Regression-Based Summarization of Email Conversations

Jan Ulrich and **Giuseppe Carenini** and **Gabriel Murray** and **Raymond Ng**
{ulrichj, carenini, gabrielm, rng}@cs.ubc.ca
Department of Computer Science
University of British Columbia, Canada

Abstract

In this paper we present a regression-based machine learning approach to email thread summarization. The regression model is able to take advantage of multiple gold-standard annotations for training purposes, in contrast to most work with binary classifiers. We also investigate the usefulness of novel features such as speech acts. This paper also introduces a newly created and publicly available email corpus for summarization research. We show that regression-based classifiers perform better than binary classifiers because they preserve more information about annotator judgements. In our comparison between different regression-based classifiers, we found that Bagging and Gaussian Processes have the highest weighted recall.

Introduction and Related Work

Email has become a ubiquitous social medium. It is a convenient form of communication due to its speed and lack of cost. However, the volume of received email entails a great cost in terms of the time required to read, sort and archive the incoming data. The problem of email overload is only going to keep increasing.

Summarization is a promising way to reducing this email triage. Email summarization has many more uses than just summarizing incoming emails. In the business world, email summarization can be used as a form of corporate memory, where the thread summaries represent all the previous business decisions that have been made. As another possibility, it also allows a new team member to more easily and quickly catch up on an ongoing conversation in a discussion forum.

In this paper, we explore new techniques in extractive email thread summarization using several regression-based classifiers and novel sentence features.

Studies on supervised summarization of email threads ((Rambow *et al.* 2004), (Carenini, Ng, & Zhou 2007)) have been carried out using binary classifiers, with definitive positive/negative class labels. It is well known in the summarization community that human judges tend to exhibit divergent opinions in creating gold-standard summaries. We therefore employ a regression framework that uses continuous classification so that we minimize the information loss of the multiple human judgments. Both the labeled annotations and the summarization predictions are continuous.

Rambow et al. pioneered email thread summarization using supervised machine learning for sentence extraction(Rambow *et al.* 2004). They focused on creating a set of sentence features by using proven text extraction features (e.g., relative position of the sentence in the document) and adding email specific features (e.g., similarity of the sentence with the email Subject field). Their results show that email specific features significantly improved summarization. We use their feature set as the baseline in our experiments, but will introduce new features that have been used in email research but not email summarization.

Other email thread summarization work has included Carenini et al. who created an unsupervised system based on *clue words*(CWS) (Carenini, Ng, & Zhou 2007). Their approach relies on the conversation structure of the emails and the repeated words throughout the thread. Representing the email conversation as a graph structure with email fragments as nodes, clue words are the highly informative words that occur in adjacent nodes. We use the unsupervised CWS as one of the baselines in our experiments. Furthermore, we use a clue-word-based score as one of our novel sentence features.

There has not been a comparative study between different classification algorithms for summarizing email conversations. We have created a framework where we can evaluate the effectiveness of various classifiers. We use regression algorithms since we are training on continuously labeled data. To justify the switch to a continuous framework, we compare binary classifiers to regression algorithms and show that the latter are indeed more accurate.

Corpora

In our experiments we utilize two corpora for training and evaluation purposes: the BC3 corpus developed for this study (Ulrich, Murray, & Carenini 2008), and the Enron email corpus annotated by (Carenini, Ng, & Zhou 2007).

The British Columbia Conversation Corpus

The BC3 corpus is a collection of multimodal conversational data. The corpus consists of email threads annotated for summarization. It contains 40 threads with an average of 5 emails per thread. The corpus provides extractive as well

as abstractive summaries of the conversations. The email threads come from the mailing list data from the W3C corpus which was derived from a crawl of the World Wide Web Consortium's sites at w3c.org. The mailing list subset is comprised of nearly 200,000 documents, and TREC participants have provided thread structure based on reply-to relations and subject overlap. The BC3 corpus is publicly available at http://www.cs.ubc.ca/labs/lci/bc3.html.

BC3 Corpus Summarization Annotation The BC3 emails have been annotated for summarization as well as labeled with sentence-level linguistic features. Annotators were asked to write an abstractive summary of the thread with links to the original content. This results in a many-to-many mapping between extractive sentences and abstractive sentences for each annotator.

In a second step the annotators were told to create an extractive summary by selecting sentences in the original text that compromise a summary. This scheme closely follows the methods used by researchers in the AMI project (http://www.amiproject.org) in annotating their meeting corpus (Carletta *et al.* 2006).

Three annotators annotated each thread. Their annotations had a kappa agreement of 0.50 for the extracted sentences. 10 annotators were recruited from the University of British Columbia. They were all proficient in English as we screened potential annotators with a small written statement.

Taking the original work of Carvalho(Carvalho & Cohen 2005) as inspiration, we decided to annotate speech acts in the new corpus. However we used a subset of the original speech acts that we consider more informative: *Propose* sentence proposes a joint activity; *Request* asks the recipient to perform an activity; *Commit* sentence commits the sender to some future course of action; and *Meeting* sentence is regarding a joint activity in time or space.

The annotation also labels subjective sentences as subjectivity was found useful for email summarization (Carenini, Ng, & Zhou 2008). The label *Subj* means that the writer is expressing an opinion or strong sentiment. Additionally, the annotators labeled meta sentences as this was found useful by Murray and Renals (Murray & Renals to appear 2008) in meeting summarization. A sentence is labeled *Meta* if it refers to the email thread that it is part of.

The Enron Corpus

From the Enron email corpus released after the legal investigation into the Enron corporation, 39 email threads were selected from the 10 largest email inbox folders and then annotated by 50 annotators as described in (Carenini, Ng, & Zhou 2007). Each thread was annotated by five annotators. The annotators were asked to pick 30% of the original sentences such that the summary contained the overall information in the email and could be understood without referencing the original email thread. Sentences were labeled as either *essential* or *optional*, where an essential sentence is vital to the understanding of the summary and an optional sentence only elaborates on the meaning of the conversation. The overall score of a sentence was then computed by aggregating the judgments of all the annotators.

Continuous Classification Setting

Often extractive summarization is simplified to a binary classification of whether a sentence will be included in a summary or not. This is problematic because sentences in the training set need to be labeled as either included or not, while frequently annotators do not fully agree on which sentences should be included in a summary as can be seen by kappa values of 0.5 in the BC3 corpus. A threshold is needed, which is often picked quite arbitrarily.

Our solution is to move to a continuous setting by using the average annotator score as our gold-standard label, rather than employing a binary scheme. We then use a regression-based classifier that outputs a continuous importance score. Summaries can be created for a desired length by simply taking the appropriate number of top scoring sentences.

Sentence Length Normalization

Feature calculation and classification are performed at the sentence level while final summaries are usually limited by word length. The classifiers should take this into account during training. We therefore normalize sentence annotator score by sentence length. The normalized score represents the information or importance content of the sentence per word.

Sentence Features

We compare different possible feature sets. We start with a feature set that was previously used in email summarization (Rambow *et al.* 2004) and add speech acts, meta labels, and subjectivity labels.

Due to the conversational nature of our corpora we decided to add speech act labels. We compare between automatically generated and manually annotated speech act features in the BC3 corpus. For automatic speech acts features generation, we use Ciranda (Carvalho & Cohen 2005). To test different levels of granularity, we ran the classifier individually on each sentence (SA_S) as well as at the email level (SA_E).

By drawing from work in related fields, we also included a feature marking meta sentences (i.e., sentences referring to the current conversation), which have been shown to be very useful in meeting summarization (Murray & Renals to appear 2008). Furthermore, we have included a feature assessing the subjectivity of the sentence as this improved summarization quality in another email summarization approach (Carenini, Ng, & Zhou 2008).

Experimental Setup

Our generated summaries were limited to 30% word length. The machine learning based summarizers rely on two software packages for their implementation. MEAD (Radev *et al.* 2004) was used as the summarization framework consisting of three stages: generating sentence features, sentence classification, and sentence reranking. WEKA was used for the implementation of the different classifiers.

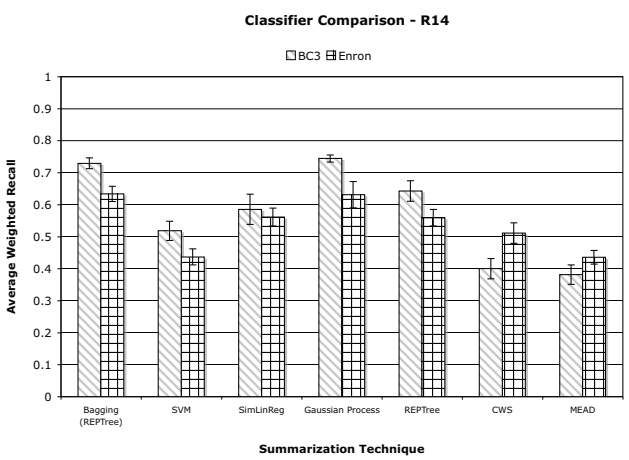

Figure 1: Classifier comparison using R14, the original features presented in Rambow et al.

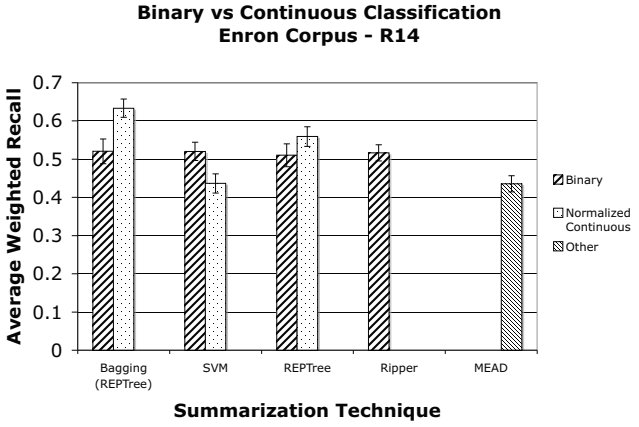

Figure 2: Performance differences between binary and continuous classifiers in the Enron corpus.

Evaluation Procedure

The summaries of the Enron and BC3 corpus generated by the different summarizers were compared to the human generated gold standards using 10-fold cross-validation with 90% training data.

Evaluation Metric: Weighted Recall

Human annotators cannot typically agree on a single perfect summary for a given email thread. We therefore have several annotators summarize an email thread and then score the machine-generated summary against all of them. In the Enron corpus there were 5 annotators per thread and in the BC3 corpus we had 3 annotators per thread. We measure the recall score against an ideal summary weighed across all annotators. The reason we use a recall score instead of an f-measure, is because the length of the summary is fixed. So formally we have:

$$WeightedRecall = \frac{\sum_{i \in Sent_{Sum}} Nscore_i}{\sum_{i \in Sent_{GS}} Nscore_i}$$

$Nscore$ is the normalized version of the corpus dependent sentence score, $Sent_{Sum}$ are the sentences in the generated summary, and $Sent_{GS}$ are the sentences in the gold standard summary.

Results

In the following sections we provide a description of our results as we compared different classifiers, continuous vs. binary labels, and different feature sets.

Effect of Classifier Choice

For comparing the different classifiers we have chosen to use the baseline feature set, R14, and continuous sentence labels normalized by sentence length because this provides the best overall results. In Figure 1 we compare the performance of 7 summarization approaches. The first 5 are supervised while the last 2 are unsupervised. It can be seen that the supervised machine learning algorithms outperform the unsupervised versions. Bagging of the REPTree classifier had the highest average weighted recall with a score of 0.63 in the Enron dataset. Bagging performs significantly better than any other approach except for Gaussian Processes. A runtime analysis was also performed between the different algorithms, and Bagging was the fastest machine learning algorithm and Gaussian Processes was significantly slower. CWS was the fastest unsupervised algorithm.

The results are similar for the two corpora. Bagging and Gaussian Processes therefore seem to be the most effective classifiers for email summarization. SVMs performed poorly here but as will be shown later, SVMs seem to be better suited for a binary framework.

Continuous vs Binary Labels

In this section we evaluate the difference between using a binary label and a continuous label that is normalized by sentence length for training data. The binary label is generated by taking the annotation score and using a threshold to decide whether to include a sentence or not. The threshold used in the Enron corpus was 8 (used in (Carenini, Ng, & Zhou 2007)) and the threshold used in the BC3 corpus was 2, as it signifies that the majority of the annotators wanted to include the sentence in the summary. We have also included Ripper in this evaluation as this was the algorithm used in previous work on extractive email summarization (Rambow *et al.* 2004).

In Figure 2 it can be seen that the best performance is achieved using a continuous normalized framework and the Bagging algorithm. The continuous labels using Bagging were significantly better with p-values of less than 0.00001 compared to binary bagging, Ripper, and MEAD individually. However not all algorithms are suited for a continuous regression setting. SVM actually performs better in the binary framework. We hypothesis that this is because SVM is a margin maximizing algorithm and that this works best when having only two classes.

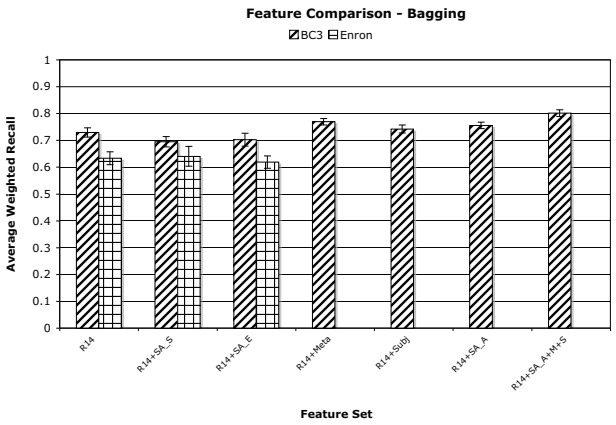

Figure 3: Feature comparison on the BC3 & Enron corpora. SA_A, Meta (M), Subj (S) are manually annotated speech acts, SA_S are sentence level Ciranda labels, and SA_E are email level Ciranda labels.

Feature Sets

The baseline feature set, R14 (from (Rambow *et al.* 2004)), was compared to additional speech act features, meta labels, and subjectivity labels. Ciranda was used to automatically label the speech acts at the email level (SA_E) and at the sentence level (SA_S). Manual speech act annotations (SA_A) as well as manual annotations for meta sentences (Meta) and subjective sentences (Subj) were also available in the BC3 corpus.

We only show results of the best performing classifier, Bagging, but these results generalize to the other classifiers.

Figure 3 shows that the generated speech act labels with Ciranda did not help with summarization. This included both email level annotation and sentence level annotation. However the manually labeled features were significantly better than the baseline as can be seen in Figure 3. The generated features did not perform well for different reasons. The email level annotations were too coarse, and the sentence level annotations were too noisy as there was not enough data for Ciranda to select the correct label. Ciranda was trained to label emails, not sentences. The fact that the manually generated labels were useful for summarization shows that while speech acts are indeed a useful feature, the generated labels we used are just too noisy or too coarse. It would be interesting to pursue automatic speech act classification at the sentence level in future work.

Meta sentences had the highest increase in weighted recall of all the new features by themselves. This was somewhat surprising as meta sentences had the lowest kappa agreement between all the annotators. This shows that it is not necessarily important for all the annotators to agree.

An additional feature, CWS, which is the clue word score as generated in previous work (Carenini, Ng, & Zhou 2007), surprisingly also did not improve weighted recall significantly. It seems that clue words are a good feature by themselves in the clue word summarizer, but they are not a useful additional feature in a supervised machine learning system.

Conclusions

We created a continuous classification framework for summarizing email threads and evaluated it on two corpora, one of which was developed for this study. Our results show that the best regression-based classifiers for email thread summarization perform better than binary classifiers because they preserve more information. In our comparison between different classifiers, we found that Bagging and Gaussian Processes have the highest weighted recall, but Bagging is more efficient. We confirm the 14 features from (Rambow *et al.* 2004) provide good results but can be improved with other features. The results on our new dataset show that speech acts are a very useful feature if they can be generated with higher accuracy at the sentence level. Meta sentences and subjectivity were also shown to be useful features for email summarization.

References

Carenini, G.; Ng, R. T.; and Zhou, X. 2007. Summarizing email conversations with clue words. *16th International World Wide Web Conference (ACM WWW'07)*.

Carenini, G.; Ng, R.; and Zhou, X. 2008. Summarizing emails with conversational cohesion and subjectivity. *The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008)*.

Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; Lathoud, G.; Lincoln, M.; Lisowska, A.; McCowan, I.; Post, W.; Reidsma, D.; and Wellner, P. 2006. *The AMI Meeting Corpus: A Pre-announcement*. Springer Berlin. 28–39.

Carvalho, V. R., and Cohen, W. W. 2005. On the collective classification of email "speech acts". In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 345–352. New York, NY, USA: ACM.

Murray, G., and Renals, S. to appear, 2008. Meta comments for summarizing meeting speech. In *Proc. of MLMI 2008, Utrecht, Netherlands*.

Radev, D.; Allison, T.; Blair-Goldensohn, S.; Blitzer, J.; Çelebi, A.; Dimitrov, S.; Drabek, E.; Hakim, A.; Lam, W.; Liu, D.; Otterbacher, J.; Qi, H.; Saggion, H.; Teufel, S.; Topper, M.; Winkel, A.; and Zhu, Z. 2004. MEAD - a platform for multidocument multilingual text summarization. In *Proceedings of LREC 2004*.

Rambow, O.; Shrestha, L.; Chen, J.; and Lauridsen, C. 2004. Summarizing email threads. In *Proceedings of HLT-NAACL 2004*.

Ulrich, J.; Murray, G.; and Carenini, G. 2008. A publicly available annotated corpus for supervised email summarization. *AAAI-2008 EMAIL Workshop*.