Open access • Journal Article • DOI:10.1016/J.IPM.2007.09.007

Single-document and multi-document summarization techniques for email threads using sentence compression — Source link

David Zajic, Bonnie J. Dorr, Jimmy Lin

Institutions: University of Maryland, College Park

Published on: 01 Jul 2008 - Information Processing and Management (Pergamon Press, Inc.)

Topics: Automatic summarization, Multi-document summarization and Sentence

Single-Document and Multi-Document Summarization Techniques for Email Threads Using Sentence Compression*

David M. Zajic[1], Bonnie J. Dorr[1], Jimmy Lin[2]
[1]Department of Computer Science
[2]College of Information Studies
University of Maryland
College Park, MD 20742
dmzajic@umiacs.umd.edu, bonnie@cs.umd.edu, jimmylin@umd.edu

Abstract

We present two approaches to email thread summarization: Collective Message Summarization (CMS) applies a multi-document summarization approach, while Individual Message Summarization (IMS) treats the problem as a sequence of single-document summarization tasks. Both approaches are implemented in our general framework driven by sentence compression. Instead of a purely extractive approach, we employ linguistic and statistical methods to generate multiple compressions, and then select from those candidates to produce a final summary. We demonstrate these ideas on the Enron collection—a very challenging corpus because of the highly technical language. Experimental results point to two findings: that CMS represents a better approach to email thread summarization, and that current sentence compression techniques do not improve summarization performance in this genre.

1 Introduction

Over the past few decades, email has become the preferred medium of communication for many organizations and individuals. As a growing portion of our lives is captured over email exchanges, the phenomenon of the overcrowded inbox is becoming an increasingly serious impediment to communications and productivity. Furthermore, large existing email archives hold valuable knowledge that is often not captured elsewhere. Systems that help users organize and access email are clearly important in modern information societies.

This work tackles a problem that contributes to the broader goal of providing users with effective applications to access large email collections—the task of summarizing email threads. Such a capability could, for example, be deployed on the output of email or desktop search systems; see, for example (Craswell et al., 2005; Cutrell et al., 2006). Summarization technology might be especially attractive for display of email on mobile devices with limited screen area. Previous work has shown that summarization techniques are useful in document retrieval tasks (Mani et al., 2002; Dorr et al., 2005). Similarly, we believe that an email thread summarization system could constitute an important component of a larger email application. Specifically, we adopt the working assumption that at least

*Please cite as: David Zajic, Bonnie Dorr, and Jimmy Lin. Single-Document and Multi-Document Summarization Techniques for Email Threads Using Sentence Compression. *Information Processing and Management*, 44(4):1600–1610, 2008. [DOI: 10.1016/j.ipm.2007.09.007] This is the pre-print version of a published article. Citations to and quotations from this work should reference that publication. If you cite this work, please check that the published form contains precisely the material to which you intend to refer. (Received 27 March 2007; revised 10 September 2007; accepted 12 September 2007) This document prepared May 31, 2008, and may have minor differences from the published version.

in certain situations, users will find reading summaries preferable to reading the entire (possibly much longer) thread—although see Lam et al. (2002) for results from a pilot study that highlight many of the challenges associated with this task.

We describe two separate approaches to email thread summarization that adapt existing techniques: one treats the problem as a sequence of single-document summarization tasks (a technique we call Individual Message Summarization, or IMS), and the other treats the problem as a variant of multi-document summarization (a technique we call Collective Message Summarization, or CMS). Both approaches involve selecting important sentences from email messages and compressing them (i.e., removing unimportant fragments). Our implemented systems were evaluated using data from the Enron corpus, using a small manually-created test corpus. Experimental results suggest two findings: that CMS is superior to IMS, and that current sentence compression techniques do not improve summarization performance in this genre (an interesting negative result). We also discuss the challenges associated with both this task and specifically with the Enron corpus.

This paper is organized as followed: we first present a general overview of email summarization, including a discussion of related work. Section 3 describes our general framework for text summarization and specific approaches we have developed for email thread summarization. Section 4 focuses on the test collection we created to support our experiments. Section 5 details system evaluation, the results of which are analyzed in Section 6. Future work is outlined in Section 7 before the conclusion.

2 Email Thread Summarization

The problem of summarizing email threads is technically challenging because email is qualitatively different from newswire text, the focus of much research effort by computational linguists. Unlike single-author journalistic writings, email threads capture the conversation among two or more individuals, across both time and space. However, the asynchronous nature of these exchanges distinguishes it from spoken dialog, an area in which there is some previous work (Zechner, 2002).

Unlike newswire text, which is meant for general consumption by a wide audience, emails are only intended for their recipients. As a result, they are much more informal and often rely on shared contexts, specialized sublanguages, and other implicit cues to facilitate efficient communication. Furthermore, email is often embedded in a larger organizational context which we cannot directly observe from the texts alone, as in the simple case of collaboration between two colleagues that occurs partially over email and partially in face-to-face meetings. Finally, email is not subjected to the careful editorial process that news articles are, thus making typos, incomplete sentences, and other grammatical oddities much more prevalent.

Email represents an instance of "informal" text—a broader genre that includes conversational speech, blogs, instant and SMS messages, etc. Interest in automated processing techniques for informal media has been growing over the past few years for many reasons. There is the recognition that an increasingly large portion of our society's knowledge is captured in informal communication channels. Serious research in this area is facilitated by the availability of large collections and the falling cost of computational and storage resources. Finally, informal media push the frontiers of human language technologies by forcing researchers to develop more general and robust algorithms that are adaptable to different domains and tasks.

An email thread is a collection of messages that form a multi-party conversation. Generally, a thread will consist of an initial email message and subsequent responses to it. We describe a first attempt at email thread summarization on a challenging corpus—the Enron email collection. This represents among the first automatic summarization work of its type on this particular corpus.[1] As a

[1] We would like to thank an anonymous reviewer for pointing out another work on Enron email summarization (Carenini et al., 2007). We note that this related work was published after the initial submission of our article for peer review.

first step, we have adapted existing document summarization techniques to tackle this problem. Our initial foray hopefully paves the way for future advances in the area.

The general problem of email summarization is not new. Previous work has employed a corpus of emails sent among the board members of the ACM chapter at Columbia University (Rambow et al., 2004). Researchers have also examined summarization of archived discussion lists (Nenkova and Bagga, 2003; Newman and Blitzer, 2003; Wan and McKeown, 2004), email gisting by means of noun-phrase extraction (Muresan et al., 2001), thread-driven email summarization (Lam et al., 2002), and summarization of other informal media (Zechner, 2002; Maskey and Hirschberg, 2003; Zhou and Hovy, 2006). The recent work of Carenini et al. (2007) examines extractive approaches to summarization on Enron data that leverage graphs defined by quoted texts. Our work adopts a more abstractive approach and focuses on a slightly different set of research questions.

In addition to the problem of generating content, there are also several presentational issues associated with email thread summarization. The usual practice of presenting an undifferentiated segment of prose does not appear to be a good idea, since email comes with a great deal of metadata (e.g., sender, recipients, time, etc.). Presentational issues potentially confound evaluations of content since associated metadata may be required for the interpretation of system output.

Finally, evaluation issues in general present challenges for text summarization. Are established methodologies for existing tasks applicable? Do automatic metrics such as ROUGE (Lin, 2004) predict human judgments? If not, are there other alternatives? Despite these open research questions, we employ existing evaluation methodologies due to the lack of alternatives. In our specific case, evaluation is rendered more complex by the highly technical domain of energy trading—we return to discuss these issues in Section 6.

3 Summarization Framework

We have developed two different approaches to the problem of email thread summarization that leverage existing work. In one case, each message can be considered a "document" in a multi-document summarization task. In the same way that current systems are given a number of documents about a topic and asked to generate a summary, this approach treats each email as a document "about" the topic. We term this the Collective Message Summarization (CMS) approach. In contrast, we can take an alternative view and treat email thread summarization as the task of generating successive single-document summaries. That is, we generate a short summary for each individual email, and then aggregate the output to form a complete summary for the thread. We call this approach Individual Message Summarization (IMS).

Prima facie, both approaches have advantages and disadvantages. While IMS will faithfully preserve thread structure, it is fairly evident that not all messages in a thread are equally important. Thus, the approach runs the risk of over-representing messages that do not contain important content. Furthermore, since summary length is largely determined by thread length, system output must be further processed to generate a summary of a desired length. The CMS approach has the opposite problems: although summary length is easier to control, it is more difficult to convey thread structure (and hence the conversational nature of email). There is little guarantee that content in different parts of the thread will be represented (but this may not be a problem).

Our basic summarization architecture is shown in Figure 1—this describes both our previous single-document and multi-document summarization systems, which we adapt for IMS and CMS. Instead of a purely extractive approach, we use sentence compression to remove unimportant fragments of otherwise important sentences. One salient feature of our work is that the sentence compression module generates multiple variants of source sentences. The advantage of this approach is that it provides the necessary flexibility to accommodate complex interactions between relevance and redundancy that cannot be captured in a single compression. Downstream processes that have access to more information are

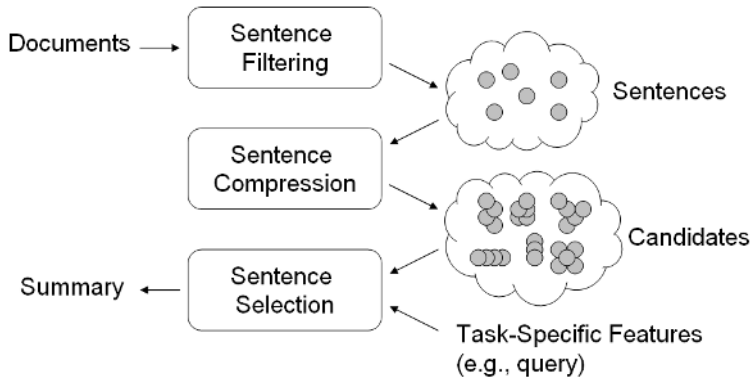

Figure 1: The basic architecture of our summarization framework.

capable of making better decisions on the choice of a final compression; this approach is also espoused by Vanderwende et al. (2006).

Specifically, a sentence selector builds the final summary by choosing among the candidates, based on features propagated from the sentence compression method, features of the candidates themselves, and features of the present summary state. In this work, we do not examine the filtering process in detail; instead, only very simple approaches are employed, e.g., retain first n sentences. Finally, we note that summaries can be influenced by task-specific considerations (e.g., query-focused vs. generic summaries)—although this is not relevant in our current task formulation.

We have previously implemented both single-document and multi-document summarization systems built around this architecture. Our single-document summarization system is generally considered the state of the art and has performed very well in previous DUC evaluations (Zajic et al., 2004). Due to the complexity of the parameter optimization process, our multi-document summarization system has been more difficult to perfect. It is currently a "middle of the pack" system based on recent DUC evaluations—not significantly better or worse than most systems (Zajic et al., 2006; Madnani et al., 2007).

In published work, we have examined two approaches to sentence compression: one based on linguistically-motivated rules that operate on parse trees ("parse-and-trim") and the other based on a noisy-channel model implementation using HMMs. We apply both methods to the problem of email thread summarization. These two compression techniques represent different tradeoffs that we think are particularly salient for informal text. Since the trimming approach requires an accurate parse tree to work with, we anticipate that parse errors will be a major source of concern because modern statistical parsers are generally trained on newswire text and perform poorly on out-of-genre text. On the other hand, we expect that the purely-statistical HMM-based approach will be more robust to text from different genres. However, given a workable parse, the trimming approach is more likely to generate fluent text since it performs linguistically-meaningful manipulations of the source text, whereas the HMM approach models language only at the level of n-grams, often resulting in disfluent text.

The sentence selector in our framework iteratively chooses from compressed variants of source sentences to generate a final summary. We adopt a weighted feature-based approach where the parameters have been tuned on test data from previous DUC evaluations. Features are either static or dynamic, in that dynamic features are recomputed after the inclusion of each additional sentence in the final summary. Such features take into account redundancy with respect to the current summary, the distribution of documents from which sentences have been selected, etc. Static features include values propagated from the sentence compression algorithm, keyword similarity measures computed with respect to the working set of documents, etc. More details are given in (Zajic et al., 2007).

Human	Avg. Size (words)
1	126.4
2	52.2
3	135.7
4	136.6
5	241.5
Avg.	138.5

Table 1: Average length in words of the reference summaries for the email threads in our test collection.

I know that you do not need numbers until late next month, but I thought you might want an early look at May.

One number is particularly interesting: VaR for the Total Return Swaps. You will notice that it decreased substantially from about $20 million in March to about $8 million in May. We had several deals that expired (Churchill, Piti Guam, and Blackbird), reducing risk, and only one new one (Motown). Most importantly, we cut back on our exposure to Rhythms from 5.4 million shares in April to 4.7 million in May, and the stock price continued to fall from $36 to $21 to $16 per share (the less the investment is worth, the less we can lose in it).

We will send you June numbers as we collect them.

Figure 2: Text of an email from thread 6.

4 Building a Test Collection

We explored the email thread summarization problem using messages from the Enron corpus, which consists of approximately half a million emails from the folders of 151 Enron employees. This corpus represents the largest available collection of real-world email traffic, and offers researchers a unique glimpse into the nature of corporate communication and the illegal activities that eventually led to the downfall of the company. Already, many topics have been explored using this data, including name reference resolution (Diehl et al., 2006), topic and role discovery (McCallum et al., 2005), and social network analysis (Diesner et al., 2005). Along with (Carenini et al., 2007), this work is among the first attempts at summarization on this collection.

Since there were no existing resources to support a summarization task, we had to create a test collection ourselves. This was performed by a master's student in the College of Information Studies at the University of Maryland, who spent several months learning about energy trading and examining the data (as part of a larger project on knowledge discovery). Our test corpus was created with the end application in mind: she first developed information needs that users might have. Using a baseline retrieval engine built on Lucene, she manually searched for relevant threads and selected them for summarization.

In total, ten threads were selected for inclusion in our test collection. The threads range in size from 3 to 30 emails, with an average size of 12.6 emails per thread. In addition to writing a reference summary for each of the threads herself, our Enron expert recruited and trained four additional individuals (also master's students in the College of Information Studies) to generate reference summaries. Since these additional subjects had no prior domain knowledge, sessions began with an overview of energy trading and other background necessary to understand the content of the threads (which took a few hours).

No length limit was placed on these human reference summaries.

Ultimately, we obtained five reference summaries for each of the ten manually-selected threads. Table 1 shows the average lengths in words of the references. Summarizer 5 was the Enron expert who assembled the threads and also trained the other human summarizers; she also had the most in-depth understanding of the domain.

Consider the sample email in Figure 2, selected from thread 6. It is apparent that the email summarization task on this dataset is very difficult, even for humans. It is also clear that one must be familiar with the arcane world of energy trading in order to comprehend the message contents. Furthermore, this highly technical domain uses plenty of jargon that is not typically found in newswire text.

All email messages were pre-processed before they were presented to our summarization systems. These processes included removal of headers and attachments. Repetitions of text from earlier messages ("quoted text") were also eliminated. We attempted to present our summarization systems with text as clean as possible.

5 Evaluation

We conducted a variety of experiments to explore the problem of email thread summarization. The system task was to generate a fixed-length summary of the thread. Two different lengths were considered: 100-word summaries, an arbitrary cutoff, and 140-word summaries, which roughly correspond to the average length of our human references.

In particular, we focused on two variables:

- Approach: IMS vs. CMS
- Compression method: linguistically-motivated rules operating on parse trees ("Trim" for short) vs. a purely statistical approach based on HMMs

In the IMS approach, our system selected the best compression of the first non-trivial sentence in each email message under 75 characters, where the first non-trivial sentence is the first sentence that is not a salutation or a content-free opening line. The character limit was adopted from previous single-document summarization task definitions. In the CMS approach, the sentence selector had access to the first five sentences of each message in the thread as well as the multiple compressions of these sentences.

Summaries generated by the IMS approach required one additional processing step. Since the length of the summaries is determined by the size of the thread, there are different ways of truncating the output in order to meet the desired length restrictions. We experimented with two different variants of the IMS approach:

- "Initial"—thread summaries consist of message summaries starting from the first message. That is, text from later messages is dropped.
- "Final"—thread summaries consist of message summaries starting from the last message. That is, text from earlier messages is dropped.

These two alternatives make opposite hypotheses about the conversational structure of email threads. One supposes that earlier messages are more important because they, for example, set up the context of the topic under discussion. The other supposes that later messages are more important because they, for example, involve the resolution of the particular issue at head. We explored

Approach	Compression	Variant	100 words	140 words
IMS	none	initial	0.0489	0.0571
IMS	HMM	initial	0.0315	0.0325
IMS	Trimmer	initial	0.0421	0.0431
IMS	none	final	0.0505	0.0541
IMS	HMM	final	0.0316	0.0328
IMS	Trimmer	final	0.0411	0.0434
CMS	none	-	0.0659	0.0918
CMS	HMM	-	0.0508	0.0782
CMS	Trimmer	-	0.0453	0.0681
Humans			**100 words**	**140 words**
Human 1			0.0770	0.0883
Human 2			0.0187	0.0222
Human 3			0.0963	0.1145
Human 4			0.0709	0.0929
Human 5			0.0955	0.1265

Table 2: ROUGE-2 recall using jackknifing: results from different experimental conditions.

both hypotheses experimentally. Note that additional truncation was not necessary with the CMS approach since summary length is directly controlled by the sentence selector, which iteratively selected candidates until the desired length had been achieved.

Finally, we compared our systems against three separate baselines: IMS without sentence compression, "initial" variant; IMS without sentence compression, "final" variant; CMS without sentence compression. The matrix setup of the complete set of experimental conditions allowed to us answer two research questions independently: Which is better, CMS or IMS, and, does sentence compression help in email thread summarization?

System output was automatically evaluated using ROUGE with the five reference summaries described in the previous section. Table 2 shows ROUGE-2 recall scores, with jackknifing. Note that since none of the threads were used in system development, they can be considered blind held-out test data. For our sentence selector, we simply employed default parameters trained on data from previous DUC evaluations. In addition, Table 2 shows the performance of the human summarizers so that we can quantify potential upper-bound performance. For fair comparison, human summaries were also truncated to the appropriate lengths (either 100 or 140 words). The bar graphs in Figures 3 and 4 present a different view of the results; error bars denote the 95% confidence intervals, which provide the reader a method for assessing the statistical significance of the differences.

For reference, sample output from the CMS approach for thread 6 (100-word condition) is shown in Figure 5—Trimmer output on top, HMM output in the middle, and no compression on the bottom. Following Rambow et al. (2004), we sort system output chronologically and prepend the author name and a timestamp to each email. Since sentence breaks are often not explicitly marked, we add a special break symbol (∘) for clarity. The insertion of metadata occurs purely for the purposes of presentation (and were not included in the ROUGE evaluations). Although the system output may be difficult to understand, we note that the source text is just as difficult to comprehend due to the prevalence of domain jargon (see Figure 2). Due to space limitations, comparable output from the IMS approach is not shown.

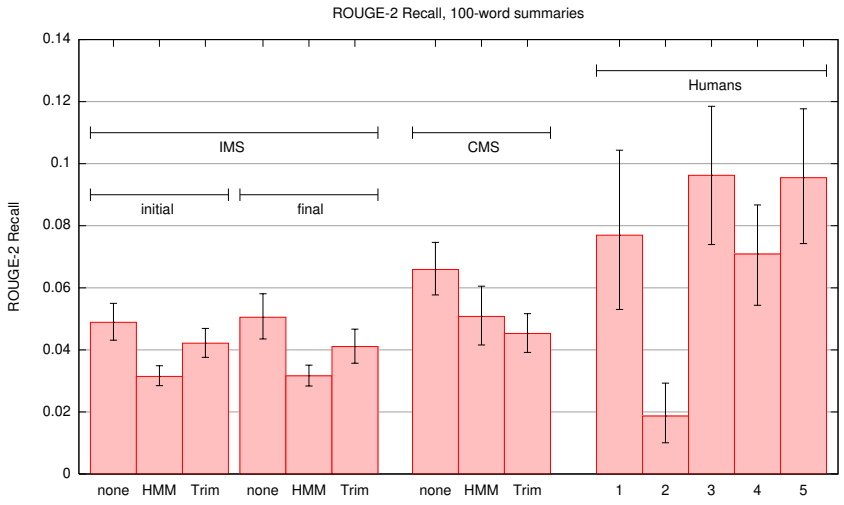

Figure 3: ROUGE-2 scores for 100-word summaries, under different experimental conditions. Results are grouped by the major approach (IMS vs. CMS).

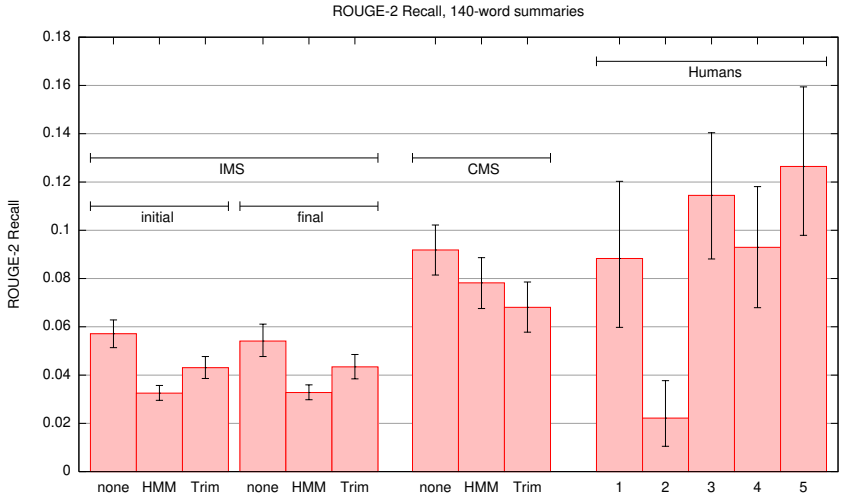

Figure 4: ROUGE-2 scores for 100-word summaries, under different experimental conditions. Results are grouped by the major approach (IMS vs. CMS).

∘ **Eugenio Perez (6/26/2000 06:40):** I know that you do not need numbers until late next month but I thought you might want an early look at May
∘ **Eugenio Perez (10/27/2000 02:50):** The good news was that September VaRs is little changed from the June numbers
∘ **Eugenio Perez (1/25/2001 09:34):** Gary and Georgeanne let me know that all but 487 shares of EOG are hedged (without the EOG leg the Cerberus total return swap is really only a loan and its VaR is about $ 500 thousand)
∘ **Eugenio Perez (2/2/2001 02:14):** AA informed me that the hedges on the New Power Company warrants that were monetized in the Hawaii 125 0 McGarret swaps were put on October 4 not in September

∘ **Eugenio Perez (6/26/2000 06:40):** you might want early look ∘ it decreased substantially ∘ the investment is worth
∘ **Eugenio Perez (10/27/2000 02:50):** New Power Company went public ∘ warrants we inserted are hugely. ∘ swaps will probably be over $30 million.
∘ **Eugenio Perez (1/25/2001 09:34):** VaR fell and $18 million ∘ Cerberus total return swap is really only a loan ∘ natural gas prices are up so much ∘ we can potentially lose
∘ **Eugenio Perez (1/31/2001 08:25):** Please disregard previous versions.
∘ **Eugenio Perez (2/2/2001 02:14):** hedges that monetized 125-0 McGarret swaps put
∘ **Eugenio Perez (2/6/2001 02:20):** we created by granting options ∘ we have long term contracts to remove variability of revenues ∘ the contracts expire ∘ for total return swaps fell from $34 to $28 million.
∘ **Adarsh Vakharia (2/8/2001 09:37):** it is little hedged by Phantom swap ∘ Regards, Adarsh and Eugenio

∘ **Eugenio Perez (6/26/2000 06:40):** One number is particularly interesting: VaR for the Total Return Swaps.
∘ **Eugenio Perez (10/27/2000 02:50):** VaR for the total return swaps will probably be over $30 million in October.
∘ **Eugenio Perez (1/25/2001 09:34):** Regards, Eugenio
∘ **Eugenio Perez (1/31/2001 05:17):** ave recalculated their VaR, and I have enclosed it above. ∘ Regards, Eugenio
∘ **Eugenio Perez (1/31/2001 08:25):** Merchant Assets has re-revised numbers. ∘ Please disregard previous versions. ∘ Regards, Eugenio
∘ **Eugenio Perez (2/2/2001 02:14):** Accordingly, I have recalculated total return swap VaR for September. ∘ Regards, Eugenio
∘ **Adarsh Vakharia (2/8/2001 09:37):** Georgeanne told Eugenio about another Enron stock swap. ∘ This one is very large, with an exposure to 12 million shares (though it is a little hedged by the Phantom swap).
∘ **Eugenio Perez (2/9/2001 03:04):** The Jedi swap VaR was moved from non-trading securities to trading securities. ∘ Regards, Eugenio

Figure 5: Output from the CMS approach for thread 6 (100-word condition): Trimmer (top), HMM (middle), no compression (bottom).

6 Discussion

It is readily apparent from our experiments that summarization of email threads from the Enron corpus is very challenging, even for humans. Overall, we observe significant variance in human performance on this task. The primary difficulty comes from the need for specialized domain knowledge in order to comprehend the email messages. Recall that to generate our reference summaries, the domain expert (Human 5) recruited and trained four other subjects for the task. These training sessions, which lasted a few hours, may not have been sufficient. For example, Human 2 found the task so difficult that one of her summaries was simply the following statement: "This thread is very hard to follow. Not sure what they are attempting to convey." This was reflected in the ROUGE score, which was significantly lower than many of our automatically-generated summaries.

Nevertheless, our experiments on this small test collection point to two major findings: that CMS is a better approach to email thread summarization than IMS, and that current sentence compression techniques do not improve summarization performance in this genre.

We conclude from the bar graphs in Figures 3 and 4 that CMS is a more effective approach to email thread summarization than IMS. For many of the contrastive pairs, CMS achieves significantly higher scores than IMS. In other words, treating email thread summarization as a multi-document summarization problem leads to higher performance than treating the task as an independent series of single-document summarization problems. This is a significant finding, as previous evaluations have shown that taking the first n words of a document is a tough baseline to beat in single-document summarization tasks (Over and Liggett, 2002). Since IMS is a straightforward extension of this baseline to email threads, we expected its performance to be competitive also.

The other major finding from this study is that current sentence compression techniques do not improve summarization performance. For both IMS and CMS, the highest scores are achieved with no sentence compression—in some cases, compression actually results in significantly lower scores. As this result is inconsistent with previous work in the newswire domain, where many groups have confirmed that sentence compression improves summarization performance (Blair-Goldensohn et al., 2004; Conroy et al., 2006), out-of-genre issues are likely the culprit. For Trimmer, proper compression depends on correct parse trees, and parsers trained on newswire text (like the one we use) are likely to make many errors. Similarly, language models for our HMMs were induced from newswire text, which obviously has different distributional characteristics. Using ill-adapted compression techniques appears to be a liability in this particular application. This result points to a need to focus on domain-adaptation issues for sentence compression before they are likely to have a positive impact on email thread summarization.

It is also interesting to note that although both sentence compression techniques result in lower scores than having no sentence compression, HMM outperforms Trimmer in the IMS case, whereas the opposite appears true in the CMS case. As a point of reference, we have previously found that Trimmer performs better than HMM for summarization of written news in both single-document and multi-document conditions (Zajic et al., 2007). Once again, this inconsistency with respect to previous results suggests that out-of-genre issues pose significant challenges in email thread summarization.

7 Future Work

Our exploration of email thread summarization on the Enron corpus has helped us better understand the nature of the problem, thus paving the way for future work.

First, we need a more precise definition of the task. What exactly is a summary of an email thread? Should such summaries be informative or indicative? (Probably a mixture of both.) How should the conversational nature of email threads be conveyed? (Probably by explicitly marking participants and turn-taking, as we have.) What is the summary itself used for? We have framed the problem in the

context of a search application, but no doubt the task can be cast in different ways. Furthermore, the highly technical nature of the domain makes developing test collections difficult, since experts are required to generate reference summaries. Our strategy of training non-experts was moderately successful, but the paucity of domain expertise remains.

Our experiments rely on the assumption that ROUGE performance correlates with human preferences. Although this is generally accepted in the summarization literature, and ROUGE scores are widely reported in lieu of opinions from human assessors, the extension of this automatic metric across domains has not been established. Previous work in email summarization have used sentence-level precision and recall to quantify performance (Rambow et al., 2004), but this is applicable only in a purely extractive framework. However, there are few other options, as manual evaluation is usually prohibitively expensive and too slow for system development. Work on alternative evaluation metrics, particularly extrinsic ones, is sorely needed to enable the advancement of summarization technology.

Recall that the IMS approach can be separated into two variants, "initial" and "final", which met the fixed length restrictions by removing words from the end of the thread and from the beginning of the thread, respectively. Experiments did not reveal any significant differences in performance between the two approaches, suggesting that there is value in content both at the beginnings and ends of email threads. These two variants, however, highlight the more general problem of determining summary length. Following many standard summarization tasks, we specified fixed-length summaries (either 100 or 140 words). However, since some email threads are longer than others, the compression ratio is highly variable. In general, the CMS approach is not affected by desired summary length, since the sentence selector iteratively chooses candidates until the desired length is achieved. However, the IMS approach is less sophisticated in controlling summary length—threads with few emails might result in a summary shorter than the desired length and threads with many emails might result in a summary that is arbitrarily truncated. This points to another issue that requires more exploration.

Finally, this work highlights the importance of genre adaptation. Both our linguistic and statistical sentence compression techniques did not appear to perform well on Enron data, due to out-of-genre issues. Both are hampered by their reliance of newswire training data, although in different ways—more work is needed to understand how these two approaches degrade and how to improve them.

8 Conclusion

We believe that the biggest contribution of this work lies in making inroads to a difficult and important problem. The fact that the Enron corpus is representative of many organizational email collections lends realism to the task that we have framed. Our initial explorations have probed this large problem with existing single-document and multi-document summarization techniques. In addition to establishing some benchmark baselines for performance, we have identified a number of challenges that lie ahead. We are encouraged by these initial results, and hope that this works provides a foundation upon which others can build.

9 Acknowledgments

This work has been supported in part by the Joint Institute for Knowledge Discovery at the University of Maryland. We would like to thank Erin Greenwell, who lead the effort to develop our test collection, and our human summarizers—this work would not be possible without their hard work. In addition, our thanks go out to an anonymous reviewer for helpful comments and suggestions. DZ wishes to thank Naomi for her support. BD would like to thank Steve, Carissa, and Ryan for their energy enablement. JL would like to thank Esther and Kiri for their kind support.

References

Sasha Blair-Goldensohn, David Evans, Vasileios Hatzivassiloglou, Kathleen McKeown, Ani Nenkova, Rebecca Passonneau, Barry Schiffman, Andrew Schlaikjer, Advaith Siddharthan, and Sergey Siegelman. 2004. Columbia University at DUC 2004. In *Proceedings of the 2004 Document Understanding Conference (DUC 2004) at HLT/NAACL 2004*, pages 23–30, Boston, Massachusetts.

Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zho. 2007. Summarizing email conversations with clue words. In *Proceedings of the Sixteenth International World Wide Web Conference (WWW2007)*, pages 91–100, Banff, Alberta, Canada.

John M. Conroy, Judith D. Schlesinger, Dianne P. OLeary, and Jade Goldstein. 2006. Back to basics: CLASSY 2006. In *Proceedings of the 2006 Document Understanding Conference (DUC 2006) at HLT/NAACL 2006*, New York, New York.

Nick Craswell, Arjen P. de Vries, and Ian Soboroff. 2005. Overview of the TREC-2005 enterprise track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, Maryland.

Edward Cutrell, Daniel C. Robbins, Susan T. Dumais, and Raman Sarin. 2006. Fast, flexible filtering with Phlat—personal search and organization made easy. In *Proceedings of SIGCHI 2006 Conference on Human Factors in Computing Systems (CHI 2006)*, pages 261–270, Montréal, Québec, Canada.

Christopher P. Diehl, Lise Getoor, and Galileo Namata. 2006. Name reference resolution in organizational email archives. In *Proceedings of the 2006 SIAM Conference on Data Mining*, Bethesda, Maryland.

Jana Diesner, Terrill Frantz, and Kathleen Carley. 2005. Communication networks from the Enron email corpus. *Journal of Computational and Mathematical Organization Theory*, 11(3):201–228.

Bonnie J. Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. 2005. A methodology for extrinsic evaluation of text summarization: Does ROUGE correlate? In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 1–8, Ann Arbor, Michigan.

Derek Lam, Steven L. Rohall, Chris Schmandt, and Mia K. Stern. 2002. Exploiting e-mail structure to improve summarization. Collaborative User Experience Technical Report 02-02, IBM T.J. Watson Research Center.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004) at ACL 2004*, pages 74–81, Barcelona, Spain.

Nitin Madnani, David Zajic, Bonnie Dorr, Necip Fazil Ayan, and Jimmy Lin. 2007. Multiple Alternative Sentence Compressions for automatic text summarization. In *Proceedings of the 2007 Document Understanding Conference (DUC 2007) at NLT/NAACL 2007*, Rochester, New York.

Inderjeet Mani, Gary Klein, David House, and Lynette Hirschman. 2002. SUMMAC: A text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

Sameer Raj Maskey and Julia Hirschberg. 2003. Automatic summarization of broadcast news using structural features. In *Proceedings of Eurospeech 2003*, Geneva, Switzerland.

Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. 2005. The author–recipient–topic model for topic and role discovery in social networks, with application to Enron and academic email. In *Proceedings of the Workshop on Link Analysis, Counterterrorism and Security at the 2005 SIAM International Conference on Data Mining.*

Smaranda Muresan, Evelyne Tzoukermann, and Judith L. Klavans. 2001. Combining linguistic and machine learning techniques for email summarization. In *Proceedings of the ACL/EACL 2001 Workshop on Computational Natural Language Learning (ConLL)*, pages 1–8, Toulouse, France.

Ani Nenkova and Amit Bagga. 2003. Facilitating email thread access by extractive summary generation. In *Proceedings of 2003 Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, Borovets, Bulgaria.

Paula S. Newman and John C. Blitzer. 2003. Summarizing archived discussions: A beginning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI 2003)*, Miami, Florida.

Paul Over and Walter Liggett. 2002. Introduction to DUC-2002: An intrinsic evaluation of generic news text summarization systems, available at http://duc.nist.gov/pubs.html#2002.

Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. 2004. Summarizing email threads. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004): Short Papers*, pages 105–108, Boston, Massachusetts.

Lucy Vanderwende, Hisami Suzuki, and Chris Brockett. 2006. Microsoft Research at DUC2006: Task-focused summarization with sentence simplification and lexical expansion. In *Proceedings of the 2006 Document Understanding Conference (DUC 2006) at HLT/NAACL 2006*, New York, New York.

Stephen Wan and Kathy McKeown. 2004. Generating overview summaries of ongoing email thread discussions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 549–555, Geneva, Switzerland.

David Zajic, Bonnie Dorr, and Richard Schwartz. 2004. BBN/UMD at DUC-2004: Topiary. In *Proceedings of the 2004 Document Understanding Conference (DUC 2004) at NLT/NAACL 2004*, pages 112–119, Boston, Massachusetts.

David Zajic, Bonnie Dorr, Jimmy Lin, and Richard Schwartz. 2006. Sentence compression as a component of a multi-document summarization system. In *Proceedings of the 2006 Document Understanding Conference (DUC 2006) at NLT/NAACL 2006*, New York, New York.

David Zajic, Bonnie Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-Candidate Reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management*, 43(6):1549–1570.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Liang Zhou and Eduard Hovy. 2006. On the summarization of dynamically introduced information: Online discussions and blogs. In *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, Stanford, California.