Classifying Dialogue Acts in One-on-one Live Chats

Su Nam Kim,♠ Lawrence Cavedon♡ and Timothy Baldwin♠
♠ Dept of Computer Science and Software Engineering, University of Melbourne
♡ School of Computer Science and IT, RMIT University
sunamkim@gmail.com, lcavedon@gmail.com, tb@ldwin.net

Abstract

We explore the task of automatically classifying dialogue acts in 1-on-1 online chat forums, an increasingly popular means of providing customer service. In particular, we investigate the effectiveness of various features and machine learners for this task. While a simple bag-of-words approach provides a solid baseline, we find that adding information from dialogue structure and inter-utterance dependency provides some increase in performance; learners that account for sequential dependencies (CRFs) show the best performance. We report our results from testing using a corpus of chat dialogues derived from online shopping customer-feedback data.

1 Introduction

Recently, *live chats* have received attention due to the growing popularity of chat services and the increasing body of applications. For example, large organizations are increasingly providing support or information services through live chat. One advantage of chat-based customer service over conventional telephone-based customer service is that it becomes possible to semi-automate aspects of the interaction (e.g. conventional openings or canned responses to standard questions) without the customer being aware of it taking place, something that is not possible with speech-based dialogue systems (as synthesised speech is still easily distinguishable from natural speech). Potentially huge savings can be made by organisations providing customer help services if we can increase the degree of automation of live chat.

Given the increasing impact of live chat services, there is surprisingly little published computational linguistic research on the topic. There has been substantially more work done on dialogue and dialogue corpora, mostly in spoken dialogue (e.g. Stolcke et al. (2000)) but also multimodal dialogue systems in application areas such as telephone support service (Bangalore et al., 2006) and tutoring systems (Litman and Silliman, 2004). Spoken dialogue analysis introduces many complications related to the error inherent in current speech recognition technologies. As an instance of written dialogue, an advantage of live chats is that recognition errors are not such an issue, although the nature of language used in chat is typically ill-formed and turn-taking is complicated by the semi-asynchronous nature of the interaction (e.g. Werry (1996)).

In this paper, we investigate the task of automatic classification of *dialogue acts* in 1-on-1 live chats, focusing on "information delivery" chats since these are proving increasingly popular as part of enterprise customer-service solutions. Our main challenge is to develop effective features and classifiers for classifying aspects of 1-on-1 live chat. Much of the work on analysing dialogue acts in spoken dialogues has relied on non-lexical features, such as prosody and acoustic features (Stolcke et al., 2000; Julia and Iftekharuddin, 2008; Sridhar et al., 2009), which are not available for written dialogues. Previous dialogue-act detection for chat systems has used bags-of-words (hereafter, BoW) as features for dialogue-act detection; this simple approach has shown some promise (e.g. Bangalore et al. (2006), Louwerse and Crossley (2006) and Ivanovic (2008)). Other features such as keywords/ontologies (Purver et al., 2005; Forsyth, 2007) and lexical cues (Ang et al., 2005) have also been used for dialogue act classification.

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 862–871,
MIT, Massachusetts, USA, 9-11 October 2010. ©2010 Association for Computational Linguistics

In this paper, we first re-examine BoW features for dialogue act classification. As a baseline, we use the work of Ivanovic (2008), which explored 1-grams and 2-grams with Boolean values in 1-on-1 live chats in the MSN Online Shopping domain (this dataset is described in Section 5). Although this work achieved reasonably high performance (up to a micro-averaged F-score of around 80%), we believe that there is still room for improvement using BoW only. We extend this work by using ideas from related research such as text categorization (Debole and Sebastiani, 2003), and explore variants of BoW based on analysis of live chats, along with feature weighting. Finally, our main aim is to explore new features based on dialogue structure and dependencies between utterances[1] that can enhance the use of BoW for dialogue act classification. Our hypothesis is that, for task-oriented 1-on-1 live chats, the structure and interactions among utterances are useful in predicting future dialogue acts: for example, conversations typically start with a greeting, and questions and answers typically appear as adjacency pairs in a conversation. Therefore, we propose new features based on structural and dependency information derived from utterances (Sections 4.2 and 4.3).

2 Related Work

While there has been significant work on classifying dialogue acts, the bulk of this has been for spoken dialogue. Most such work has considered: (1) defining taxonomies of dialogue acts; (2) discovering useful features for the classification task; and (3) experimenting with different machine learning techniques. We focus here on (2) and (3); we return to (1) in Section 3.

For classifying dialogue acts in spoken dialogue, various features such as dialogue cues, speech characteristics, and n-grams have been proposed. For example, Samuel et al. (1998) utilized the characteristics of spoken dialogues and examined speaker direction, punctuation marks, cue phrases and n-grams for classifying spoken dialogues. Jurafsky et al. (1998) used prosodic, lexical and syntactic features for spoken dialogue classification. More recently, Julia and Iftekharuddin (2008) and Sridhar et al. (2009) achieved high performance using acoustic and prosodic features. Louwerse and Crossley (2006), on the other hand, used various n-gram features—which could be adapted to both spoken and written dialogue—and tested them using the Map Task Corpus (Anderson et al., 1991). Extending the discourse model used in previous work, Bangalore et al. (2006) used n-grams from the previous 1–3 utterances in order to classify dialogue acts for the target utterance.

There has been substantially less effort on classifying dialogue acts in written dialogue: Wu et al. (2002) and Forsyth (2007) have used keyword-based approaches for classifying online chats; Ivanovic (2008) tested the use of n-gram features for 1-on-1 live chats with MSN Online Shopping assistants.

Various machine learning techniques have been investigated for the dialogue classification task. Samuel et al. (1998) used transformation-based learning to classify spoken dialogues, incorporating Monte Carlo sampling for training efficiency. Stolcke et al. (2000) used Hidden Markov Models (HMMs) to account for the structure of spoken dialogues, while Wu et al. (2002) also used transformation- and rule-based approaches plus HMMs for written dialogues. Other researchers have used Bayesian based approaches, such as naive Bayes (e.g. (Grau et al., 2004; Forsyth, 2007; Ivanovic, 2008)) and Bayesian networks (e.g. (Keizer, 2001; Forsyth, 2007)). Maximum entropy (e.g. (Ivanovic, 2008)), support vector machines (e.g. (Ivanovic, 2008)), and hidden Markov models (e.g. (Bui, 2003)) have also all been applied to automatic dialogue act classification.

3 Dialogue Acts

A number of dialogue act taxonomies have been proposed, designed mainly for spoken dialogue. Many of these use the Dialogue Act Markup in Several Layers (DAMSL) scheme (Allen and Core, 1997). DAMSL was originally applied to the TRAINS corpus of (transcribed) spoken task-oriented dialogues, but various adaptations of it have since been proposed for specific types of dialogue. The Switchboard corpus (Godfrey et al., 1992) defines 42 types of dialogue acts from human-to-human telephone conversations. The HCRC Map Task corpus (Ander-

[1] An *utterance* is the smallest unit to deliver a participant's message(s) in a turn.

son et al., 1991) defines a set of 128 dialogue acts to model task-based spoken conversations.

For casual online chat dialogues, Wu et al. (2002) define 15 dialogue act tags based on previously-defined dialogue act sets (Samuel et al., 1998; Shriberg et al., 1998; Jurafsky et al., 1998; Stolcke et al., 2000). Forsyth (2007) defines 15 dialogue acts for casual online conversations, based on 16 conversations with 10,567 utterances. Ivanovic (2008) proposes 12 dialogue acts based on DAMSL for 1-on-1 online customer service chats.

Ivanovic's set of dialogue acts for chat dialogues has significant overlap with the dialogue act sets of Wu et al. (2002) and Forsyth (2007) (e.g. GREETING, EMOTION/EXPRESSION, STATEMENT, QUESTION). In our work, we re-use the set of dialogue acts proposed in Ivanovic (2008), due to our targeting the same task of 1-on-1 IM chats, and indeed experimenting over the same dataset. The definitions of the dialogue acts are provided in Table 1, along with examples.

4 Feature Selection

In this section, we describe our initial dialogue-act classification experiments using simple BoW features, and then introduce two groups of new features based on structural information and dependencies between utterances.

4.1 Bag-of-Words

n-gram-based BoW features are simple yet effective for identifying similarities between two utterances, and have been used widely in previous work on dialogue act classification for online chat dialogues (Louwerse and Crossley, 2006; Ivanovic, 2008). However, chats containing large amounts of noise such as typos and emoticons pose a greater challenge for simple BoW approaches. On the other hand, keyword-based features (Forsyth, 2007) have achieved high performance; however, keyword-based approaches are more domain-dependent. In this work, we chose to start with a BoW approach based on our observation that commercial live chat services contain relatively less noise; in particular, the commercial agent tends to use well-formed, formulaic prose.

Previously, Ivanovic (2008) explored Boolean 1-gram and 2-gram features to classify MSN Online Shopping live chats, where a user requests assistance in purchasing an item, in response to which the commercial agent asks the customer questions and makes suggestions. Ivanovic (2008) achieved solid performance over this data (around 80% F-score). While 1-grams performed well (as live chat utterances are generally shorter than, e.g., sentences in news articles), we expect 2- and 3-grams are needed to detect formulaic expressions, such as *No problem* and *You are welcome*. We would also expect a positive effect from combining n-grams due to increasing the coverage of feature words. We thus test 1-, 2- and 3-grams individually, as well as the combination of 1- and 2-grams together (i.e. 1+2-grams) and 1-, 2- and 3-grams (i.e. 1+2+3-grams); this results in five BoW sets. Also, unlike Ivanovic (2008), we test both raw words and lemmas; we expect the use of lemmas to perform better than raw words as our data is less noisy. As the feature weight, in addition to simple Boolean, we also experiment with *TF*, *TF·IDF* and *Information Gain (IG)*.

4.2 Structural Information

Our motivation for using structural information as a feature is that the location of an utterance can be a strong predictor of the dialogue act. That is, dialogues are sequenced, comprising turns (i.e. a given user is sending text), each of which is made up of one or more messages (i.e. strings sent by the user). Structured classification methods which make use of this sequential information have been applied to related tasks such as tagging semantic labels of key sentences in biomedical domains (Chung, 2009) and post labels in web forums (Kim et al., 2010).

Based on the nature of live chats, we observed that the utterance position in the chat, as well as in a turn, plays an important role when identifying its dialogue act. For example, an utterance such as *Hello* will occur at the beginning of a chat while an utterance such as *Have a nice day* will typically appear at the end. The position of utterances in a turn can also help identify the dialogue act; i.e. when there are several utterances in a turn, utterances are related to each other, and thus examining the previous utterances in the same turn can help correctly predict the target utterance. For example, the greeting (*Welcome to ..*) and question (*How may I help you?*) could occur in

Dialogue Act, Definition and Examples
CONVENTIONAL_CLOSING: Various ways of ending a conversation e.g. *Bye Bye*
CONVENTIONAL_OPENING: Greeting and other ways of starting a conversation e.g. *Hello Customer*
DOWNPLAYER: A backwards-linking label often used after THANKS to down play the contribution e.g. *You are welcome, my pleasure*
EXPRESSIVE: An acknowledgement of a previous utterance or an indication of the speaker's mood. e.g. *haha,* $: -)$ *wow*
NO_ANSWER: A backward-linking label in the form of a negative response to a YESNO-QUESTION e.g. *no, nope*
OPEN_QUESTION: A question that cannot be answered with only a *yes* or *no*. The answer is usually some form of explanation or statement. e.g. *how do I use the international version?*
REQUEST: Used to express a speaker's desire that the learner do something – either performing some action or simply waiting. e.g. *Please let me know how I can assist you on MSN Shopping today.*
RESPONSE_ACK: A backward-linking acknowledgement of the previous utterance. Used to confirm that the previous utterance was received/accepted. e.g. *Sure*
STATEMENT: Used for assertions that may state a belief or commit the speaker to doing something e.g. *I am sending you the page which will pop up in a new window on your screen.*
THANKS: Conventional thanks e.g. *Thank you for contacting us.*
YES_ANSWER: A backward-linking label in the form of an affirmative response to a YESNO-QUESTION e.g. *yes, yeah*
YESNO_QUESTION: A closed question which can be answered in the affirmative or negative. e.g. *Did you receive the page, Customer?*

Table 1: The set of dialogue acts used in this research, taken from Ivanovic (2008)

the same turn. We also noticed that identifying the utterance author can help classify the dialogue act (previously used in Ivanovic (2008)).

Based on these observations, we tested the following four structural features:

- Author information,
- Relative position in the chat,
- Author + Relative position,
- Author + Turn-relative position among utterances in a given turn.

We illustrate our structural features in Table 2, which shows an example of a 1-on-1 live chat. The participants are the agent (A) and customer (C); *Uxx* indicates an utterance (U) with ID number xx. This conversation has 42 utterances in total. The relative position is calculated by dividing the utterance number by the total number of utterances in the dialogue; the turn-relative position is calculated by dividing the utterance position by the number of utterances in that turn. For example, for utterance 4 (U4), the relative position is $\frac{4}{42}$, while its turn-relative position is $\frac{2}{3}$ since *U4* is the second utterance among *U3,4,5* that the customer makes in a single turn.

4.3 Utterance Dependency

In recent work, Kim et al. (2010) demonstrated the importance of dependencies between post labels in web forums. The authors introduced series of features based on structural dependencies among posts. They used relative position, author information and automatically predicted labels from previous post(s) as dependency features for assigning a semantic label to the current target post.

Similarly, by examining our chat corpus, we observed significant dependencies between utterances. First, 1-on-1 (i.e. agent-to-user) dialogues often contain dependencies between adjacent utterances by different authors. For example, in Table 2, when the agent asks *Is that correct?*, the expected response from the user is a *Yes* or *No*. Another example is that when the agent makes a greeting, such as *Have a nice day*, then the customer will typically respond with a greeting or closing remark, and not a *Yes* or *No*. Second, the flow of dialogues is in general cohesive, unless the topic of utterances changes dramatically (e.g. *U5: Are you still there?*, *U22: brb in 1 min* in Table 2). Third, we observed that between utterances made by the same author (either agent or user), the target utterance relies on previous utterances made by the same author, especially when

ID	Utterance
A:U1	Hello Customer, welcome to MSN Shopping.
A:U2	My name is Krishna and I am your online Shopping assistant today.
C:U3	Hello!
C:U4	I'm trying to find a sports watch.
C:U5	are you still there?
A:U6	I understand that you are looking for sports watch.
A:U7	Is that correct?
C:U8	yes, that is correct.
..	
C:U22	brb in 1 min
C:U23	Thank you for waiting
..	
A:U37	Thank you for allowing us to assist you regarding wrist watch.
A:U38	I hope you found our session today helpful.
A:U39	If you have any additional questions or you need additional information, please log in again to chat with us. We are available 24 hours a day, 7 days a week for your help.
A:U40	Thank you for contacting MSN Shopping.
A:U41	Have a nice day! Good Bye and Take Care.
C:U42	You too.

Table 2: An example of a 1-on-1 live chat, with turn and utterance structure

the agent and user repeatedly question and answer. With these observations, we checked the likelihood of dialogue act pairings between two adjacent utterances, as well as between two adjacent utterances made by the same author. Overall, we found strong co-occurrence (as measured by number of occurrences of labels across adjacency pairs) between certain pairs of dialogue acts (e.g. (YesNo_Question →Yes_Answer/No_Answer) and (Request →Yes_Answer)). Statement, on the other hand, can associate with most other dialogue acts.

Based on this, we designed the following five utterance dependency features; by combining these, we obtain 31 feature sets.

1. Dependency of utterances regardless of author
   (a) Dialogue act of previous utterance
   (b) Accumulated dialogue act(s) of previous utterances
   (c) Accumulated dialogue acts of previous utterances in a given turn
2. Dependency of utterances made by a single author
   (a) Dialogue act of previous utterance by same author; a dialogue act can be in the same turn or in the previous turn
   (b) Accumulated dialogue acts of previous utterances by same author; dialogue acts can be in the same turn or in the previous turn

To capture utterance dependency, Bangalore et al. (2006) previously used n-gram BoW features from the previous 1–3 utterances. In contrast, instead of using utterances which indirectly encode dialogue acts, we directly use the dialogue act classifications, as done in Stolcke et al. (2000). The motivation is that, due to the high performance of simple BoW features, using dialogue acts directly would capture the dependency better than indirect information from utterances, despite introducing some noise. We do not build a probabilistic model of dialogue transitions the way Stolcke et al. (2000) does, but follow an approach similar to that used in Kim et al. (2010) in using predicted dialogue act(s) labels learned in previous step(s) as a feature.

5 Experiment Setup

As stated earlier, we use the data set from Ivanovic (2008) for our experiments; it contains 1-on-1 live chats from an information delivery task. This dataset contains 8 live chats, including 542 manually-segmented utterances. The maximum and minimum number of utterances in a dialogue are 84 and 42, respectively; the maximum number of utterances in a turn is 14. The live chats were manually tagged with the 12 dialogue acts described in Section 3. The utterance distribution over the dialogue acts is described in Table 3.

For our experiments, we calculated TF, TF·IDF and IG (Information Gain) over the utterances, which were optionally lemmatized with the `morph` tool (Minnen et al., 2000). We then built a dialogue act classifier using three different machine learners: SVM-HMM (Joachims, 1998),[2] naive Bayes

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

Dialogue Act	Utterance number
CONVENTIONAL_CLOSING	15
CONVENTIONAL_OPENING	12
DOWNPLAYER	15
EXPRESSIVE	5
NO_ANSWER	12
OPEN_QUESTION	17
REQUEST	28
RESPONSE _ACK	27
STATEMENT	198
THANKS	79
YES_ANSWER	35
YESNO_QUESTION	99

Table 3: Dialogue act distribution in the corpus

Index	Learner	Ours		Ivanovic	
		Feature	Acc.	Feature	Acc.
Word	SVM	1+2+3/B	.790	1/B	.751
	NB	1/B	.673	1/B	.673
	CRF	1/IG	.839	1/B	.825
Lemma	SVM	1+2+3/IG	.777	N/A	N/A
	NB	1/B	.672	N/A	N/A
	CRF	1/B	**.862**	N/A	N/A

Table 4: Best accuracy achieved by the different learners over different feature sets and weighting methods (1 = 1-gram; 1+2+3 = 1/2/3-grams; B = Boolean; IG = information gain)

from the WEKA machine learning toolkit (Witten and Frank, 2005), and Conditional Random Fields (CRF) using CRF++.[3] Note that we chose to test CRF and SVM-HMM as previous work (e.g. (Samuel et al., 1998; Stolcke et al., 2000; Chung, 2009)) has shown the effectiveness of structured classification models on sequential dependencies. Thus, we expect similar effects with CRF and SVM-HMM. Finally, we ran 8-fold cross-validation using the feature sets described above (partitioning across the 8 sessions). All results are presented in terms of classification accuracy. The accuracy of a zero-R (i.e. majority vote) baseline is 0.36.

6 Evaluation

6.1 Testing Bag-of-Words Features

Table 4 shows the best accuracy achieved by the different learners, in combination with BoW representations and feature weighting methods. Note that the CRF learner ran using 1-grams only, as CRF++ does not accept large numbers of features. As a benchmark, we also tested the method in Ivanovic (2008) and present the best performance over words (rather than lemmas). Overall, we found using just 1-grams produced the best performance for all learners, although SVM achieved the best performance when using all three n-gram orders (i.e. 1+2+3). Since the utterances are very short, 2-grams or 3-grams alone are too sparse to be effective. Among the feature weighting methods, Boolean and IG achieved higher accuracy than TF and TF·IDF. Likewise, due to the short utterances, simple Boolean values were often the most effective. However, as IG was computed using the training data, it also achieved high performance. When comparing the learners, we found that CRF produced the best performance, due to its ability to capture inter-utterance dependencies. Finally, we confirmed that using lemmas results in higher accuracy.

n-gram	Boolean	TF	TF·IDF	IG
1	.731	.511	.517	.766
2	.603	.530	.601	.614
3	.474	.463	.472	.482
1+2	.756	.511	.522	**.777**
1+2+3	.773	.511	.528	**.777**

Table 5: Accuracy of different feature representations and weighting methods for SVM-HMM

Table 5 shows the accuracy over all feature sets; for brevity, we show this for SVM only since the pattern is similar across all learners.

6.2 Using Structural Information

In this section, we describe experiments using structural information—i.e. author and/or position—with BoWs. As with the base BoW technique, we used 1-gram lemmas with Boolean values, based on the results from Section 6.1. Table 6 shows the results: *Pos* indicates the relative position of an utterance in the whole dialogue, *Author* means author information, and *Pos*$_{turn}$ indicates the relative position of the utterance in a turn. All methods outperformed the baseline; methods that surpassed the results for the simple BoW method (for the given learner) at a

[3] http://crfpp.sourceforge.net/

Feature	Learners		
	CRF	SVM	NB
BoW	.862	.731	.672
BoW+Author	.860	.655	.649
BoW+Pos	.862	.721	.655
BoW+Pos$_{absolute}$	**.863**	.631	.524
BoW+Author+Pos	**.875**	.700	.642
BoW+Author+Pos$_{turn}$	**.871**	.651	.631

Table 6: Accuracy with structural information

level of statistical significance (based on randomised estimation, $p < 0.05$) are boldfaced.

Overall, using CRFs with Author and Position information produced better performance than using BoW alone. Clearly, the ability of CRFs to natively optimise over structural dependencies provides an advantage over other learners.

Relative position cannot of course be measured directly in an actual online application; hence Table 6 also includes the use of "absolute position" as a feature. We see that, for CRF, the absolute position feature shows an insignificant drop in accuracy as compared to the use of relative position. (However, we do see a significant drop in performance when using this feature with SVM and NB.)

6.3 Using Utterance Dependency

We next combined the inter-utterance dependency features with the BoW features. Since we use the dialogue acts directly in utterance dependency, we first experimented using gold-standard dialogue act labels. We also tested using the dialogue acts which were automatically learned in previous steps.

Table 7 shows performance using both the gold-standard and learned dialogue acts. The different features listed are as follows: *LabelList/L* indicates those corresponding to all utterances in a dialogue preceding the target utterance; *LabelPrev/P* indicates a dialogue act from a previous utterance; *LabelAuthor/A* indicates a dialogue act from a previous utterance by the same author; and *LabelPrev$_t$/LabelAuthor$_t$* indicates the previous utterance(s) and previously same-authored utterance(s) in a turn, respectively. Since the accuracy for SVM and NB using learned labels is similar to that using gold standard labels, for brevity we report the performance for CRF using learned labels only. Results that exceed the BoW accuracy at a level of statistical significance ($p < 0.05$) are boldfaced.

Features	Dialogue Acts			
	Goldstandard			**Learned**
	CRF	HMM	NB	CRF
BoW	.862	.731	.672	.862
BoW+LabelList(L)	.795	.435	.225	.803
BoW+LabelPrev(P)	**.875**	.661	.364	**.876**
BoW+LabelAuthor(A)	**.865**	.633	.559	**.865**
BoW+LabelPrev$_t$(P$_t$)	**.873**	.603	.557	**.873**
BoW+LabelAuthor$_t$(A$_t$)	.862	.587	.535	.851
BoW+L+P	.804	.428	.227	.808
BoW+L+A	.799	.404	.225	.804
BoW+L+P$_t$	.803	.413	.229	.804
BoW+L+A$_t$	.808	.408	.216	.801
BoW+P+A	**.873**	.631	.517	**.869**
BoW+P+P$_t$	**.878**	.579	.539	**.875**
BoW+P+A$_t$	**.871**	.603	.519	**.867**
BoW+A+P$_t$	.847	.594	.519	.849
BoW+A+A$_t$	**.869**	.594	.530	**.871**
BoW+P$_t$+A$_t$	**.871**	.592	.519	**.867**
BoW+L+P+A	.812	.419	.231	.804
BoW+L+P+P$_t$	.816	.423	.229	.812
BoW+L+P+A$_t$	.808	.397	.225	.806
BoW+L+A+P$_t$	.810	.388	.225	.810
BoW+L+A+A$_t$	.812	.415	.216	.801
BoW+L+P$_t$+A$_t$	.810	.375	.205	.816
BoW+P+A+P$_t$	**.875**	.602	.522	**.876**
BoW+P+A+A$_t$	.862	.609	.511	**.864**
BoW+P+P$_t$+A$_t$	**.873**	.594	.515	**.867**
BoW+A+P$_t$+A$_t$	**.865**	.594	.517	**.864**
BoW+L+P+A+P$_t$	.817	.410	.231	.810
BoW+L+P+A+A$_t$	.814	.411	.223	.810
BoW+L+P+P$_t$+A$_t$	.816	.382	.205	.806
BoW+L+A+P$_t$+A$_t$	.812	.406	.203	.808
BoW+P+A+P$_t$+A$_t$	**.865**	.583	.513	**.865**
BoW+L+P+A+P$_t$+A$_t$	.816	.399	.205	.803

Table 7: Accuracy for the different learners with dependency features

Utterance dependency features worked well in combination with CRF only. Individually, *Prev* and *Prev$_t$* (i.e. BoW+P+P$_t$) helped to achieve higher accuracies, and the *Author* feature was also beneficial. However, *List* decreased the performance, as the flow of dialogues can change, and when a larger history of dialogue acts is included, it tends to introduce noise. Comparing use of gold-standard and learned dialogue acts, the reduction in accuracy was not statistically significant, indicating that we can

Feature	CRF	SVM	NB
C+LabelList	.9557	.4613	.2565
C+LabelPrev	.9649	.6365	.5720
C+LabelAuthor	**.9686**	.6310	.5424
C+LabelPrev$_t$	**.9686**	.5738	.5738
C+LabelAuthor$_t$	.9561	.6125	.5332

Table 8: Accuracy with Structural and Dependency Information: *C* means lemmatized Unigram+Position+Author

achieve high performance on dialogue act classification even with interactively-learned dialogue acts. We believe this demonstrates the robustness of the proposed techniques.

Finally, we tested the combination of features from structural and dependency information. That is, we used a base feature (unigrams with Boolean value), relative position, author information, combined with each of the different dependency features – LabelList, LabelPrev, LabelAuthor, LabelPrev$_t$ and LabelAuthor$_t$.

Table 8 shows the performance when using these combinations, for each dependency feature. As we would expect, CRFs performed well with the combined features since CRFs can incorporate the structural and dependency information; the achieved the highest accuracy of 96.86%.

6.4 Error Analysis and Future Work

Finally, we analyzed the errors of the best-performing feature set (i.e. *BoW+Position+Author+LabelAuthor*). In Table 9, we present a confusion matrix of errors, for CONVENTIONAL_CLOSING (Cl), CONVENTIONAL_OPENING (Op), DOWNPLAYER (Dp), EXPRESSIVE (Ex), NO_ANSWER (No), OPEN_QUESTION (Qu), REQUEST (Rq), RESPONSE_ACK (Ack), STATEMENT (St), THANKS (Ta), YES_ANSWER (Yes), and YESNO_QUESTION (YN). Rows indicate the correct dialogue acts and columns indicate misclassified dialogue acts.

Looking over the data, STATEMENT is a common source of misclassification, as it is the majority class in the data. In particularly, a large number of REQUEST and RESPONSE_ACK utterances were tagged as STATEMENT. We did not include punctuation such as question marks in our feature sets; including this would likely improve results further.

In future work, we plan to investigate methods for automatically cleansing the data to remove typos, and taking account of temporal gaps that can sometimes arise in online chats (e.g. in Table 2, there is a time gap between *C:U22 brb in 1 min* and *C:U23 Thank you for waiting*).

7 Conclusion

We have explored an automated approach for classifying dialogue acts in 1-on-1 live chats in the shopping domain, using bag-of-words (BoW), structural information and utterance dependency features. We found that the BoW features perform remarkably well, with slight improvements when using lemmas rather than words. Including structural and inter-utterance dependency information further improved performance. Of the learners we experimented with, CRFs performed best, due to their ability to natively capture sequential dialogue act dependencies.

Acknowledgements

This research was supported in part by funding from Microsoft Research Asia.

References

J.Allen and M.Core. Draft of DAMSL: Dialog Act Markup in Several Layers. The Multiparty Discourse Group. University of Rochester, Rochester, USA. 1997.

A. Anderson, M. Bader, E. Bard, E. Boyle G.M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H.S. Thompson, R. and Weinert. The HCRC Map Task Corpus. *Language and Speech.* 1991, 34, pp. 351–366.

J. Ang, Y. Liu and E. Shriberg. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2005, pp, 1061–1064.

S. Bangalore, G. Di Fabbrizio and A. Stent. Learning the Structure of Task-Driven Human-Human Dialogs. *Proceedings of the 21st COLING and 44th ACL.* 2006, pp. 201–208.

H. H. Bui. A general model for online probabilistic plan recognition. *IJCAI.* 2003, pp. 1309–1318.

G.Y Chung. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*. 2009, 9(10), pp. 1–13.

F. Debole and F. Sebastiani. Supervised term weighting for automated text categorization. *18th ACM Symposium on Applied Computing*. 2003, pp. 784–788.

	Cl	Op	Dp	Ex	No	Qu	Rq	Ack	St	Ta	Yes	YN
Op	0	0	0	0	0	0	0	0	0	0	0	2
Cl	0	0	0	0	0	0	0	0	1	1	0	0
Dp	0	0	0	0	0	0	0	0	0	0	0	0
Ex	0	0	0	0	0	0	0	0	0	0	0	0
No	0	0	0	0	0	0	0	0	0	0	0	0
Qu	0	0	0	0	0	0	0	0	0	0	0	0
Rq	0	0	0	0	0	0	0	0	3	0	0	0
Ack	0	0	1	0	1	0	0	0	5	0	0	0
St	0	0	0	0	0	0	1	0	0	0	0	0
Ta	1	0	0	0	0	0	0	0	0	0	0	0
Yes	0	0	0	0	0	0	0	0	0	0	0	0
YN	0	1	0	0	0	0	0	0	0	0	0	0

Table 9: Confusion matrix for errors from the CRF with *BoW+Position+Author+LabelAuthor* (rows = correct classification; columns = misclassification; Conventional_Closing = Cl; Conventional_Opening = Op; Downplayer = Dp; Expressive = Ex; No_Answer = No; Open_Question = Qu; Request = Rq; Response_Ack = Ack; Statement = St; Thanks = Ta; Yes_Answer = Yes; and YesNo_Question = YN)

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

J. Godfrey and E. Holliman and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1992, pp. 517–520.

S. Grau, E. Sanchis, M. Jose and D. Vilar. Dialogue act classification using a Bayesian approach. *Proceedings of the 9th Conference on Speech and Computer*. 2004.

P. A. Heeman and J. Allen. The Trains 93 Dialogues. *Trains Technical Note 94-2*. Computer Science Dept., University of Rochester, March 1995.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Proceedings of European Conference on Machine Learning*. 1998, pp. 137–142.

M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker and P. Maloor. MATCH: An Architecture for Multimodal Dialogue Systems. *Proceedings of 40th ACL*. 2002, pp. 376–383.

F. N. Julia and K. M. Iftekharuddin. Dialog Act classification using acoustic and discourse information of MapTask Data. *Proceedings of the International Joint Conference on Neural Networks*. 2008, pp. 1472–1479.

D. Jurafsky, E. Shriberg, B Fox and T. Curl. Lexical, Prosodic, and Syntactic Cues for Dialog Acts. *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*. 1998, pp. 114–120.

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

S. Keizer. A Bayesian Approach to Dialogue Act Classification. *5th Workshop on Formal Semantics and Pragmatics of Dialogue*. 2001, pp. 210–218.

S.N. Kim and L. Wang and T. Baldwin. Tagging and Linking Web Forum Posts. *Fourteenth Conference on Computational Natural Language Learning*. 2010.

J. Lafferty, A. McCallum and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML*. 2001, pp. 282–289.

D. J. Litman and S. Silliman. ITSPOKE: An Intelligent Tutoring Spoken Dialogue SYstem. *Proceedings of the HLT/NAACL*. 2004.

M. M. Louwerse and S. Crossley. Dialog Act Classification Using N-Gram Algorithms. *FLAIRS Conference*, 2006, pp. 758–763.

G. Minnen, J. Carroll and D. Pearce. Applied morphological processing of English *Natural Language Engineering* 2000, 7(3), pp. 77–80.

M. Purver, J. Niekrasz and S. Peters. Ontology-Based Discourse Understanding for a Persistent Meeting Assistant. *Proc. CHI 2005 Workshop on The Virtuality Continuum Revisited*. 2005.

K. Samuel, Sandra Carberry and K. Vijay-Shanker. Dialogue Act Tagging with Transformation-Based Learning. *Proceedings of COLING/ACL 1998*. 1998, pp. 1150-1156.

E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer and C. Van

Ess-Dykema. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?. *Language and Speech*. 1998, 41(3-4), pp. 439–487.

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

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema and M. Meteer. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*. 2000, 26(3), pp. 339–373.

A. Stolcke and E. Shriberg. Markovian Combination of Language and Prosodic Models for better Speech Understanding and Recognition . Invited talk at the IEEE Workshop on Speech Recognition and Understanding, Madonna di Campiglio, Italy, December 2001 2001,

C. C. Werry. Linguistic and interactional features of Internet Relay Chat. In S. C. Herring (ed.). *Computer-Mediated Communication*. Benjamins, 1996.

I. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

T. Wu, F. M. Khan, T. A. Fisher, L. A. Shuler and W. M. Pottenger. Posting act tagging using transformation-based learning. *Proceedings of the Workshop on Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining*. 2002.