73.77	65.43	77.17
+Distance$_{Absolute}$	70.05	74.79	78.87	67.47	55.52	76.09
+TermCount	75.17	72.59	79.26	72.27	**71.06**	**78.93**
+UserID	68.18	48.01	79.11	64.47	**71.90**	**78.56**
+User=Host?†	77.12	75.11	**79.40**	–	–	–

Table 4: **F-score when adding Structural Features**: *Relative/Absolute* means the distance from the first utterance by relative or absolute position, *TermCount* indicates the number of words in an utterance. Features tested for Library Forum Chats only are marked with †. Results exceeding the baseline are bold-faced.

tribution of term frequencies over dialogue acts improved the performance over both datasets. From the results, we believe that term distribution information is a useful “data-independent” feature to use, compared to heuristically hand-crafted keywords.

4.5 Result 4: Interaction among Utterances

Table 6 shows the performance using utterance interactions. As baseline systems, we used *stemmed unigram* and *keywords* for library forum chats, while we used the same *stemmed unigram with POS tags* for NPS causal chats — this choice was made because keyword features improved the performance only over library forum chats. We found that this group of features did not help improve performance, in contrast to the findings of Bangalore et al. (2006) and Kim et al. (2010a). We expect this is due to similar reasons as above — i.e., although we found some degree of interaction among utterances, entanglement caused by having multiple participants meant that interactions between dialogue acts were not directly detected, even when using the CRF. Further, errors in predicted dialogue acts exacerbate the errors. However, we found that when using dialogue acts from the gold-standard data, the results for the CRF improved dramatically. Further, among the five individual features in this group, we saw that *TextUser* improved the performance slightly using CRF, since it resolves entanglement to some degree. From these observations, we conclude that utterance interaction features work well even in multi-party live chats when predicted dialogue acts are less noisy, and entanglement issues are resolved. Thus, we believe that disentanglement would be a necessary step to achieve higher performance on dialogue act classification in multi-party live chats.

5 Error Analysis

From the results above, we observed that the results over both datasets are similar. In particular, while analyzing the errors over library forum chats, we found that the majority of errors are from pairs of dialogue acts such as REQUEST → STATEMENT, STATEMENT → RESPONSE-ACK, REQUEST, RESPONSE-ACK → STATEMENT, and YES-ANSWER → RESPONSE-ACK. We noticed that REQUEST is similar to STATEMENT, except that the structure of the utterance is imperative. This could potentially be resolved by adding utterance-structure information. We also found that some terms often occur in multiple dialogue acts, e.g. *yes* in YES-ANSWER and RESPONSE-ACK. In addition, excessive use of markers such as ? and ! (even found in STATE-

	Library Forum			NPS		
Feature	NB	SVM	CRF	NB	SVM	CRF
Baseline	77.58	76.48	79.26	74.28	68.61	78.39
+Keywords$_{all}$†	**81.61**	**81.77**	**82.77**	51.27	61.35	74.77
+Keywords$_{part}$	–	–	–	**74.79**	65.28	78.00
+InfoDistribution$_{Raw}$	56.96	**80.29**	74.07	49.97	**79.39**	72.33
+InfoDistribution$_{Percent}$	**77.63**	72.73	**79.49**	**75.58**	**71.46**	**78.62**
+InfoDistribution$_{Label}$	75.09	71.47	**79.59**	67.64	51.24	78.32
+InfoDistribution$_{Raw.5}$	55.72	**80.52**	75.25	53.29	**77.52**	75.69
+InfoDistribution$_{Percent.5}$	**77.75**	73.94	**79.47**	**75.98**	**70.74**	**78.59**
+InfoDistribution$_{Label.5}$	75.27	72.21	**79.40**	68.10	59.56	78.38

Table 5: **F-score when adding Keyword Features**: *Raw/Percent* mean raw/relative term counts over the 15 labels, respectively. *Label* means the label which has the highest count. *Keywords$_{all}$* indicates the system using all keyword features described in Forsyth (2007), and *Keywords$_{part}$* is the system using only keywords available at the time of conversation. Results exceeding the baseline are bold-faced.

		Sentence			PredictLabel			GoldLabel		
Data	Feature	NB	SVM	CRF	NB	SVM	CRF	NB	SVM	CRF
Library	Baseline	81.61	81.77	82.77	–	–	–	–	–	–
	Prev1	78.92	81.67	82.03	81.24	80.59	82.58	81.05	80.74	**97.80**
	Prev2	76.67	81.46	77.62	80.06	80.44	82.66	79.91	80.61	**94.98**
	Prev3	74.87	81.12	75.42	78.75	80.61	82.64	78.41	80.63	**90.85**
	User	78.81	81.41	79.17	80.82	80.67	82.37	80.72	80.88	**85.35**
	TextUser	79.74	81.80	81.94	80.02	81.65	**82.79**	79.89	81.65	**82.79**
NPS	Baseline	74.28	68.61	78.39	–	–	–	–	–	–
	Prev1	69.73	67.82	73.15	70.24	53.59	77.79	70.11	50.71	**99.03**
	Prev2	68.48	59.40	47.96	69.55	50.57	77.42	69.51	49.73	**96.64**
	Prev3	67.47	65.98	43.67	69.24	50.35	77.07	69.05	51.06	**92.66**
	User	70.16	59.03	61.99	72.04	55.43	77.52	72.13	60.58	77.58
	TextUser	72.47	51.04	73.13	68.68	60.75	78.10	68.70	59.81	78.18

Table 6: **F-score when adding Dialogue Interaction**: *User* means the label from the previous utterance by the same author, and *TextUser* means the label from immediate utterance by user mentioned in the target utterance. *Label.G* indicates using gold-standard labels. Results exceeding the baseline are bold-faced.

MENT) caused confusion.

Tables 7 and 8 show the the performance of each label produced by *stemmed unigram, keywords, TextUserL* features. We observed that some dialogue acts, such as EXPRESSION, OPENING, THANKING, are relatively easy to detect; others, such as NO-ANSWER, REQUEST, RESPONSE-ACK are hard to predict accurately. We also noticed that the lower recall produced the lower F-score for those dialogue acts which are hard to detect.

Finally, we conducted randomized estimation to calculate whether any performance differences between methods are statistically significant (Yeh, 2000). We found that the keyword features led to statistically significant improvements over the baseline system ($p < 0.05$).

6 Conclusion

We have investigated the task of classifying dialogue acts in multi-party chats, and proposed features to automatically classify dialogue acts based on context, structure, keyword, and interactions among utterances. We found that the system using contextual and keyword features performed the best. Further, we have shown that features from structure and interactions did not perform well, unlike their effectiveness over 1-on-1 live chats in Kim et al. (2010a). Our evaluation suggests that entanglement amongst utterances from different participants caused lower performance using structural and dialogue interaction features. We thus conclude that disentangle-

	Ope.	Clo.	Bac.	Tha.	Exp.	Sta.	Req.	Res.	WhQ	YNQ	Yes	No	Don.	Oth.
Precision	89.19	80.95	89.78	97.66	99.25	81.16	60.28	69.58	83.45	89.69	84.13	33.33	100	0.00
Recall	82.50	62.20	80.48	96.81	97.30	91.79	32.44	66.88	67.44	84.74	60.23	13.33	34.62	0.00
F-score	85.71	70.34	84.87	97.23	98.26	86.15	42.18	68.20	74.60	87.15	70.20	19.05	51.43	0.00

Table 7: Results over individual dialogue acts in the Library Forum Chats: The features used are *stemmed unigram+keyword+TextUserL.*

	Acc.	Bye	Cla.	Con.	Emo.	Emp.	Gre.	nAn.	Oth.	Rej.	Sta.	Sys.	whQ.	yAn.	ynQ.
Precision	33.78	84.73	0.00	12.00	70.47	57.95	91.15	46.15	100.0	30.91	68.51	97.38	77.32	40.38	68.86
Recall	21.46	56.92	0.00	3.57	85.90	26.84	90.68	16.67	14.29	10.69	82.26	96.12	63.98	19.44	55.09
F-score	26.25	68.10	0.00	5.50	77.42	36.69	90.92	24.49	25.00	15.89	74.76	96.75	70.02	26.25	61.21

Table 8: Results over individual dialogue act in NPS Casual Chats: features used are *stemmed unigram+keyword+TextUserL.*

ment of utterances is needed to improve the accuracy of dialogue act classification—we consider this task to be important future work.

References

James Allen and Mark Core. 1997. Draft of damsl: Dialog act markup in several layers. Technical report, University of Rochester, Rochester, USA. The Multiparty Discourse Group.

Srinivas Bangalore, Giuseppe Di Fabbrizio, and Amanda Stent. 2006. Learning the structure of task-driven human-human dialogs. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 201–208, Sydney, Australia.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into speech acts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 309–316, Barcelona, Spain.

Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *Proceedings of 31st International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR08)*, pages 467–474, Singapore.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of the 46th Annual Meeting of the ACL: HLT (ACL 2008)*, pages 834–842, Columbus, USA.

Eric N. Forsyth. 2007. Improving automated lexical and discourse analysis of online chat dialog. Master's thesis, Naval Postgraduate School.

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

Daniel Jurafsky, Elizabeth Shriberg, Barbara Fox, and Traci Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pages 114–120, Montreal, Canada.

Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 2009. Extracting social meaning: identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2009)*, pages 638–646, Boulder, USA.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010a. Classifying dialogue acts in 1-to-1 live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 862–871, Boston, USA.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010b. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202, Uppsala, Sweden.

Andrew Lampert, Robert Dale, and Cecile Paris. 2008. The nature of requests and ecommitments in email messages. In *Proceedings of the AAAI 2008 Workshop on Enhanced Messaging*, pages 42–47, Chicago, USA.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 367–374, New York, USA.

Ken Samuel, Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of COLING/ACL 1998*, pages 1150–1156, Montreal, Canada.

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.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. 2011. Predicting thread discourse structure over technical web forums. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 13–25, Edinburgh, UK.

Christopher C. Werry. 1996. Linguistic and interactional features of internet relay chat. In Susan C. Herring, editor, *Computer-Mediated Communication*. John Benjamins, Amsterdam, the Netherlands.

Ian Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, USA.

Tianhao Wu, Faisal M. Khan, Todd A. Fisher, Lori A. Shuler, and William M. Pottenger. 2002. Posting act tagging using transformation-based learning. In *Proceedings of the Workshop on Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining*, Maebashi City, Japan.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 947–953, Saarbrücken, Germany.