Building a Dataset for Summarization and Keyword Extraction from Emails

Vanessa Loza[1], Shibamouli Lahiri[1], Rada Mihalcea[2], Po-Hsiang Lai[3]

[1] Computer Science and Engineering, University of North Texas
[2] Computer Science and Engineering, University of Michigan
[3] Samsung Research America
vanessalozaponce@my.unt.edu, shibamoulilahiri@my.unt.edu, mihalcea@umich.edu, s.lai@sta.samsung.com

Abstract

This paper introduces a new email dataset, consisting of both single and thread emails, manually annotated with summaries and keywords. A total of 349 emails and threads have been annotated. The dataset is our first step toward developing automatic methods for summarization and keyword extraction from emails. We describe the email corpus, along with the annotation interface, annotator guidelines, and agreement studies.

Keywords: Email Processing, Keyword Extraction, Summarization

1. Introduction

Email constitutes an important means of communication in our daily exchanges, used not only for personal conversation but also as a repository of corporate information. Given the overwhelming number of emails that we have to handle on an everyday basis, it is becoming increasingly important to have efficient access to the important information contained in emails.

Summarization and keyphrase extraction are two complementary techniques which, given a natural language text, extract the most important sentences and the most important words or phrases. Therefore, when applied to an email or an email thread, it is expected that these two procedures, when suitably implemented, would give us the most important sentences and words/phrases contained in them, thereby effectively giving us a "snippet" of the text and mostly reducing the time needed to read an entire email.

Once a user reads the snippet, (s)he can either choose to read the complete email/email thread, or (s)he can choose to read it later. This is similar to the current techniques for Web search, where we read the snippets to determine the purported "value/relevance" of a particular web page rather than going inside each and every web page to determine its value. So effectively, our technology will make the task of processing email more efficient, by assisting people to prioritize email and by giving people a choice between reading an email right away, removing it, or postponing it for future.

Several datasets have been released for general-purpose summarization and keyword extraction (Hasan and Ng, 2010), but very few of them specifically deal with emails. Emails have a special graph structure (Carenini et al., 2007) that warrants more intricate treatment than that needed by other types of text documents. The only email summarization corpus we are aware of is due to (Ulrich et al., 2008). This corpus (BC3) comprises 40 email threads (3222 sentences) with annotations for extractive and abstractive summarization, speech act, meta sentences, and subjectivity. While important for being a path-breaker in email summarization research, the corpus is relatively small, it does not give a ranked list of extracted sentences, there is no control over the number of sentences extracted, and perhaps most importantly for our goals, the corpus does not include keywords. The only corpus for keyword extraction from emails (Turney, 2000) has never been released publicly.

This paper describes our efforts to alleviate this problem. We designed our own annotation scheme (Section 4.) and annotation interface (Section 4.2.), and annotated a large corpus of emails and threads using this scheme (Section 3.). The corpus, consisting of a total of more than 100,000 words, is available upon request, and thus it is likely to enable new research in this area.

2. Related Work

Among the very first studies of email summarization were (Muresan et al., 2001) and (Rambow et al., 2004). Muresan et al. reduced the problem of summarization to extracting important phrases from emails. They used linguistic and content features to classify noun phrases for saliency. Classification results were evaluated by a single human judge. Combining classifiers improved accuracy, and linguistic filtering was important for collecting salient noun phrases. Further, noun phrases were found to be better candidates than n-grams.

Rambow et al. (2004) on the other hand dealt with the problem of thread summarization. They cast the problem as salient sentence extraction, and used three sets of content and structural features – basic, basic+, and full – to classify thread sentences as "relevant" and "not relevant". Two independent annotators wrote an abstractive summary for each thread. Notably, annotators were not asked to generate extractive summaries.

Nenkova and Bagga (2003) and Wan and McKeown (2004) contributed further to thread summarization research. Nenkova and Bagga followed a scoring-based extractive summarization approach to generate "thread overviews" on the Pine-Info mailing list. Sentences were scored based on part-of-speech overlap with the subject line and the root message. An implicit assumption in Nenkova and Bagga's work is that topical consistency can be maintained by selecting sentences with higher part-of-speech overlap with the root message.

Wan and McKeown (2004) viewed thread summary gener-

ation as an online group decision-making process. They used thread structure and singular value decomposition (SVD) on bags of words to come up with a unique sentence-scoring mechanism. Issue detection was used to uncover the hidden dialog structure in threads. Application of centroid, SVD centroid, and oracle methods on a corpus of 300 threads gave insights into the methods that performed best for thread summarizatio.

Corston-Oliver et al. (2004) described SmartMail, a system for identifying "action items" in a message. SmartMail presents to the user a task-focused summary consisting of a list of action items. The system identified speech acts of each sentence in a message using a supervised classifier, and performed linguistic post-processing to frame sentences as task descriptions. A big corpus of 15,741 emails was constructed, with each sentence being represented by 53,000 features. Linear SVM classifiers were trained to identify "task" sentences. Finally, task sentences were leveraged to obtain logical forms and generate task descriptions.

Zajic et al. (2007) introduced *Multi-candidate Reduction* as a framework for abstractive multi-document summarization. The framework filters sentences and compresses them in two ways – a "parse-and-trim" approach, and a Hidden Markov Model approach. This framework was used to summarize email threads in (Zajic et al., 2008). Each thread was summarized in two ways – regarding constituent emails as separate documents (*Individual Message Summarization*), and treating the whole thread as a single large document (*Collective Message Summarization*). The authors, while manually building a test collection of ten threads from the Enron corpus (Klimt and Yang, 2004), remarked that "the email summarization task on this dataset is very difficult, even for humans."

Carenini et al. (2007) ranked sentences in a thread using *clue words*. They constructed a *Fragment Quotation Graph* to capture the flow of conversation in a thread, and used this graph to score each sentence. 20 different Enron threads were selected to build a test corpus. The authors' approach (CWS) outperformed two state-of-the-art baselines – MEAD and RIPPER – on this test set. The CWS algorithm was extended in (Carenini et al., 2008) to incorporate a *Sentence Quotation Graph*, and semantic similarity between sentences.

Murray et al. (2010) introduced *interpretation* and *transformation* for summarizing emails. In the interpretation step, an ontology is populated by entities and relationships mentioned in the email. The ontology can be learned very accurately with classifiers trained on a large set of features. In the transformation step, this ontology is used to generate a summary that maximizes an objective function relating sentence and entity weights.

Unlike summarization, keyword extraction from emails has received significantly less attention. The only research we are aware of is due to (Turney, 2000), who treated the problem of keyword extraction in a supervised setting. He used a decision tree (C4.5), and a genetic-algorithm-based classifier (GenEx) to classify phrases in a document as keyphrase or not. Turney's datasets include a corporate email corpus with 311 documents. This corpus was not made public.

3. The Email Corpus

For our email collection, we primarily used the Enron dataset (Klimt and Yang, 2004), which is a large collection of email messages made public during the legal investigation concerning the Enron corporation.[1] The raw Enron corpus contains 619,446 messages belonging to 158 users. We used the Enron Corpus prepared by CALO,[2] which does not contain attachments. Moreover, some messages have been removed following the request of affected employees. This dataset consists of 150 mailboxes, each of them containing a folder distribution specific to each employee. Among the variety of topics discussed over the collection, mainly energy trading, we also found a considerable number of emails representing private and personal communication between employees, employees and friends, or employees and their family. We thus decided to use the Enron Corpus as a source for both private and corporate emails.

To select the emails to be included in the "single emails" collection, we used line counting to determine the emails that met our selection criteria. Only the lines that were not part of the header were considered. The complete text included greetings and signature, as well as some privacy notes at the bottom. As a preselection step, we only considered the emails containing between 10 and 50 lines.

To select the emails to be included in the "threads" collection, we began with the list of all the files containing more than one email. From this group we counted the number of emails included in each file. Finally, every file with at least three emails was included in the thread group.

Emails were then classified as either corporate or private.

Corporate Emails. We use the term "corporate" to refer to any communication within work environment. Given Enron's business nature, the topics discussed extensively are generally energy market, energy trading, human resources, and legal advice. It is worth mentioning that the discussion incorporates an important amount of technical terms which are very specific to the energy field.

Private Emails. We collected two different sets of private emails. The first set is obtained from the Enron collection. To identify emails that potentially belonged to the private category among a large set of Enron emails, we used clue words in the folder names as a hint. For instance, we looked for emails classified under folders such as "personal_stuff", "family", "personal_mail". We also collected a second set of private emails, mainly provided by volunteers from their own private mailboxes. No topic was specified. Any personal references were removed from the text, and replaced with a different random word. Additionally, email addresses were replaced and modified. This set was processed similar to the Enron set of emails.

The final selection of private and corporate emails was made after manual inspection of the email content. During this manual inspection, we encountered several issues

[1] This original dataset, with a complete explanation, is available at http://www2.cs.cmu.edu/ enron/

[2] Cognitive Assistant that Learns and Organizes. http://www.ai.sri.com/project/CALO

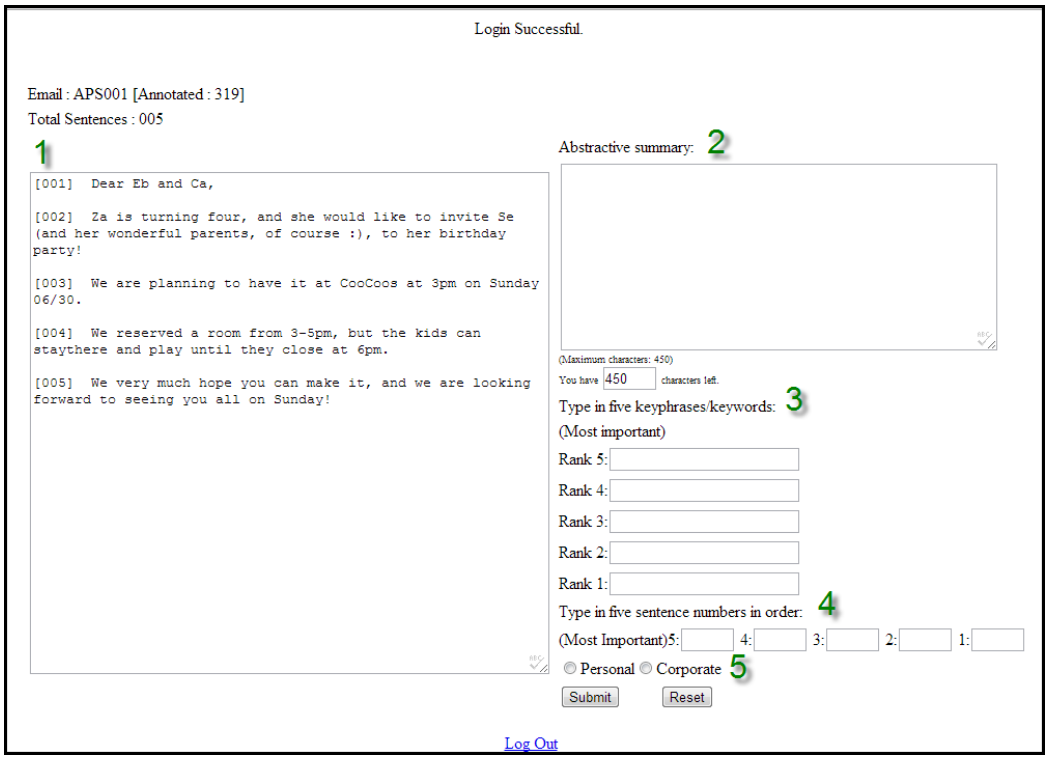

Figure 1: Screenshot of the annotation web interface. The numbers show different sections of the interface.

such as: diversity of formats, progressive emails (i.e., mails in between threads), and repeated emails (stored in different folders). The manual classification process was further complicated by the fact that some emails display topic drift, moving from a personal topic to a corporate topic, and vice versa, or moving between several topics, with threads being particularly prone to such drifts. We discarded emails that appeared to be written in a different language, subscriptions, spam mails, emails containing inappropriate content, and electronic receipts. The signature lines were also manually marked and differentiated from the email text.

The final count of emails belonging to each category is shown in Tables 1 and 2. The corpus contains a total of more than 100,000 words and close to 7,000 sentences.

Group	Type	Count	Avg. Lines	Std. Dev.
Single	Private	103	15	8
Single	Corporate	109	10	6
Thread	Private	45	26	12
Thread	Corporate	62	38	22

Table 1: Email count by type and category from Enron

Group	Type	Count	Avg. Lines	Std. Dev.
Single	Private	13	9	5
Thread	Private	17	43	35

Table 2: Email count by type and category from volunteer emails

Category	Avg. CR (%)	Std. Dev. (%)
Corporate Single	61.66	23.55
Corporate Thread	22.25	9.36
Private Single	42.79	18.82
Private Thread	16.47	6.86
Corporate	47.37	27.28
Private	34.79	20.19
Single	52.49	23.37
Thread	19.82	8.87
ALL	41.53	25.05

Table 3: Compression Ratio (CR) of different email categories from **Enron**. Lower CR values are more desirable.

Category	Avg. CR (%)	Std. Dev. (%)
Single	63.77	25.14
Thread	20.63	14.33
ALL	39.33	29.11

Table 4: Compression Ratio (CR) of different email categories from **volunteer emails**.

Note from Table 1 that on average, single emails consist of 10 to 15 lines, whereas threads consist of 26 to 38 lines. This indicates that summarization and keyword extraction from emails are indeed likely to help users focus their attention on relevant parts of emails rather than wading through a lot of unnecessary information.

Category	% containing first sentence	% containing first two sentences
Corporate Single	66.06	44.04
Corporate Thread	29.03	3.23
Private Single	33.01	13.59
Private Thread	35.56	8.89
Corporate	52.63	29.24
Private	33.78	12.16
Single	50.00	29.25
Thread	31.78	5.61
All together	43.89	21.32

Table 5: Percentage of emails that contain the first sentence and the firsht two sentences in their extractive summary, across different email categories from **Enron**.

Category	% containing first sentence	% containing first two sentences
Single	38.46	38.46
Thread	5.88	5.88
All together	20.00	20.00

Table 6: Percentage of emails that contain the first sentence and the first two sentences in their extractive summary, across different email categories from **volunteer emails**.

The emails selected for inclusion in our collection, for both the "single email" and "thread" categories are stored using an XML format. We use the XML format previously used in the BC3 Corpus (Ulrich et al., 2008), with some tags modified to meet our purposes. The format is the same for both threads and single emails, the latter being considered as threads of size one.

```
<root>
<thread>
<fileName></fileName>
<name></name>
<id></id>
<email order="">
<date></date>
<from></from>
<to></to>
<subject></subject>
<text>
<sentence id="">
</sentence>
<signature></signature>
</text>
</email>
</thread>
</root>
```

Figure 2: XML format of the email files.

Note from Figure 2 that each thread is assigned a unique identifier, and is associated with a unique filename. Further, email order within the threads is made chronological using the "order" attribute. Each sentence is assigned a separate unique identifier to ensure easy access and retrieval. The "subject" field holds the title of an email, and the "name" field holds the title of the thread. "from" and "to" fields are email addresses of the sender and the recipient, respectively. Instead of removing signatures, we included them in a separate "signature" field.

4. Annotations and Guidelines

The emails were manually annotated by two independent annotators, who generated four types of annotations for each single email or thread: an abstractive summary, a set of important sentences, a set of keyphrases, and a classification of the email as either corporate or private.

4.1. Annotation Guidelines

Abstractive summary. The abstractive summary is limited to a maximum of 450 characters, preserving the most important information of the original message. The guidelines asked that the summaries be written in the third person, regardless of the writing style of the original message. The annotators were explicitely allowed to use some excerpts from the original text, although they were encouraged to write most of the summary in their own words. While annotating the threads, signature lines could be included in the summaries, to facilitate the task of identifying the flow of the conversation.

Sentence extraction. Following the abstractive summary, the annotators were asked to select the five most significant sentences that contained the most important information in the email, and also rank the sentences in reverse order of their importance. For threads, the sentences selected as important could belong to any email in the thread. Note from Table 3 that on average private emails achieved a lower compression ratio than corporate emails. In general, Enron private threads achieved the lowest average compression ratio (16.47%). Further, Tables 3 and 4 indicate that on threads achieved a lower compression ratio on average than single emails. This indicates that in general Enron private emails are longer than Enron corporate emails, and threads are longer than single emails.
It is interesting to note that only 43.89% of Enron emails and 20.00% of volunteer emails contain the first sentence in their extractive summary (cf. Tables 5 and 6). Tables 5 and 6 also show that while single emails may benefit from having the first sentence/first two sentences in their summary, threads are less likely to benefit from such inclusion (only 5.88% volunteer threads have the first two sentences in their summary, as compared to 38.46% singles), perhaps due to the fact that threads are composed of several emails, and identifying which one is the "first" sentence is difficult.

Keyphrases. The objective of this annotation was to identify five single words and/or phrases that are the most representative for the conversation in the thread or single email. For consistency purposes, the annotators were suggested to try to select keyphrases that consist of noun phrases or named entities (rather than e.g., verbs or other parts of speech). The set of keyphrases was also ranked by impor-

tance, following the same guidelines as used in sentence ranking.

Corporate/private classification. Finally, to validate our own classification of an email as either private or corporate, the annotators were asked to classify each email/thread into one of these two categories.

4.2. Annotation Interface

To facilitate the annotation task, an online interface was developed, intended to be very simple to use. A snapshot of the interface (written in PHP with a MySQL database backend) is shown in Figure 1. After logging in, the annotator is presented with the raw email/thread, and is required to complete all four annotation tasks described above. The email text is available during the entire annotation session and shows each email/thread separated into labeled sentences to assist the user in identifying them. All the annotations are saved into the MySQL database.

4.3. Annotation Files

```
<root>
<annotation email="" annotator="">
<abstractive></abstractive>
<extractive_sentences>
<sentence rank="5"></sentence>
<sentence rank="4"></sentence>
<sentence rank="3"></sentence>
<sentence rank="2"></sentence>
<sentence rank="1"></sentence>
</extractive_sentences>
<keyword_keyphrase>
<keyword rank="5"/>
<keyword rank="4"></keyword>
<keyword rank="3"></keyword>
<keyword rank="2"></keyword>
<keyword rank="1"></keyword>
</keyword_keyphrase>
</annotation>
</root>
```

Figure 3: XML format of the annotations.

The annotations are exported into an XML format, using the structure shown in Figure 3. Note from Figure 3 that we store annotator IDs, abstractive summaries, ranked sentences and keyphrases into several fields. The private/corporate classification was not stored explicitly because its relevance to the summarization and keyword extraction task is yet to be explored.

4.4. Agreement Study

During the annotation process, several disagreements between the annotators were brought into discussion, and several observations were made. First, it was noted that private emails are easier to read but are topically more diverse, which makes abstractive summarization and sentence selection harder. On the other hand, corporate emails, albeit sometimes more difficult to interpret given the technical nature of the conversation, are mainly focused on a single topic, or at most a few topics, thereby making content selection and abstractive summarization somewhat easier.

The summarization style was a discussion point, which resulted in the decision to write abstractive summaries in the third person in order to avoid confusions between the senders of the various emails in a thread. Even though the use of the third person helps untangling pronoun references and ownership of sentences, it is sometimes hard to see what the authors actually contribute to the conversation flow.

For keyphrase selection we mainly considered noun phrases and named entities, although sometimes keyphrases of other types such as verb phrases and adverb phrases were selected. In addition, since the size of the keyphrases can also play an important role, we decided to suggest a limit of at most four words in a keyphrase, which can positively contribute to an increased agreement between annotators.

Considering one annotator as the ground truth, and another as the "system", the agreement on the keyword extraction task was 25.33% precision, 25.33% recall, 25.33% F-score, and 14.50% Jaccard similarity. For sentence extraction task the values were 51.33% precision, 51.33% recall, 51.33% F-score. Precision, recall, and F-score values are the same because both annotators annotated the same number of keyphrases and sentences per document.

The validation of the private/corporate classification during the annotation stage shows a 95% concordance for both annotators, and an inter-annotator agreement of 88%.

5. Conclusions

Summarization and keyword extraction are important problems in natural language processing, where text documents are represented by the most informative sentences and the most informative words or phrases. While general-purpose summarization and keyword extraction has a rich history and many standard datasets available, work on email summarization and keyword extraction has been considerably sparse. The only publicly available dataset annotated for email summarization (Ulrich et al., 2008) is relatively small, and there are no public datasets available for email keyword extraction.

In this work, we present a corpus of emails and threads annotated with abstractive summaries, extractive summaries, keyphrases, and private/corporate classification information. Our exracted sentences and keyphrases are ranked from most important to least important. We have also constructed an annotation website that can be used very easily in further email annotation studies. The interface is simple and modular, thereby yielding a high level of interoperability.

Further, as part of this work, we have designed an annotation scheme and an XML format (adapted from (Ulrich et al., 2008)) appropriate for storing emails/threads and their annotations. We hope that the email corpus described in this paper will spur further research in email summarization and keyword extraction. We also hope that our annotation interface, along with the XML format, will be used by future researchers to not only annotate emails, but also to annotate other forms of conversations, such as online forum threads and tweet streams.

6. Acknowledgements

We are grateful to the annotators who made this work possible. We also acknowledge the anonymous reviewers whose comments significantly improved the content of this paper. This material is based in part upon work supported by Samsung Research America under agreement GN0005468 and by the National Science Foundation under IIS award #1018613. Any opinions, findings, conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of Samsung Research America or the National Science Foundation.

7. References

Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zhou. 2007. Summarizing Email Conversations with Clue Words. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 91–100, New York, NY, USA. ACM.

Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zhou. 2008. Summarizing Emails with Conversational Cohesion and Subjectivity. In *Proceedings of ACL-08: HLT*, pages 353–361, Columbus, Ohio, June. Association for Computational Linguistics.

Simon Corston-Oliver, Eric Ringger, Michael Gamon, and Richard Campbell. 2004. Task-Focused Summarization of Email. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 43–50, Barcelona, Spain, July. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2010. Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 365–373. Association for Computational Linguistics.

Bryan Klimt and Yiming Yang. 2004. Introducing the Enron Corpus. In *First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA.

Smaranda Muresan, Evelyne Tzoukermann, and Judith L. Klavans. 2001. Combining Linguistic and Machine Learning Techniques for Email Summarization. In *Proceedings of the 2001 workshop on Computational Natural Language Learning - Volume 7*, ConLL '01, pages 19:1–19:8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Interpretation and Transformation for Abstracting Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 894–902, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ani Nenkova and Amit Bagga. 2003. Facilitating Email Thread Access by Extractive Summary Generation. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *RANLP*, volume 260 of *Current Issues in Linguistic Theory (CILT)*, pages 287–296. John Benjamins, Amsterdam/Philadelphia.

Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. 2004. Summarizing Email Threads. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 105–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter D. Turney. 2000. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(4):303–336, May.

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *AAAI08 EMAIL Workshop*, Chicago, USA. AAAI.

Stephen Wan and Kathy McKeown. 2004. Generating Overview Summaries of Ongoing Email Thread Discussions. In *Proceedings of Coling 2004*, pages 549–555, Geneva, Switzerland, Aug 23–Aug 27. COLING.

David Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-Candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks. *Information Processing & Management*, 43(6):1549–1570.

David M. Zajic, Bonnie J. Dorr, and Jimmy Lin. 2008. Single-Document and Multi-Document Summarization Techniques for Email Threads Using Sentence Compression. *Inf. Process. Manage.*, 44(4):1600–1610.