Dialogue Act Recognition in Synchronous and Asynchronous Conversations

Maryam Tavafi[†], Yashar Mehdad[†], Shafiq Joty[‡], Giuseppe Carenini[†], Raymond Ng[†]
[†]Department of Computer Science, University of British Columbia, Vancouver, Canada
[‡]Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar
[†]{tavafi, mehdad, carenini, rng}@cs.ubc.ca [‡]sjoty@qf.org.qa

Abstract

In this work, we study the effectiveness of state-of-the-art, sophisticated supervised learning algorithms for dialogue act modeling across a comprehensive set of different spoken and written conversations including: emails, forums, meetings, and phone conversations. To this aim, we compare the results of SVM-multiclass and two structured predictors namely SVM-hmm and CRF algorithms. Extensive empirical results, across different conversational modalities, demonstrate the effectiveness of our SVM-hmm model for dialogue act recognition in conversations.

1 Introduction

Revealing the underlying conversational structure in dialogues is important for detecting the human social intentions in spoken conversations and in many applications including summarization (Murray, 2010), dialogue systems and dialogue games (Carlson, 1983) and flirt detection (Ranganath, 2009). As an additional example, Ravi and Kim (2007) show that dialogue acts can be used for analyzing the interaction of students in educational forums.

Recently, there have been increasing interests for dialogue act (DA) recognition in spoken and written conversations, which include meetings, phone conversations, emails and blogs. However, most of the previous works are specific to one of these domains. There are potentially useful features and algorithms for each of these domains, but due to the underlying similarities between these types of conversations, we aim to identify a domain-independent DA modeling approach that can achieve good results across all types of conversations. Such a domain-independent dialogue act recognizer makes it possible to automatically recognize dialogue acts in a wide variety of conversational data, as well as in conversations spanning multiple domains/modalities; for instance a conversation that starts in a meeting and then continues via email.

While previous work in DA modeling has focused on studying only one (Carvalho, 2005; Shrestha, 2004; Ravi, 2007; Ferschke, 2012; Kim, 2010a; Sun, 2012) or, in a few cases, a couple of conversational domains (Jeong, 2009; Joty, 2011), in this paper, we analyze the performance of supervised DA modeling on a comprehensive set of different spoken and written conversations that includes: emails, forums, meetings, and phone conversations. More specifically, we compare the performance of three state-of-the-art, sophisticated machine learning algorithms, which include SVM-multiclass and two structured predictors SVM-hmm and Conditional Random Fields (CRF) for DA modeling. We present an extensive set of experiments studying the effectiveness of DA modeling on different types of conversations such as emails, forums, meeting, and phone discussions. The experimental results show that the SVM-hmm algorithm outperforms other supervised algorithms across all datasets.

2 Related Work

There have been several studies on supervised dialogue act (DA) modeling. To the best of our knowledge, none of them compare the performance of DA recognition on different synchronous (e.g., meeting and phone) and asynchronous (e.g., email and forum) conversations. Most of the works analyze DA modeling in a specific domain. Carvalho and Cohen (2005) propose classifying emails into their dialogue acts according to two ontologies for nouns and verbs. The ontologies are used for determining the speech acts of each single email with verb-noun pairs. Shrestha and McKeown (2004) also study the

Proceedings of the SIGDIAL 2013 Conference, pages 117–121,
Metz, France, 22-24 August 2013.

problem of DA modeling in email conversations considering the two dialogue acts of *question* and *answer*. Likewise, Ravi and Kin (2007) present a DA recognition method for detecting questions and answers in educational discussions. Ferschke et al. (2012) apply DA modeling to Wikipedia discussions to analyze the collaborative process of editing Wikipedia pages. Kim et al. (2010a) study the task of supervised classification of dialogue acts in one-to-one online chats in the shopping domain.

All these previous studies focus on DA recognition in one or two domains, and do not systematically analyze the performance of different dialog act modeling approaches on a comprehensive set of conversation domains. As far as we know, the present work is the first that proposes domain-independent supervised DA modeling techniques, and analyzes their effectiveness on different modalities of conversations.

3 Dialogue Act Recognition

3.1 Conversational structure

Adjacent utterances in a conversation have a strong correlation in terms of their dialogue acts. As an example, if speaker 1 asks a question to speaker 2, it is a high probability that the next utterance of the conversation would be an answer from speaker 2. Therefore, the conversational structure is a paramount factor that should be taken into account for automatic DA modeling. The conversational structure differs in spoken and written discussions. In spoken conversations, the discussion between the speakers is synchronized. The speakers hear each other's ideas and then state their opinions. So the temporal order of the utterances can be considered as the conversational structure in these types of conversations. However, in written conversations such as email and forum, authors contribute to the discussion in different order, and sometimes they do not pay attention to the content of previous posts. Therefore, the temporal order of the conversation cannot be used as the conversational structure in these domains, and appropriate techniques should be used to extract the underlying structure in these conversations.

To this aim, when reply links are available in the dataset, we use them to capture the conversation structure. To obtain a conversational structure that is often even more refined than the reply links, we build the Fragment Quotation Graph. To this end, we follow the procedure proposed by Joty et al. (2011) to extract the graph structure of a thread.

3.2 Features

In defining the feature set, we have two primary criteria, being domain independent and effectiveness in previous works. Lexical features such as unigrams and bigrams have been shown to be useful for the task of DA modeling in previous studies (Sun, 2012; Ferschke, 2012; Kim, 2010a; Ravi, 2007; Carvalho, 2005). In addition, unigrams have been shown to be the most effective among the two. So, as the lexical feature, we include the frequency of unigrams in our feature set.

Moreover, length of the utterance is another beneficial feature for DA recognition (Ferschke, 2012; Shrestha, 2004; Joty, 2011), which we add to our feature set. The speaker of an utterance has shown its utility for recognizing speech acts (Sun, 2012; Kim, 2010a; Joty, 2011). Sun and Morency (2012) specifically employ a speaker-adaptation technique to demonstrate the effectiveness of this feature for DA modeling. We also include the relative position of a sentence in a post for DA modeling since most of previous studies (Ferschke, 2012; Kim, 2010a; Joty, 2011) prove the efficiency of this feature.

3.3 Algorithms

Since most top performing DA models use supervised approaches (Carvalho, 2005; Shrestha, 2004; Ravi, 2007; Ferschke, 2012; Kim, 2010a), to analyze the performance of DA modeling on a comprehensive set of different spoken and written conversations, we compare the state-of-the-art supervised algorithms.

We employ three state-of-the-art, sophisticated supervised learning algorithms:

SVM-hmm predicts labels for the examples in a sequence (Tsochantaridis, 2004). This approach uses the Viterbi algorithm to find the highest scoring tag sequence for a given observation sequence. Being a Hidden Markov Model (HMM), the model makes the Markov assumption, which means that the label of a particular example is assigned only by considering the label of the previous example. This approach is considered an SVM because the parameters of the model are trained discriminatively to separate the label of sequences by a large margin.

CRF is a probabilistic framework to label and segment sequence data (Lafferty, 2001). The main advantage of CRF over HMM is that it relaxes the assumption of conditional independence of observed data. HMM is a generative model that assigns a joint distribution over label and observation sequences. Whereas, CRF defines the conditional probability distribution over label sequences given a particular observation sequence.
SVM-multiclass is a generalization of binary SVM to a multiclass predictor (Crammer, 2001). The SVM-multiclass does not consider the sequential dependency between the examples.

4 Corpora

Gathering conversational corpora for DA modeling is an expensive and time-consuming task. Due to the privacy issues, there are few available conversational datasets.

For asynchronous conversations, we use available corpora for email and forum discussions. For synchronous domains we employ available corpora in multi-party meeting and phone conversations.

BC3 (Email): As the labeled dataset for email conversations, we use BC3 (Ulrich, 2008), which contains 40 threads from W3C corpus. The BC3 corpus is annotated with twelve domain-independent dialogue acts, which are mainly adopted from the MRDA tagset, and it has been used in several previous works (e.g., (Joty, 2011)).

CNET (Forum): As the labeled forum dataset, we use the available CNET corpus, which is annotated with eleven domain-independent dialogue acts in a post-level (Kim et al, 2010b). This corpus consists of 320 threads and a total of 1332 posts, which are mostly from technical forums.

MRDA (Meeting): ICSI-MRDA dataset is used as labeled data for meeting conversation, which contains 75 meetings with 53 unique speakers (Shriberg, 2004). The ICSI-MRDA dataset requires one general tag per sentence followed by variable number of specific tags. There are 11 general tags and 39 specific tags in the annotation scheme. We reduce their tagset to the eleven general tags to be consistent with the other datasets.

SWBD (Phone): In addition to multi-party meeting conversations, we also report our experimental results on Switchboard-DAMSL (SWBD), which is a large-scale corpus containing telephone speech (Jurafsky, 1997). This corpus is annotated with the SWBD-DAMSL tagset, which consists of 220 tags. We use the mapping table presented by Jeong (2009) to reduce the tagset to 16 domain-independent dialogue acts.

All the available corpora are annotated with dialogue acts at the sentence-level. The only exception is the CNET forum dataset, on which we apply DA classification at the post-level.

5 Experiments and Results

5.1 Experimental settings

In our experiments, we use the SVM-hmm[1] and SVM-multiclass[2] packages developed with the SVM-light software. We use the Mallet package[3] for the CRF algorithm. The results of supervised classifications are compared to the baseline, which is the majority class of each dataset. We apply 5-fold cross-validation for the supervised learning methods to each dataset, and compare the results of different methods using micro-averaged and macro-averaged accuracies.

5.2 Results

Table 1 shows the results of supervised classification on different conversation modalities. We observe that SVM-hmm and CRF classifiers outperform SVM-multiclass classifier in all conversational domains. Both SVM-hmm and CRF classifiers consider the sequential structure of conversations, while this is ignored in the SVM-multiclass classifier. This shows that the sequential structure of the conversation is beneficial independently of the conversational modality. We can also observe that the SVM-hmm algorithm results in the highest performance in all datasets. As shown in (Altun, 2003), generalization performace of SVM-hmm is superior to CRF. This superiority also applies to the DA modeling task across all the conversational modalities. However, as it was investigated by Keerthi and Sundararajan (2007), the discrepancy in the performance of these methods may arise from different feature functions that these two methods use, and they might perform similarly when they use the same feature functions.

Comparing the results across different datasets, we can also note that the largest improvement of SVM-hmm and CRF is on the SWBD, the

[1] http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html
[2] http://svmlight.joachims.org/svm_multiclass.html
[3] http://mallet.cs.umass.edu

Corpus	Baseline		SVM-multiclass		SVM-hmm		CRF	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
BC3	69.56	8.34	73.57 (4.01)	8.34 (0)	**77.75 (8.19)**	**18.20 (9.86)**	72.18 (2.62)	14.9 (6.56)
CNET	36.75	9.09	34.8 (-1.95)	9.3 (0.21)	**58.7 (21.95)**	**17.1 (8.01)**	40.3 (3.55)	11.5 (2.41)
MRDA	66.47	9.09	66.47 (0)	9.09 (0)	**80.5 (14.03)**	**32.4 (23.31)**	77.8 (11.33)	22.9 (13.81)
SWBD	46.44	6.25	46.5 (0.06)	6.25 (0)	**74.32 (27.88)**	**30.13 (23.88)**	73.04 (26.6)	24.05 (17.8)

Table 1: Results of supervised DA modeling; columns are micro-averaged and macro-averaged accuracies with difference with baseline in parentheses.

phone conversation dataset. Moreover, supervised DA recognition on synchronous conversations achieves a better performance than on asynchronous conversations. We can argue that this is due to the less complex sequential structure of synchronous conversations. A lower macro-averaged accuracy in asynchronous conversations (i.e., forums and emails) can be justified in the same way.

By looking at the results in asynchronous conversations, we observe a larger improvement of micro-averaged accuracy over the CNET corpus. This might be due to two reasons: *i)* the DA tagsets in both corpora are different (i.e., no overlap in tagsets); and *ii)* the conversational structure in forums and emails is different.

5.3 Discussion

We analyze the strengths and weakness of supervised DA modeling with SVM-hmm in different conversations individually.

BC3: SVM-hmm succeeds in classifying most of the *statement* and *yes-no question* speech acts in the BC3 corpus. However, it does not show a high accuracy for classifying *polite mechanisms* such as 'thanks' and 'regards'. Through the error analysis, we observed that in most of these cases the error arose from the voting algorithm. Moreover, the improvement of supervised DA modeling on the BC3 corpus is smaller than the other datasets. This may suggest that email conversation is a challenging domain for DA recognition.

CNET: The inventory of dialogue acts in the CNET dataset can be considered as two groups of *question* and *answer* dialogue acts, and we would need more sophisticated features in order to classify the posts into the fine-grained dialogue acts. The SVM-hmm succeeds in predicting the labels of *question-question* and *answer-answer* dialogue acts, but it performs poorly for the other labels. The improvement of DA modeling over the baseline is significant for this dataset. To further improve the performance, a hierarchical DA classification can be applied. In this way, the posts would be classified into *question* and *non-question* dialogue acts in the first level.

MRDA: SVM-hmm performs well for predicting the classes of *statement*, *floor holder*, *backchannel*, and *wh-question*. *Floor holders* and *backchannels* are mostly the short utterances such as 'ok', 'um', and 'so', and we believe the length and unigrams features are very effective for predicting these dialogue acts. On the other hand, SVM-hmm fails in predicting the other types of questions such as *rhetorical questions* and *open-ended questions* by classifying them as *statements*. Arguably by adding more sophisticated features such as POS tags, SVM-hmm would perform better for classifying these speech acts.

SWBD: The improvement of supervised DA recognition on the SWBD is higher than the other domains. Supervised DA classification correctly predicts most of the classes of *statement*, *reject response*, *wh-question*, and *backchannel*. However, SVM-hmm cannot predict some specific dialogue acts of phone conversations such as *self-talk* and *signal-non-understanding*. There are a few utterances in the corpus with these dialogue acts, and most of them are classified as *statements*.

6 Conclusion and Future Work

We have studied the effectiveness of sophisticated supervised learning algorithms for DA modeling across a comprehensive set of different spoken and written conversations. Through an extensive experiment, we have shown that our proposed SVM-hmm algorithm with the domain-independent feature set can achieve high results on different synchronous and asynchronous conversations.

In future, we will incorporate other lexical and syntactic features in our supervised framework. We also plan to augment our feature set with domain-specific features like prosodic features for spoken conversations. We will also investigate the performance of our domain-independent approach in a semi-supervised framework.

References

Congkai Sun and Louise-Philippe Morency. 2012. *Dialogue Act Recognition using Reweighted Speaker Adaptation*. 13th Annual SIGdial Meeting on Discourse and Dialogue.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. *Switchboard SWBD-DAMSL labeling project coder's manual, draft 13*. Technical report, Univ. of Colorado Institute of Cognitive Science.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. *The ICSI Meeting Recorder Dialog Act (MRDA) Corpus*. HLT-NAACL SIGDIAL Workshop.

Gabriel Murray, Giuseppe Carenini, and Raymond T. Ng. 2010. *Generating and validating abstracts of meeting conversations: a user study*. INLG'10.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. *Support vector machine learning for interdependent and structured output spaces.* Proceedings of the 21st International Conference on Machine Learning (ICML).

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. *A publicly available annotated corpus for supervised email summarization*. EMAIL'08 Workshop. AAAI.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Intl. Conf. on Machine Learning.

Koby Crammer and Yoram Singer. 2001. *On the algorithmic implementation of multiclass kernel-based vector machines*. Journal of Machine Learning Research.

Lari Carlson. 1983. *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel.

Lokesh Shrestha and Kathleen McKeown. 2004. *Detection of question-answer pairs in email conversations*. Proceedings of the 20th Biennial Int. Conf. on Computational Linguistics.

Minwoo Jeong, Chin-Yew Lin, and Gary G. Lee. 2009. *The Semi-supervised speech act recognition in emails and forums*. Proceedings of the 2009 Conf. Empirical Methods in Natural Language Processing.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. *Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages*. Proceedings of the 13th Conference of the European Chapter of the ACL.

Rajesh Ranganath, Dan Jurafsky, and Dan Mcfarland. 2009. *Its not you, its me: Detecting flirting and its misperception in speed-dates*. EMNLP-09.

S. S. Keerthi and S. Sundararajan. 2007. *CRF versus SVM-Struct for sequence labeling*. Technical report, Yahoo Research.

Shafiq R. Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. *Unsupervised modeling of dialog acts in asynchronous conversations*. IJCAI.

Su N. Kim, Lawrence Cavedon, and Timothy Baldwin. 2010a. *Classifying dialogue acts in one-on-one live chats*. EMNLP'10.

Su N. Kim, Li Wang, and Timothy Baldwin. 2010b. *Tagging and linking web forum posts*. Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL '10.

Sujith Ravi and Jihie Kim. 2007. *Profiling student interactions in threaded discussions with speech act classifiers*. AIED'07, LA, USA.

Vitor R. Carvalho and William W. Cohen. 2005. *On the collective classification of email "speech acts"*. Proceedings of the 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval.

Yasemin Altun and Ioannis Tsochantaridis and Thomas Hofmann. 2003. *Hidden Markov Support Vector Machines*. Proceedings of the 20th International Conference on Machine Learning.

7 Appendix A. Frequency of Dialogue Acts in the Corpora

Tag	Dialogue Acts	Email (BC3)	Forum (CNET)	Meeting (MRDA)	Phone (SWBD)
A	Accept response	2.07%	–	–	6.96%
AA	Acknowledge and appreciate	1.24%	–	–	2.12%
AC	Action motivator	6.09%	–	–	0.38%
P	Polite mechanism	6.97%	–	–	0.12%
QH	Rhetorical question	0.75%	–	0.34%	0.25%
QO	Open-ended question	1.32%	–	0.17%	0.3%
QR	Or/or-clause question	1.10%	–	–	0.2%
QW	Wh-question	2.29%	–	1.63%	0.95%
QY	Yes-no question	6.75%	–	4.75%	2.62%
R	Reject response	1.06%	–	–	1.03%
S	Statement	**69.56%**	–	**66.47%**	**46.44%**
U	Uncertain response	0.79%	–	–	0.15%
Z	Hedge	–	–	–	11.55%
B	Backchannel	–	–	14.44%	26.62%
D	Self-talk	–	–	–	0.1%
C	Signal-non-understanding	–	–	–	0.14%
FH	Floor holder	–	–	7.96%	–
FG	Floor grabber	–	–	2.96%	–
H	Hold	–	–	0.76%	–
QRR	Or clause after yes-no question	–	–	0.38%	–
QR	Or question	–	–	0.2%	–
QQ	Question-question	–	27.92%	–	–
QA	Question-add	–	11.67%	–	–
QCN	Question-confirmation	–	3.89%	–	–
QCC	Question-correction	–	0.36%	–	–
AA	Answer-answer	–	**36.75%**	–	–
AD	Answer-add	–	8.84%	–	–
AC	Answer-confirmation	–	0.36%	–	–
RP	Reproduction	–	0.71%	–	–
AO	Answer-objection	–	1.07%	–	–
RS	Resolution	–	7.78%	–	–
O	Other	–	0.71%	–	–

Table 2: Dialogue act categories and their relative frequency.

Table 2 indicates the dialogue acts of each corpus and their relative frequencies in that dataset. The table shows that the distribution of dialogue acts in the datasets are not balanced. Most of the utterances in the datasets are labeled as *statements*. Consequently, during the classification step, most of the utterances are labeled as the *statement* dialogue act. This always affects the performance of a classifier in dealing with low frequency classes. A possible approach to tackle this problem is to cluster the correlative dialogue acts into the same group and apply a DA modeling approach in a hierarchical manner.