Dialogue Act Tagging for Instant Messaging Chat Sessions

Edward Ivanovic
Department of Computer Science and Software Engineering
University of Melbourne
Victoria 3010, Australia
edwardi@csse.unimelb.edu.au

Abstract

Instant Messaging chat sessions are real-time text-based conversations which can be analyzed using dialogue-act models. We describe a statistical approach for modelling and detecting dialogue acts in Instant Messaging dialogue. This involved the collection of a small set of task-based dialogues and annotating them with a revised tag set. We then dealt with segmentation and synchronisation issues which do not arise in spoken dialogue. The model we developed combines naive Bayes and dialogue-act n-grams to obtain better than 80% accuracy in our tagging experiment.

1 Introduction

Instant Messaging (IM) dialogue has received relatively little attention in discourse modelling. The novelty and popularity of IM dialogue and the significant differences between written and spoken English warrant specific research on IM dialogue. We show that IM dialogue has some unique problems and attributes not found in transcribed spoken dialogue, which has been the focus of most work in discourse modelling. The present study addresses the problems presented by these differences when modelling dialogue acts in IM dialogue.

Stolcke et al. (2000) point out that the use of dialogue acts is a useful first level of analysis for describing discourse structure. Dialogue acts are based on the illocutionary force of an utterance from speech act theory, and represent acts such as assertions and declarations (Austin, 1962; Searle, 1979). This theory has been extended in dialogue acts to model the conversational functions that utterances can perform. Dialogue acts have been used to benefit tasks such as machine translation (Tanaka and Yokoo, 1999) and the automatic detection of dialogue games (Levin et al., 1999). This deeper level of discourse understanding may help replace or assist a support representative using IM dialogue by suggesting responses that are more sophisticated and realistic to a human dialogue participant.

The unique problems and attributes exhibited by IM dialogue prohibit existing dialogue act classification methods from being applied directly. We present solutions to some of these problems along with methods to obtain high accuracy in automated dialogue act classification. A statistical discourse model is trained and then used to classify dialogue acts based on the observed words in an utterance. The training data are online conversations between two people: a customer and a shopping assistant, which we collected and manually annotated. Table 1 shows a sample of the type of dialogue and discourse structure used in this study.

We begin by considering the preliminary issues that arise in IM dialogue, why they are problematic when modelling dialogue acts, and present their solutions in §2. With the preliminary problems solved, we investigate the dialogue act labelling task with a description of our data in §3. The remainder of the paper describes our experiment involving the training of a naive Bayes model combined with a n-gram discourse model (§4). The results of this model and evaluation statistics are presented in §5. §6 contains a discussion of the approach we used including its strengths, areas of improvement, and issues for future research followed by the conclusion in §7.

Proceedings of the ACL Student Research Workshop, pages 79–84,
Ann Arbor, Michigan, June 2005.

Turn	Msg	Sec	Speaker	Message
5	8	18	Customer	[i was talking to mike and my browser crashed]$^{U_8:\text{STATEMENT}}$ - [can you transfer me to him again?]$^{U_9:\text{YES-NO-QUESTION}}$
5	9	7	Customer	[he found a gift i wanted]$^{U_{10}:\text{STATEMENT}}$
6	10	35	Sally	[I will try my best to help you find the gift,]$^{U_{11}:\text{STATEMENT}}$ [please let me know the request]$^{U_{12}:\text{REQUEST}}$
6	11	9	Sally	[Mike is not available at this point of time]$^{U_{13}:\text{STATEMENT}}$
7	12	1	Customer	[but mike already found it]$^{U_{14}:\text{STATEMENT}}$ [isn't he there?]$^{U_{15}:\text{YES-NO-QUESTION}}$
8	13	8	Customer	[it was a remote control car]$^{U_{16}:\text{STATEMENT}}$
9	14	2	Sally	[Mike is not available right now.]$^{U_{17}:\text{NO-ANSWER}}$ [I am here to assist you.]$^{U_{18}:\text{STATEMENT}}$
10	15	28	Sally	[Sure Customer,]$^{U_{19}:\text{RESPONSE-ACK}}$ [I will search for the remote control car.]$^{U_{20}:\text{STATEMENT}}$

Table 1: An example of unsynchronised messages occurring when a user prematurely assumes a turn is finished. Here, message ("Msg") 12 is actually in response to 10, not 11 since turn 6 was sent as 2 messages: 10 and 11. We use the seconds elapsed ("Sec") since the previous message as part of a method to re-synchronise messages. Utterance boundaries and their respective dialogue acts are denoted by U_n.

2 Issues in Instant Messaging Dialogue

There are several differences between IM and transcribed spoken dialogue. The dialogue act classifier described in this paper is dependent on preprocessing tasks to resolve the issues discussed in this section.

Sequences of words in textual dialogue are grouped into three levels. The first level is a Turn, consisting of at least one Message, which consists of at least one Utterance, defined as follows:

Turn: Dialogue participants normally take turns writing.

Message: A message is defined as a group of words that are sent from one dialogue participant to the other as a single unit. A single turn can span multiple messages, which sometimes leads to accidental interruptions as discussed in §2.2.

Utterance: This is the shortest unit we deal with and can be thought of as one complete semantic unit—something that has a meaning. This can be a complete sentence or as short as an emoticon (e.g. ":-)" to smile).

Several lines from one of the dialogues in our corpus are shown as an example denoted with Turn, Message, and Utterance boundaries in Table 1.

2.1 Utterance Segmentation

Because dialogue acts work at the utterance level and users send messages which may contain more than one utterance, we first need to segment the messages by detecting utterance boundaries. Messages in our data were manually labelled with one or more dialogue act depending on the number of utterances each message contained. Labelling in this fashion had the effect of also segmenting messages into utterances based on the dialogue act boundaries.

2.2 Synchronising Messages in IM Dialogue

The end of a turn is not always obvious in typed dialogue. Users often divide turns into multiple messages, usually at clause or utterance boundaries, which can result in the end of a message being mistaken as the end of that turn. This ambiguity can lead to accidental turn interruptions which cause messages to become unsynchronised. In these cases each participant tends to respond to an earlier message than the immediately previous one, making the conversation seem somewhat incoherent when read as a transcript. An example of such a case is shown in Table 1 in which Customer replied to message 10 with message 12 while Sally was still completing turn 6 with message 11. If the resulting discourse is read sequentially it would seem that the customer ignored the information provided in message 11. The time between messages shows that only 1 second elapsed between messages 11 and 12, so message 12 must in fact be in response to message 10.

Message M_i is defined to be *dependent* on message M_d if the user wrote M_i having already seen and presumably considered M_d. The importance of unsynchronised messages is that they result in the dialogue acts also being out of order, which is

problematic when using bigram or higher-order n-gram language models. Therefore, messages are re-synchronised as described in §3.2 before training and classification.

3 The Dialogue Act Labelling Task

The domain being modelled is the online shopping assistance provided as part of the MSN Shopping site. People are employed to provide live assistance via an IM medium to potential customers who need help in finding items for purchase. Several dialogues were collected using this service, which were then manually labelled with dialogue acts and used to train our statistical models.

There were 3 aims of this task: 1) to obtain a realistic corpus; 2) to define a suitable set of dialogue act tags; and 3) to manually label the corpus using the dialogue act tag set, which is then used for training the statistical models for automatic dialogue act classification.

3.1 Tag Set

We chose 12 tags by manually labelling the dialogue corpus using tags that seemed appropriate from the 42 tags used by Stolcke et al. (2000) based on the Dialog Act Markup in Several Layers (DAMSL) tag set (Core and Allen, 1997). Some tags, such as UNINTERPRETABLE and SELF-TALK, were eliminated as they are not relevant for typed dialogue. Tags that were difficult to distinguish, given the types of utterances in our corpus, were collapsed into one tag. For example, NO ANSWERS, REJECT, and NEGATIVE NON-NO ANSWERS are all represented by NO-ANSWER in our tag set.

The Kappa statistic was used to compare inter-annotator agreement normalised for chance (Siegel and Castellan, 1988). Labelling was carried out by three computational linguistics graduate students with 89% agreement resulting in a Kappa statistic of 0.87, which is a satisfactory indication that our corpus can be labelled with high reliability using our tag set (Carletta, 1996).

A complete list of the 12 dialogue acts we used is shown in Table 2 along with examples and the frequency of each dialogue act in our corpus.

Tag	Example	%
STATEMENT	I am sending you the page now	36.0
THANKING	Thank you for contacting us	14.7
YES-NO-QUESTION	Did you receive the page?	13.9
RESPONSE-ACK	Sure	7.2
REQUEST	Please let me know how I can assist	5.9
OPEN-QUESTION	how do I use the international version?	5.3
YES-ANSWER	yes, yeah	5.1
CONVENTIONAL-CLOSING	Bye Bye	2.9
NO-ANSWER	no, nope	2.5
CONVENTIONAL-OPENING	Hello Customer	2.3
EXPRESSIVE	haha, :-), grr	2.3
DOWNPLAYER	my pleasure	1.9

Table 2: The 12 dialogue act labels with examples and frequencies given as percentages of the total number of utterances in our corpus.

3.2 Re-synchronising Messages

The typing rate is used to determine message dependencies. We calculate the typing rate by $\frac{time(M_i)-time(M_d)}{length(M_i)}$, which is the elapsed time between two messages divided by the number of characters in M_i. The dependent message M_d may be the immediately preceding message such that $d = i-1$ or any earlier message where $0 < d < i$ with the first message being M_1. This algorithm is shown in Algorithm 1.

Algorithm 1 Calculate message dependency for message i

$d \leftarrow i$
repeat
  $d \leftarrow d-1$
  $typing\_rate \leftarrow \frac{time(M_i)-time(M_d)}{length(M_i)}$
until $typing\_rate < typing\_threshold$ or $d = 1$ or $speaker(M_i) = speaker(M_d)$

The $typing\_threshold$ in Algorithm 1 was calculated by taking the 90th percentile of all observed typing rates from approximately 300 messages that had their dependent messages manually labelled resulting in a value of 5 characters per second. We found that 20% of our messages were unsynchro-

nised, giving a baseline accuracy of automatically detecting message dependencies of 80% assuming that $M_d = M_{i-1}$. Using the method described, we achieved a correct dependency detection accuracy of 94.2%.

4 Training on Speech Acts

Our goal is to perform automatic dialogue act classification of the current utterance given any previous utterances and their tags. Given all available evidence E about a dialogue, the goal is to find the dialogue act sequence U with the highest posterior probability $P(U|E)$ given that evidence. To achieve this goal, we implemented a naive Bayes classifier using bag-of-words feature representation such that the most probable dialogue act $\hat{d}$ given a bag-of-words input vector $\bar{v}$ is taken to be:

$$\hat{d} = \operatorname*{argmax}_{d \in D} \frac{P(\bar{v}|d)P(d)}{P(\bar{v})} \quad (1)$$

$$P(\bar{v}|d) \approx \prod_{j=1}^{n} P(v_j|d) \quad (2)$$

$$\hat{d} = \operatorname*{argmax}_{d \in D} P(d) \prod_{j=1}^{n} P(v_j|d) \quad (3)$$

where v_j is the jth element in $\bar{v}$, D denotes the set of all dialogue acts and $P(\bar{v})$ is constant for all $d \in D$.

The use of $P(d)$ in Equation 3 assumes that dialogue acts are independent of one another. However, we intuitively know that if someone asks a YES-NO-QUESTION then the response is more likely to be a YES-ANSWER rather than, say, CONVENTIONAL-CLOSING. This intuition is reflected in the bigram transition probabilities obtained from our corpus.[1]

To capture this dialogue act relationship we trained standard n-gram models of dialogue act history with add-one smoothing for the calculation of $P(v_j|d)$. The bigram model uses the posterior probability $P(d|H)$ rather than the prior probability $P(d)$ in Equation 3, where H is the n-gram context vector containing the previous dialogue act or previous 2 dialogue acts in the case of the trigram model.

[1]Due to space constraints, the dialogue act transition table has been omitted from this paper and is made available at http://www.cs.mu.oz.au/∼edwardi/papers/da_transitions.html

Model	Min	Max	Mean	Hit %	Px
Baseline	—	—	36.0%	—	—
Likelihood	72.3%	90.5%	80.1%	—	—
Unigram	74.7%	90.5%	80.6%	100	7.7
Bigram	75.0%	92.4%	81.6%	97	4.7
Trigram	69.5%	94.1%	80.9%	88	3.3

Table 3: Mean accuracy of labelling utterances with dialogue acts using n-gram models. Shown with hit-rate results and perplexities ("Px")

5 Experimental Results

Evaluation of the results was conducted via 9-fold cross-validation across the 9 dialogues in our corpus using 8 dialogues for training and 1 for testing. Table 3 shows the results of running the experiment with various models replacing the prior probability, $P(d)$, in Equation 3. The Min, Max, and Mean columns are obtained from the cross-validation technique used for evaluation. The baseline used for this task was to assign the most frequently observed dialogue act to each utterance, namely, STATEMENT.

Omitting $P(d)$ from Equation 3 such that only the likelihood (Equation 2) of the naive Bayes formula is used resulted in a mean accuracy of 80.1%. The high accuracy obtained with only the likelihood reflects the high dependency between dialogue acts and the actual words used in utterances. This dependency is represented well by the bag-of-words approach. Using $P(d)$ to arrive at Equation 3 yields a slight increase in accuracy to 80.6%.

The bigram model obtains the best result with 81.6% accuracy. This result is due to more accurate predictions with $P(d|H)$. The trigram model produced a slightly lower accuracy rate, partly due to a lack of training data and to dialogue act adjacency pairs not being dependent on dialogue acts further removed as discussed in §4.

In order to gauge the effectiveness of the bigram and trigram models in view of the small amount of training data, hit-rate statistics were collected during testing. These statistics, presented in Table 3, show the percentage of conditions that existed in the various models. Conditions that did not exist were not counted in the accuracy measure during evaluation.

The perplexities (Cover and Thomas, 1991) for the various n-gram models we used are shown in

Table 3. The biggest improvement, indicated by a decreased perplexity, comes when moving from the unigram to bigram models as expected. However, the large difference between the bigram and trigram models is somewhat unexpected given the theory of adjacency pairs. This may be a result of insufficient training data as would be suggested by the lower trigram hit rate.

6 Discussion and Future Research

As indicated by the Kappa statistics in §3.1, labelling utterances with dialogue acts can sometimes be a subjective task. Moreover, there are many possible tag sets to choose from. These two factors make it difficult to accurately compare various tagging methods and is one reason why Kappa statistics and perplexity measures are useful. The work presented in this paper shows that using even the relatively simple bag-of-words approach with a naive Bayes classifier can produce very good results.

One important area not tackled by this experiment was that of utterance boundary detection. Multiple utterances are often sent in one message, sometimes in one sentence, and each utterance must be tagged. Approximately 40% of the messages in our corpus have more than one utterance per message. Utterances were manually marked in this experiment as the study was focussed only on dialogue act classification given a sequence of utterances. It is rare, however, to be given text that is already segmented into utterances, so some work will be required to accomplish this segmentation before automated dialogue act tagging can commence. Therefore, utterance boundary detection is an important area for further research.

The methods used to detect dialogue acts presented here do not take into account sentential structure. The sentences in (1) would thus be treated equally with the bag-of-words approach.

(1) a. john has been to london
b. has john been to london

Without the punctuation (as is often the case with informal typed dialogue) the bag-of-words approach will not differentiate the sentences, whereas if we look at the ordering of even the first two words we can see that "john has ..." is likely to be a STATEMENT whereas "has john ..." would be a question. It would be interesting to research other types of features such as phrase structure or even looking at the order of the first x words and the parts of speech of an utterance to determine its dialogue act.

Aspects of dialogue macrogame theory (DMT) (Mann, 2002) may help to increase tagging accuracy. In DMT, sets of utterances are grouped together to form a *game*. Games may be nested as in the following example:

A: May I know the price range please?
B: In which currency?
A: $US please
B: 200–300

Here, B has nested a clarification question which was required before providing the price range. The bigram model presented in this paper will incorrectly capture this interaction as the sequence YES-NO-QUESTION, OPEN-QUESTION, STATEMENT, STATEMENT, whereas DMT would be able to extract the nested question resulting in the correct pairs of question and answer sequences.

Although other studies have attempted to automatically tag utterances with dialogue acts (Stolcke et al., 2000; Jurafsky et al., 1997; Kita et al., 1996) it is difficult to fairly compare results because the data was significantly different (transcribed spoken dialogue versus typed dialogue) and the dialogue acts were also different ranging from a set of 9 (Kita et al., 1996) to 42 (Stolcke et al., 2000). It may be possible to use a standard set of dialogue acts for a particular domain, but inventing a set that could be used for all domains seems unlikely. This is primarily due to differing needs in various applications. A superset of dialogue acts that covers all domains would necessarily be a large number of tags (at least the 42 identified by Stolcke et al. (2000)) with many tags not being appropriate for other domains.

The best result from our dialogue act classifier was obtained using a bigram discourse model resulting in an average tagging accuracy of 81.6% (see Table 3). Although this is higher than the results from 13 recent studies presented by Stolcke et al. (2000) with accuracy ranging from $\approx 40\%$ to 81.2%, the tasks, data, and tag sets used were all quite different, so any comparison should be used as only a guideline.

7 Conclusion

In this paper, we have highlighted some unique characteristics in IM dialogue that are not found in transcribed spoken dialogue or other forms of written dialogue such as e-mail; namely, utterance segmentation and message synchronisation. We showed the problem of unsynchronised messages can be readily solved using a simple technique utilising the typing-rate and time stamps of messages. We described a method for high-accuracy dialogue act classification, which is an essential part for a deeper understanding of dialogue. In our experiments, the bigram model performed with the highest tagging accuracy which indicates that dialogue acts often occur as adjacency pairs. We also saw that the high tagging accuracy results obtained by the likelihood from the naive Bayes model indicated the high correlation between the actual words and dialogue acts. The Kappa statistics we calculated indicate that our tag set can be used reliably for annotation tasks.

The increasing popularity of IM and automated agent-based support services is ripe with new challenges for research and development. For example, IM provides the ability for an automated agent to ask clarification questions. Appropriate dialogue modelling will enable the automated agent to reliably distinguish questions from statements. More generally, the rapidly expanding scope of online support services provides the impetus for IM dialogue systems and discourse models to be developed further. Our findings have demonstrated the potential for dialogue modelling for IM chat sessions, and opens the way for a comprehensive investigation of this new application area.

Acknowledgments

We thank Steven Bird, Timothy Baldwin, Trevor Cohn, and the anonymous reviewers for their helpful and constructive comments on this paper. We also thank Vanessa Smith, Patrick Ye, and Jeremy Nicholson for annotating the data.

References

John L. Austin. 1962. *How to do Things with Words*. Clarendon Press, Oxford.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Mark Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley, New York.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. *Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding*, pages 88–95.

Kenji Kita, Yoshikazu Fukui, Masaaki Nagata, and Tsuyoshi Morimoto. 1996. Automatic acquisition of probabilistic dialogue models. *Proceedings of the Fourth International Conference on Spoken Language*, 1:196–199.

Lori Levin, Klaus Ries, Ann Thyme-Gobbel, and Alon Lavie. 1999. Tagging of speech acts and dialogue games in spanish call home. *Towards Standards and Tools for Discourse Tagging (Proceedings of the ACL Workshop at ACL'99)*, pages 42–47.

William Mann. 2002. Dialogue macrogame theory. *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*, pages 129–141.

John R. Searle. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, Cambridge, UK.

Sidney Siegel and N. John Castellan, Jr. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, second edition.

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

Hideki Tanaka and Akio Yokoo. 1999. An efficient statistical speech act type tagging system for speech translation systems. In *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 381–388. Association for Computational Linguistics.