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

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 our techniques on the Enron collection—a very challenging corpus because of the highly technical language. Results suggest that CMS represents a better approach and additional findings pave the way for future explorations.

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 well-defined 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, for example, (Dumais et al., 2003; Craswell et al., 2005). 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 access application.

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, and the other treats the problem as a variant of multi-document summarization. Both approaches involve selecting important sentences from email messages and compressing them (i.e., removing unimportant portions). Our implemented systems were evaluated using data from the Enron collection, using a small manually-created test corpus. Results show promise, although we discuss the challenges associated with both this task and the Enron corpus.

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 in-
individuals, across both time and space. However, the asynchronous nature of these exchanges distinguishes it further from spoken dialog—another area that has received attention.

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 dataset. To our knowledge, this represents the first summarization study of its type on this particular collection. As a first step, we have adapted existing document summarization techniques to tackle this problem. This first foray 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 (Maskey and Hirschberg, 2003; Zechner, 2002; Zhou and Hovy, 2006). However, our work is unique in examining the Enron collection.

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. Are established methodologies for existing summarization 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 processes due to the lack of alternatives. In our specific case, evaluation is rendered more complex by our highly technical domain—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 traditional 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 aggre-
Figure 1: The basic architecture of our summarization framework.

We aggregate the output to form a complete summary. 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 obvious 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 given 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 capable of making better decisions on the choice of a final compression. 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. 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.¹

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.

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

¹References have been omitted to facilitate blind review.
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 chosen, 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 (Anonymous).

4 The Data

We explored the email thread summarization problem using messages from the Enron dataset, 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). However, to our knowledge this work represents the first attempt 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 library and information science (LIS) 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 LIS) 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.

In the end, 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 also the Enron expert who assembled the threads and trained the other subjects, and had the greatest understanding of the domain.

Consider the sample email in Figure 2, selected from thread 6. It is apparent that the email thread summarization task on this dataset is very difficult, even for humans. It is obvious that one must be familiar with the arcane world of energy trading in order to comprehend the message contents. Furthermore, this specialized and 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”) was 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 attempted here was to generate one hundred word summaries of threads.

In particular, we focused on two variables:

- Approach: IMS vs. CMS
- Compression method: linguistically-motivated rules operating on parse trees (“Trimmer” for short) vs. HMM

In the IMS approach, our system chose 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 text in the entire email thread.

Summaries generated by the IMS approach required one additional processing step. Since the length of summaries is determined by the size of the thread, we simply retained the first 100 words if the system output was longer than the desired length. Note that additional truncation was not necessary with the CMS approach since summary length is directly controlled by the sentence selector, which iteratively chooses candidates until the desired length has been achieved.

Finally, we tested our systems against the following baseline: the first 75 characters of each email message were selected to form a summary. This essentially represented an IMS approach, except without any sentence compression. Since the length of this baseline output is also dependent on thread size, we discarded all but the first 100 words.

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 100 words. Figure 3 offers a different view of the results, with the different conditions sorted in increasing order of ROUGE-2 scores. Error bars denote the 95% confidence intervals.

For reference, sample output from the CMS approach for thread 6 is shown in Figure 4—Trimmer output on top and HMM on the bottom. Following Rambow et al. (2004), we sort system output

<table>
<thead>
<tr>
<th>Run</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS Trimmer</td>
<td>0.0421</td>
</tr>
<tr>
<td>IMS HMM</td>
<td>0.0315</td>
</tr>
<tr>
<td>CMS Trimmer</td>
<td>0.0453</td>
</tr>
<tr>
<td>CMS HMM</td>
<td>0.0508</td>
</tr>
<tr>
<td>baseline</td>
<td>0.0489</td>
</tr>
<tr>
<td>Human 1</td>
<td>0.0770</td>
</tr>
<tr>
<td>Human 2</td>
<td>0.0187</td>
</tr>
<tr>
<td>Human 3</td>
<td>0.0963</td>
</tr>
<tr>
<td>Human 4</td>
<td>0.0709</td>
</tr>
<tr>
<td>Human 5</td>
<td>0.0963</td>
</tr>
</tbody>
</table>

Table 2: ROUGE recall scores using jackknifing from different system runs.

Figure 3: ROUGE-2 scores for different conditions, sorted in increasing order.
<table>
<thead>
<tr>
<th>Author</th>
<th>Date</th>
<th>Email Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eugenio Perez</td>
<td>6/26/00 06:40</td>
<td>I know that you do not need numbers until late next month but I thought you might want an early look at May.</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>10/27/00 02:50</td>
<td>The good news was that September VaRs is little changed from the June numbers.</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>1/25/01 09:34</td>
<td>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).</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>2/2/01 02:14</td>
<td>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.</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>6/26/00 06:40</td>
<td>you might want early look it decreased substantially the investment is worth</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>10/27/00 02:50</td>
<td>New Power Company went public warrants we inserted are hugely. swaps will probably be over $30 million.</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>1/25/01 09:34</td>
<td>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</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>1/31/01 08:25</td>
<td>Please disregard previous versions.</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>2/2/01 02:14</td>
<td>hedges that monetized 125-0 McGarret swaps put</td>
</tr>
<tr>
<td>Eugenio Perez</td>
<td>2/6/01 02:20</td>
<td>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.</td>
</tr>
<tr>
<td>Adarsh Vakharia</td>
<td>2/8/01 09:37</td>
<td>it is little hedged by Phantom swap Regards, Adarsh and Eugenio</td>
</tr>
</tbody>
</table>

Figure 4: Output from the CMS approach: using Trimmer (top) and HMM-based (bottom).

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). For brevity, comparable output from the IMS approach is not shown.

6 Analysis

How are we to interpret these results? We note two important observations: that the task is exceedingly difficult and that the baseline seems to perform well, at least in terms of ROUGE scores.

Summarization of email threads from the Enron dataset is very challenging, even for humans. 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, subject 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 our systems’ (see Figure 3).

Overall, we observe significant variance in human performance on this task. Furthermore, it unclear that humans are actually better than machines (at least in terms of ROUGE scores). Only 2 of 5 humans scored significantly higher on ROUGE-2 recall than the best automated system, and one human performed significantly lower (Subject 2).

Our second major observation is the the baseline is highly competitive in terms of ROUGE-2 scores, beating all system variants except for CMS HMM (although many of the differences are not statistically significant). In some ways, this is not surprising, since similar baselines have been tough to beat in previous DUC evaluations. In some cases, systems did not perform better than simple baselines until a few years after researchers started tackling the problem (Over and Liggett, 2002). All things considered, we are encouraged by system performance in this first attempt at email thread summarization on the Enron corpus.

We note that the baseline is essentially a variant
of IMS that does not utilize sentence compression. Therefore, it is surprising that the baseline outperforms both IMS HMM and IMS Trimmer. We interpret this finding to suggest that our sentence compression algorithms are not functioning as expected. However, since results in summarizing newswire data have demonstrated the value of sentence compression (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.

Nevertheless, it does appear that CMS represents a better approach to email thread summarization than IMS. Our CMS HMM variant outperforms the baseline, although differences are not statistically significant. Overall, we are encouraged by the CMS performance, because the HMM variant performed better than its IMS counterpart, and the same for the Trimmer variant. In the first case, the difference was statistically significant, but not so in the second case.

We also note with interest that Trimmer does not perform significantly better than HMM in either CMS or IMS approaches for our task, even though we have demonstrated that Trimmer performs better than HMM for summarization of written news in both single-document and multi-document summarization (Anonymous 2003, Anonymous 2007). HMM-based techniques might be a more attractive choice for sentence compression in noisy environments where parser performance is compromised. However, based on our experiments, the HMM-based technique fared worse on out-of-genre text in the IMS case. Statistical methods may not be as robust as we have previously thought, given that they still rely on language models to capture fluency. Since many n-grams in the Enron collection are simply not observed on newswire training data, these language models may not be portable.

Recall that with the IMS approach, only the first 100 words of system output were retained. For longer threads, this resulted in summaries that only covered email messages towards the beginning. This might be problematic, since we expect messages towards the end to contain important information also. For example, the final messages in a thread might discuss the resolution of a particular issue. To test this hypothesis, we tried selecting words from the end of system output, for both the IMS and baseline cases. Unfortunately, results were inconclusive, as ROUGE scores remained essentially unchanged.

7 Future Work

Our exploration of the email thread summarization on the Enron dataset 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.

Finally, this work highlights the importance of genre adaptation. Both our linguistic and statis-
tical sentence compression techniques did not appear to perform well on Enron data, due to out-of-genre issues. Both are hampered by their reliance on newswire training data, although in different ways—more work is needed to understand how these two approaches degrade. Nevertheless, this work affirms that robustness and adaptability remain two highly-valued characteristics of text processing algorithms.

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 interested in making our test collection publicly available so that others can build on our work.

References


