Extractive Summarization and Dialogue Act Modeling on Email Threads: An Integrated Probabilistic Approach

Tatsuro Oya and **Giuseppe Carenini**
Department of Computer Science
University of British Columbia
Vancouver, B.C. Canada
{toya, carenini}@cs.ubc.ca

Abstract

In this paper, we present a novel supervised approach to the problem of summarizing email conversations and modeling dialogue acts. We assume that there is a relationship between dialogue acts and important sentences. Based on this assumption, we introduce a sequential graphical model approach which simultaneously summarizes email conversation and models dialogue acts. We compare our model with sequential and non-sequential models, which independently conduct the tasks of extractive summarization and dialogue act modeling. An empirical evaluation shows that our approach significantly outperforms all baselines in classifying correct summary sentences without losing performance on dialogue act modeling task.

1 Introduction

Nowadays, an overwhelming amount of text information can be found on the web. Most of this information is redundant and thus the task of document summarization has attracted much attention. Since emails in particular are used for a wide variety of purposes, the process of automatically summarizing emails might be of great benefit in dealing with this excessive amount of information. Much work has already been conducted on email summarization. The first research on this topic was conducted by Rambow *et al*. (2004), who took a supervised learning approach to extracting important sentences. A study on the supervised summarization of email threads was also performed by Ulrich *et al*. (2009). This study used the regression-based method for classification. There have been studies on unsupervised summarization of email threads as well. Zhou *et al*. (2007, 2008) proposed a graph-based unsupervised approach to email conversation summarization using clue words, i.e., recurring words contained in replies.

In addition, the task of labeling sentences with dialogue acts has become important and has been employed in many conversation analysis systems. For example, applications such as meeting summarization and collaborative task learning agents use dialogue acts as their underlying structure (Allen *et al*., 2007; Murray *et al*., 2010). In a previous work, Cohen *et al*. (2004) defined a set of "email acts" and employed text classification methods to detect these acts in emails. Later, Carvalho *et al*. (2006) employed a combination of n-gram sequences as features and then used a supervised machine learning method to improve the accuracy of this email act classification. In addition, Shafiq *et al*. (2011) presented unsupervised dialogue act labeling methods. In their work, they introduced a graph-based method and two probabilistic sequence-labeling methods for modeling dialogue acts.

However, little work has been done on discovering the relationship between dialogue acts and extractive summaries. If there is a relationship between them, combining these approaches so as to model both simultaneously will yield better results. In this paper, we investigate this hypothesis by introducing a new sequential graphical model approach that performs dialogue act modeling and extractive summarization jointly on email threads.

2 Related Work

While email summarization and dialogue act modeling have been effectively studied, in most previous work, these tasks were studied independently. This section provides related work for each task separately.

Proceedings of the SIGDIAL 2014 Conference, pages 133–140,
Philadelphia, U.S.A., 18-20 June 2014.

2.1 Extractive Summarization

Rambow *et al*. (2004) introduced sentence extraction techniques that work for email threads. In their work, they introduced email-specific features and used a machine learning method to classify whether or not a sentence should be incorporated into a summary. Their experiments demonstrated that their features were highly effective for email summarization.

Ulrich *et al*. (2009) proposed a regression-based machine learning approaches to email thread summarization. They compared regression-based classifiers to binary classifiers and showed that their approach significantly improves the summarization accuracy. They employed the feature set introduced by Rambow *et al*. (2004) as their baseline and introduced new features that are also effective for email summarization. Some of their features refer to dialogue acts but the assumption is that they are computed before the summarization task is performed. Our work is aimed at a much closer integration of the two tasks by modeling them simultaneously.

Carenini *et al*. (2007) developed a fragment quotation graph that can capture a fine-grain conversation structure in email threads, which we will describe in detail in Section 3. They then introduced a ClueWordSummarizer (CWS), a graph-based unsupervised summarization approach based on the concept of clue words, which are recurring words found in email replies. Their experiment showed that the CWS performs better than the email summarization approach in Rambow *et al*. (2004).

Extractive summarization using a sequential labeling technique has also been studied. While this is not an email summarization, Shen *et al*. (2007) proposed a linear-chain Conditional Random Field (CRF) based approach for extractive document summarization. In their work, they treated the summarization task as a sequence labeling problem to take advantage of interaction relationships between sentences; their approach showed significant improvement when compared with non-sequential classifiers.

2.2 Dialogue Act Modeling

The first studies on the dialogue act modeling in emails were performed by Cohen *et al*. (2004). They defined "email speech acts" (e.g., Request, Deliver, Propose, and Commit) and used machine learning methods to classify emails according to the intent of the sender.

Carvalho *et al*. (2006) further developed this initial proposal by using contextual information such as combinations of n-gram sequences in emails as their features for a supervised learning approach. The experiment showed that their approach reduced classification error rates by 26.4%. Shafiq *et al*. (2011) proposed unsupervised dialogue act modeling in email threads and on forums. They introduced a graph-based and two probabilistic unsupervised approaches for modeling dialogue acts. By comparing those approaches, they demonstrated that the probabilistic approaches were quite effective and performed better than the graph-based one.

While the following work is not done on the email domain, Kim *et al*. (2010) introduced a dialogue act classification on one-on-one online chat forums. To be able to capture sequential dialogue act dependency on chats, they applied a CRF model. They demonstrated that, compared with other classifiers, their CRF model performed the best. In their later work (Kim *et al*., 2012), they extended the domain to multi-party live chats and proposed new features for that domain.

3 Capturing Conversation Structure in Email Threads

In this section, we describe how to build a fragment quotation graph which captures the conversation structure of any email thread at finer granularity. This graph was developed and shown to be effective by Carenini *et al*. (2011). A key assumption of this approach is that in order to effectively perform summarization and dialogue act modeling, a fine graph representation of the underlying conversation structure is needed.

Here, we start with the sample email conversation shown in Figure 1 (a). For convenience, the content of the emails is represented as a sequence of fragments.

First, we identify all new and quoted fragments. For example, email E1 is composed of one new fragment, 'b', and one quoted fragment, 'a'. As for email E3, since we do not yet know whether or not 'd' and 'e' are different fragments, we consider E3 as being composed of one new fragment, 'de' and one quoted fragment, 'b'.

Second, we identify distinct fragments. To do this, we first identify overlaps by comparing fragments with each other. If necessary, we split the fragments and remove any duplicates from them. For example, a fragment, 'de', in E3 is

split into ‘d’ and ‘e’ after being compared with fragments in E4 and the duplicates are removed. By applying this process to all of the emails, seven distinct fragments, a, b ..., and, g remain in this example.

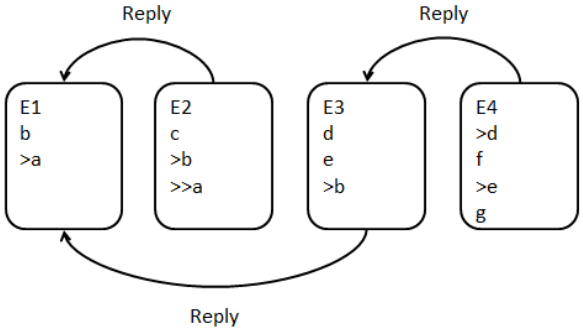
In the third step, edges which represent the replying relationships among the fragments are created. These edges are determined based on the assumption that any fragment is a reply to neighboring quotations (the quoted fragments immediately preceding or following the current one). For example, the neighboring nodes of ‘f’ in E4 are ‘d’ and ‘e’. Thus, we create two edges from node ‘f’ in E4 to node ‘d’ and ‘e’ in E3. In the same way, we see that the neighboring node of ‘g’ in E4 is ‘e’. Hence, there is one edge from node ‘g’ to ‘e’. If no quotation is contained in a reply email, we connect the fragments in the email to fragments in emails to which it reply.

In email threads, there are cases in which the original email with its quotations is missing from the user’s folder, as in the case of ‘a’ in Figure 1 (a). These types of emails are called hidden emails. Carenini *et al*. (2005) studied in detail how these email types might be treated and their influence on email summarization.

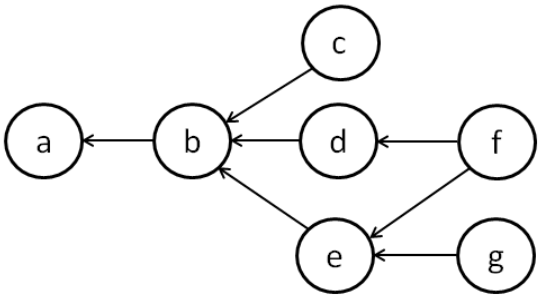
Figure 1 (b) shows the completed fragment quotation graph of the email thread shown in Figure 1 (a). In the fragment quotation graph structure, all paths (e.g., a-b-c, a-b-d-f, a-b-e-f, and a-b-e-g in Figure 1 (b)) capture the adjacent relationships between email fragments. Hence, we use every path that can be derived from the graph as our dataset. However, in this case, when we run the labeling task on these paths, we obtain multiple labels for some of the sentences because the sentences in fragments such as ‘a’, ‘b’, and ‘f’ in Figure 1 (b) are shared among multiple paths. Therefore, to assign a label to one of these sentences, we take the label more frequently assigned to that sentence when all its paths are considered (i.e., the majority vote).

4 Features

For both dialogue act modeling and extractive summarization, many effective sentence features have been discovered so far. Interestingly, some common features are shown to be effective in both tasks. This section explains the features used in our model. We begin with the features for extractive summarization and then describe how we derive the features for dialogue act modeling. All the features explained in this section, whether they belong to extractive summarization or dialogue act modeling, are included in our model.



(a) A possible configuration of an email conversation (E2 and E3 reply to E1, and E4 replies to E3)



(b) An example of a fragment quotation graph

Figure 1: A fragment quotation graph derived from a possible configuration of an email conversation

4.1 Extractive Summarization Features

The features we use for extractive summarization are mostly from Carenini *et al*. (2008) and Rambow *et al*. (2004) and have proven to be effective on conversational data. Details of these features are described below. Note that all sentences in an email thread are ordered based on paths derived from a fragment quotation graph.

Length Feature: The number of words in each sentence.
Relative Position Feature: The number of sentences preceding the current divided by the total number of sentences in one path.
Thread Name Overlaps Feature: The number of overlaps of the content words between the email thread title and a sentence.
Subject Name Overlaps Feature: The number of overlaps of the content words between the subject of the email and a sentence.
Question Feature: A binary feature that indicates whether or not a sentence has a question mark.
CC Feature: A binary feature that indicates whether or not an email contains CC.

Participation Dominance Feature: The number of utterances each person makes in one path.

Finally, we also include a simplified version of the ClueWordScore (CWS) developed by Carenini *et al*. (2007), which is listed below.

Simplified CWS Feature: The number of overlaps of the content words that occur in both the current and adjacent sentences in the path, ignoring stopwords.

4.2 Dialogue Act Features

The relative positions and length features have proven to be beneficial to both tasks (Jeong et al., 2009; Carenini *et al*., 2008). Hence, these are categorized as both dialogue acts and extractive summarization features. In addition, we use word and POS n-grams as our features for dialogue act modeling. These features are extracted by the following process explained in Carvalho *et al.* (2006). However, we extend the original approach in order to further abstract n-gram features to avoid making them too sparse to be effective. In this section, we describe the derivation process in detail.

A multi-step approach is used to generate word n-gram features. First, all words are tagged with the named entity using the Stanford Named Entity Recognizer (Finkel *et al.*, 2005), and are then replaced with these tags. Second, a sequence of word-replacement tasks is applied to all email messages. Initially, some types of punctuation marks (e.g., <>()[];:. and ,) and extra spaces are removed. Then, shortened phrases such as "I'm" and "We'll" are substituted for more formal versions such as "I am" and "We will". Next, other replacement tasks are performed. Some of them are described in Table1. In the third step, unigrams and bigrams are extracted. In this paper, unigrams and bigrams refer to all possible sequences of length one and two terms. After extracting all unigrams and bigrams for each dialogue act, we then compute Information Gain Score (Forman, 2003) and select the n-grams whose scores are in the top five greatest on the training set. In this way, we can automatically detect features that represent the characteristics of each dialogue act. In addition to word n-grams, we also include POS n-grams in our features. In a similar way, we first tag each word in sentences with POS using the Stanford POS tagger (Toutanova *et al.*, 2003). Then, for each dialogue act, we extract bigrams and trigrams, all of which are scored by the Information Gain. Based on their scores, we select the POS bigram and trigram features whose scores are within the top five greatest. One example of word n-gram features for a Question dialogue act selected by this derivation method is shown in Table 2.

Pattern	**Replacement**
'why', 'where', 'who', 'what' 'when'	[WWHH]
nominative pronouns	[I]
objective pronouns	[ME]
'it', 'those', 'these', 'this', 'that'	[IT]
'will', 'would', 'shall', 'should', 'must'	[MODAL_STRONG]
'can', 'could', 'may', 'might'	[MODAL_WEAK]
'do', 'does', 'did', 'done'	[DO]
'is', 'was', 'were', 'are', 'been' 'be', 'am'	[BE]
'after' , 'before', 'during'	[AAAFTER]
'Jack", "Wendy"	[Personal_PRONOUN]
"New York"	[LOCATION]
"*Acme Corp."*	[ORGANIZATION]

Table 1: Some Preprocessing Replacement Pattern

Word Unigram	**Word Bigram**
?	[MODAL_STRONG] [I]
anyone	[IT] ?
WWHH	[DO] anyone
deny	[WWHH] [BE]
[Personal _PRONOUN]	[BE] [IT]

Table 2: Sample word n-grams selected as the features for Question dialogue act

5 The Sequential Labeling Task

We use a Dynamic Conditional Random Field (DCRF) (Sutton *et al*., 2004) for labeling tasks. A DCRF is a generalization of a linear-chain CRF which allows us to represent complex interaction between labels. To be more precise, it is a conditionally-trained undirected graphical model whose structure and parameters are repeated over a sequence. Hence, it is the most appropriate method for performing multiple labeling tasks on the same sequence.

Our DCRF uses the graph structure shown in Figure 2 with one chain (the top X nodes) modeling extractive summary and the other (the middle Y nodes) modeling dialogue acts. Each node in the observation sequence (the bottom Z nodes) corresponds to each sentence in a path of the fragment quotation graph of the email thread. As shown in Figure 2, the graph structure captures the relationship between extractive summaries and dialogue acts by connecting their nodes.
We use Mallet[1] (McCallum, 2002) to implement our DCRF model. It uses l2-based regularization to avoid overfitting, and a limited BFGS fitting algorithm to learn the DCRF model parameters. Also, it uses tree-based reparameterization (Wainwright *et al*., 2002) to compute the posterior marginal, or inference.

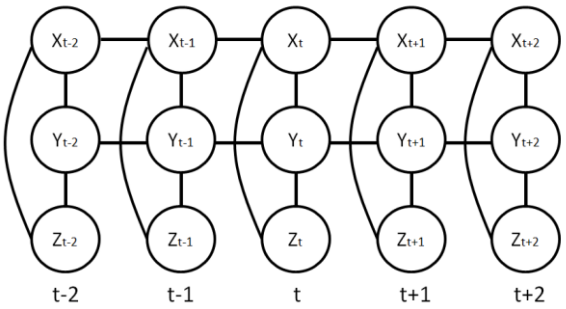


Figure 2: The DCRF model used to create extractive summaries and model dialogue acts

6 Empirical Evaluations

6.1 Dataset Setup

In our experiment, the publically available BC3 corpus[2] (Ulrich *et al*., 2008) is used for training and evaluation purposes. The corpus contains email threads from the World Wide Web Consortium (W3C) mailing list. It consists of 40 threads with an average of five emails per thread. The corpus provides extractive summaries of each email thread, all of which were annotated by three annotators. Hence, we use sentences that are selected by more than one annotator as the gold standard summary for each conversation.

In addition, all sentences in the 39 out of 40 threads are annotated for dialogue act tags. The tagset consists of five general and 12 specific tags. All of these tags are based on Jeong *et al*. (2009). For our experiment, considering that our data is relatively small, we decide to use the coarser five tag set. The details are shown in Table 3.

[1] http://mallet.cs.umass.edu
[2] http://www.cs.ubc.ca/nest/lci/bc3.html

Tag	Description	Relative Frequency (%)
S	Statement	73.8
Q	Question	7.92
R	Reply	5.23
Su	Suggestion	5.62
M	Miscellaneous	7.46

Table 3: Dialogue act tag categories and their relative frequency in the BC3 corpus

After removing quoted sentences and redundant information such as senders and addresses, 1300 distinct sentences remain in the 39 email threads. The detailed content of the corpus is summarized in Table 4.

	Total Dataset
No. of Threads	39
No. of Sentences	1300
No. of Extractive Summary Sentences	521
No. of S Sentences	959
No. of Q Sentences	103
No. of R Sentences	68
No. of Su Sentences	73
No. of M Sentences	97

Table 4: Detailed content of the BC3 corpus

6.2 Evaluation Metrics

Here, we introduce evaluation metrics for our joint model of extractive summarization and dialogue act recognition.

The CRF model has been shown to be the effective one in both dialogue act modeling and extractive summarization (Shen *et al*., 2007; Kim *et al*., 2010; Kim *et al*., 2012). Hence, for comparison, we implement two different CRFs, one for extractive summarization and the other for dialogue act modeling. When classifying extractive summaries using the CRF, we only use its extractive summarization features. Similarly, when modeling dialogue acts, we only use its dialogue act features. In addition, we also com-

pare our system with a non-sequential classifier, a support vector machine (SVM), with the same settings as those described above. For these implementations, we use Mallet and SVM-light package[3] (Joachims, 1999).

In our experiment, we first measure separately the performance of extractive summarization and dialogue act modeling. The performance of extractive summarization is measured by its averaged precision, recall, and F-measure. For dialogue acts**,** we report the averaged-micro and macro accuracies as well as the averaged accuracies of each dialogue act.

Second, we evaluate the combined performance of extractive summarization and dialogue act modeling tasks. In general, we are interested in the dialogue acts in summary sentences because they can be later used as input for other natural language processing applications such as automatic abstractive summarization (Murray *et al*., 2010). Therefore, we measure the performance of our model with the following modified precision (*Pre'*), recall (*Rec'*), and F-measure (*F'*):

$$Pre' = \frac{\{No.\ of\ correctly\ classified\ \ sentences\ \}}{\{No.of\ sentences\ classified\ as\ summary\ setences\}} \quad (1)$$

$$Rec' = \frac{\{No.\ of\ correctly\ classified\ \ sentences\ \}}{\{No.of\ true\ summary\ sentences\}} \quad (2)$$

$$F' = \frac{2 \times Pre' \times Rec'}{Pre' + Rec'} \quad (3)$$

where a *correctly classified sentence* refers to a true summary sentence that is classified as such and whose dialogue acts are also correctly classified.

6.3 Experiment Procedure

For all cases, we run five sets of 10-fold cross validation to train and test the classifiers on a shuffled dataset and calculate the average of the results. For each cross validation run, we extract all features following the process described in Section 4 on the training set. When comparing these two baselines with our model, we report p-values obtained from a student paired t-test on the results to determine their significance.

[3] http://www.cs.cornell.edu/people/tj/svm_light

6.4 Results

The performances of extractive summarization and dialogue act modeling using the three methods are summarized in Table 5 and 6, respectively.

	DCRF	**CRF**	**SVM**
F-measure	**0.485**	0.428	0.397
t-test's p-value		0.00046	2.5E-07
Precision	0.562	0.591	**0.675**
Recall	**0.457**	0.370	0.308

Table 5: A comparison of the extractive summarization performance of our DCRF model and the two baselines based on precision, recall, and F-measure

	DCRF	**CRF**	**SVM**
Micro Accuracy	**0.785**	0.779	0.775
t-test's p-value		0.116	0.036
Macro Accuracy	**0.516**	**0.516**	0.304
t-test's p-value		0.950	5.2E-32
S Accuracy	0.901	0.892	**0.999**
Q Accuracy	**0.832**	0.809	0.465
R Accuracy	**0.580**	0.575	0.05
Su Accuracy	**0.139**	0.108	0.00
M Accuracy	0.126	**0.198**	0.00

Table 6: A comparison of the dialogue act modeling performance of our DCRF model and the two baselines based on averaged accuracies

From Table 5, we observe that, in terms of extractive summarization results, our DCRF model significantly outperforms the two baselines. Noticeable improvements can be seen for the recall and F-measure. In terms of F-measure, compared with the CRF and SVM, our model improves by 5.7% and 8.8% respectively. The p-values obtained from the t-test indicate that our results are statistically significantly different ($p < 0.05$) from those of the two baselines.

Regarding dialogue act modeling, the results are summarized in Table 6. While no improvement is shown for the micro-averaged accuracy, our model and the CRF significantly outperform the SVM in terms of the macro-averaged accura-

cy. Both our model and the CRF consider the sequential structure of the conversation, which is not captured in the SVM model. Clearly, this indicates that the sequential models are effective in modeling dialogue acts due to their ability to capture the inter-utterance relations of conversations.

Compared with the CRF, our DCRF model outperforms it in most cases except in classifying the 'M' dialogue act. However these improvements are not significant as t-test of both macro and micro-averaged accuracies indicate that the differences are not statistically significant ($p > 0.05$).

Another item to be mentioned here is that the accuracies of classifying 'R', 'Su' and 'M' dialogue acts are relatively low. This issue applies to all classifiers and is plausibly due to the small dataset. There are only 68, 73 and 97 sentences, respectively, out of 1300 that are labeled as 'R', 'Su' and 'M' in the BC3 corpus. Since our dialogue act classifiers rely heavily on n-gram features, were the data small, these features would be too sparse to effectively represent the characteristics of the dialogue acts. However, compared with the SVM results, our joint model and the CRF perform significantly better in classifying these dialogue acts. This also explains why the sequential model is preferable in dialogue act modeling.

Note that despite the small dataset, all the classifiers are relatively accurate in classifying 'Q'. This is because n-gram features selected for 'Q' such as '?' and 'WWHH' are very specific to this dialogue act, which makes the task of 'Q' classification easier compared to those of others.

Next, we discuss the result of the combined performance. The performances of our model and the two baselines are summarized in Table 7.

	DCRF	**CRF**	**SVM**
F-measure'	**0.352**	0.324	0.292
t-test's p-value		0.015	3.3E-05
Precision'	0.407	0.450	**0.501**
Recall'	**0.335**	0.280	0.227

Table 7: A comparison of the overall performance of our DCRF model and the two baselines based on modified precision, recall and F-measure

We see that our DCRF model significantly outperforms the two baselines. While our model yields the lowest *Pre'* of all, its *Rec'* is much greater than the other two baselines and this leads to its achieving the highest *F'*. Compared with the CRF and SVM, the *F'* obtained from our system improves by 2.8% and 6% respectively. In addition, the p-values show that the results of our model are statistically significant ($p < 0.05$) compared with those of the two baselines.

Overall, these experiments clearly indicate that our model is effective in classifying both dialogue acts and summary sentences.

7 Conclusions and Future Work

In this work, we have explored a new automated approach for extractive summarization and dialogue act modeling on email threads. In particular, we have presented a statistical approach for jointly modeling dialogue acts and extractive summarization in a single DCRF. The empirical results demonstrate that our approach outperforms the two baselines on the summarization task without loss of performance on the dialogue act modeling one. In the future, we would like to extend our approach by exploiting more effective features. We also plan to apply our approach to different domains possessing large dataset.

Acknowledgements

We are grateful to Yashar Mehdad, Raimond Ng, Maryam Tavafi and Shafiq Joty for their comments and UBC LCI group and ICICS for financial support.

References

J. Allen, N. Chambers, G. Ferguson, L. Galescu, H. Jung, and W. Taysom. Plow: A collaborative task learning agent. In AAAI-07, pages 22–26, 2007.

Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2011. Methods for Mining and Summarizing Text Conversations. *Morgan Claypool*.

Giuseppe Carenini, Raymond Ng, and Xiaodong Zhou. 2005. Scalable discovery of hidden emails from large folders. In *ACM SIGKDD'05*, pages 544–549.

Giuseppe Carenini, Raymond Ng, and Xiaodong Zhou. 2008. Summarizing Emails with Conversational Cohesion and Subjectivity In *proceeding 46th Annual Meetint Assoc.for Computational Linguistics*, page 353-361.

Giuseppe Carenini, Raymond Ng, and Xiaodong Zhou. 2007. Summarizing email conversations with clue words. *16th International World Wide Web Conference (ACM WWW'07).*

Vitor R. Carvalho and William W. Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech, ACTS '09*, pages 35–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research, 3:1289–1305.

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

Thorsten Joachims. 1999 Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), *MIT-Press, 1999.*

Shafiq Joty, Giuseppe Carenini, and Lin, Chin-Yew Lin. 2011. Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. In *Proceedings of the twenty second International Joint Conference on Artificial Intelligence (IJCAI) 2011*. Barcelona, Spain.

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2009 Finding Topics in Emails: Is LDA enough? *NIPS-2009 workshop on applications for topic models: text and beyond.* Whistler, Canada.

McCallum, A. Kachites, 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

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, Lawrence Cavedon and Timothy Baldwin (2012) Classifying Dialogue Acts in Multiparty Live Chats, In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26)*, Bali, Indonesia, pp. 463—472.

Gabriel Murray and Giuseppe Carenini. 2008. Summarizing Spoken and Written Conversations. *Empirical Methods in NLP (EMNLP 2008)*, Waikiki, Hawaii, 2008.

Gabriel Murray and Giuseppe Carenini. 2010. Summarizing Spoken and Written Conversations. *Generating and Validating Abstracts of Meeting Conversations: a User study (INLG 2010)*, Dublin, Ireland, 2010.

Gabriel Murray, Renals Steve, and Carletta Jean. 2005a. Extrative summarization of meeting recordings. In *Proceeding of Interspeech 2005*, Lisbon, Portugal, pages 593-596.

Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. 2004. Summarizing email threads. In *Proceedings of HLTNAACL 2004*.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In Proc. of IJCAI, volume 7, 2862–2867.

Charles Sutton, Khashayar Rohanimanesh, and Andrew McCallum. 2004. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *Proc. ICML*.

Maryam Tavafi, Yashar Mehdad, Shafiq Joty, Giuseppe Carenini and Raymond Ng. 2013. Dialogue Act Recognition in Synchronous and Asynchronous Conversations. *In Proceedings of the SIGDIAL 2013 Conference*, pages 117–121, Metz, France. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

Jan Ulrich, Giuseppe Carenini, Gabriel Murray, and Raymond T. Ng: Regression-Based Summarization of Email Conversations. ICWSM 2009

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

Martin J. Wainwright, Tommi Jaakkola, and Alan S. Willsky. 2002. Treebased Reparameterization for Approximate Inference on Loopy Graphs. In *Advances in Neural Information Processing Systems 14*, pages 1001 1008. MIT Press.