The EnronSent Corpus

Will Styler
University of Colorado - Boulder
william.styler@colorado.edu

May 26th, 2011

Contents

1 Introduction

1.1 The Enron Email Corpus

The Enron Email Corpus is a massive dataset, containing ~500,000 messages from senior management executives at the Enron Corporation. Enron was a large American corporation which was investigated by the Federal Energy Regulatory Commission (FERC) in 2001 following its rather spectacular bankruptcy and dissolution.

The Enron Email Dataset website[1] gives the following history of the corpus:

> This dataset was collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation.
>
> The email dataset was later purchased by Leslie Kaelbling at MIT, and turned out to have a number of integrity problems. A number of folks at SRI, notably Melinda Gervasio, worked hard to correct these problems, and it is thanks to them (not me) that the dataset is available. The dataset here does not include attachments, and some messages have been deleted "as part of a redaction effort due to requests from affected employees". Invalid email addresses were converted to something of the form user@enron.com whenever possible (i.e., recipient is specified in some parse-able format like "Doe, John" or "Mary K. Smith") and to no_address@enron.com when no recipient was specified.

1.2 Difficulties with the Enron Corpus for linguistic research

In its raw form, the Enron corpus is a vast set of folders containing 2.2 Gigabytes of messages in MBOX format, all kept individually and numbered sequentially (each folder has 1., 2., 3. 4...), then sorted by folders and users. These messages had not been reformatted at all, and were "straight from the server", spam, computer generated messages and all. Although originally the corpus contained all of the company email as seized by the FERC, the maintainers have removed the home folders of those who explicitly requested it, as well as redacting some messages which were similarly flagged by their authors or recipients.

The Enron Email Corpus is an excellent resource, and is a rare glimpse into the language usage patterns present in email (among other purposes), but for corpus-based linguistic research, it suffers from a few downfalls.

[1] http://www-2.cs.cmu.edu/~enron/, accessed May 26, 2011

First and foremost, a large portion of the text in the corpus is not human-generated at all. In addition to machine-generated headers attached to each message, Enron's servers were as clogged with spam, company newsletters, automatically appended signatures and machine-generated notification emails as any others. So, search for any given term was likely to yield as many machine-generated lines as human generated lines, especially given the rather complex methods that spammers use to approximate human-created text.

Second, even among human-generated content, the amount of duplications of message text was staggering. Because of our habit of quoting original messages in replies, a single message may include text from several prior messages. As such, word and expression counts are often inaccurate, as a single email may be quoted tens of times throughout a long email thread, and messages from important individuals may be quoted or discussed more heavily still. So, if an otherwise vanishingly rare word occurs in an important email from an important person, the number of messages which technically contain that word (by virtue of quoting) will make that word seem much more common, even if it was seldom used intentionally in original messages.

Finally, the formatting of the corpus itself makes work difficult for linguists. Many linguists are not familiar with email formats, and are even less familiar with email headers and with examining plaintext, MBOX formatted messages. This formatting (and the associated headers, quoting, etc) makes it extremely difficult to search the corpus with tools designed for searching plaintext, and as such, the corpus is further inaccessible to linguists. In addition, the corpus itself in its raw form is nearly 2.2 gigabytes, containing thousands of directories, which itself is a barrier to efficient or casual examination.

1.3 Goals of the EnronSent corpus

Given those troubles with the full Enron Email Corpus, I set out to create a sub-corpus which would adhere to four fundamental principles:

1. Create a corpus which can be used freely and easily with conventional corpus linguistics tools (like Regular Expressions, grep, and script-parsing)
2. Clean out as much of the machine-generated emails and text as possible
3. Keep the corpus as large enough to provide a good sample, but not so large as to inhibit ease of use
4. Bring the final result into an easily read and searched plaintext format

After several months of work, strict adherence to these principles resulted in the creation of the EnronSent corpus. The process of generating this corpus is described below.

2 Creating the EnronSent Corpus

2.1 Retrieving and simplifying the raw emails

The EnronSent corpus is based on the Enron Email Corpus, as retrieved from http://www.cs.cmu.edu/~enron/ in April of 2006. As previously mentioned, this is a 2.2 GB corpus spread across thousands of individual folders for 150+ employees.

The first step required was to narrow the scope of the processing, both to improve the quality of messages within, and to reduce the daunting scale of a corpus search on the dataset. Because of the high frequency of spam, bulk marketing messages, company newsletters, and automatically generated emails ("This is an automated reminder that your progress report for X is due..."), the highest quality source of messages turned out to be the "sent" and "sent_mail" folders for each user.

In most workflows, the user will never copy any spam or company newsletters into their sent folder, and most programs which send automated messages will do so using a separate email client, usually without sent message archiving. So, by using only mail from the "sent" and "sent_mail" folders in the corpus, I was able to automatically extract ~96,000 files which were far more likely to be from a human, to a human, and using actual human language patterns.

Unfortunately, the corpus still consisted of ~96,000 files spread across 150+ user folders. Given that the identity of the sender is completely irrelevant to most linguistic research, and that performing searches across that many files will often lead you into argument overflow errors, it seemed that the best choice in terms of ease of use was to concatenate the messages across all "sent" and "sent_mail" folders in the corpus into an extraordinarily large plaintext, containing one raw message after another (separated with two blank lines).

2.2 Cleaning the corpus

In order to remove non-human-generated and quoted text, a state machine was written in Python to remove undesired texts and sections. In this process, hours were spent moving through the corpus to identify common machine-added, redundant or otherwise undesirable message types, and the state machine was modified to recognize and remove sections matching these patterns. The sections and patterns identified and removed were:

- All Email Headers. Each message had 10 or more lines of nearly identical headers that (excepting the subject line) contained no human language. Although by removing headers, the ability to tie one message/sender to the next message/recipient was lost, those wanting to study interactional patterns will be better off using the full corpus, and to

preserve them would add more bulk than good. In the process, a small subcorpus, EnronSubjects, was created.

- Quoted and Forwarded Messages. Because the same message was duplicated many times when replies and forwards were left in, all quoted sections and forwarded messages were removed.
- HTML Messages. Frequently, they are tagged so heavily that they become unreadable and cause undue clutter in the corpus, and they're easily removed. Also, many of them were spam messages or retail newsletters, which, although written by humans, do not provide an accurate view of natural email communication.
- "Subscribe to our Mail Service today" messages from Yahoo, MSN, Gmail and other providers. Although they are generally short and formulaic, they still add bulk to the corpus.
- Enron Specific formulaic signatures and letterheads. The corpus was full of "Enron Legal Team. This message is intended only for..." paragraphs, and many of them, especially when combined with other signatures and notices, were easily long enough to affect word counts.

This script was designed to be more aggressive and to err on the side of removing data, because the sheer amount of data in the corpus allowed it. Of course, no automated method will completely remove all of these elements, and doubtless many examples of each remain (along with other missed sorts of machine generated text), but this process alone removed nearly 2.5 million lines of text, leaving behind the corpus' current 220,500 lines of text (roughly 14 million words).

It is worth noting that no attempt has been made to strip e-mail and home addresses, websites, phone numbers, and other personal information. Given that the headers have been stripped and there's no easy way to associate an email with a given author (short of non-formulaic signatures), and also given there is no content included from those employees who requested their messages withheld from the original corpus, the author felt no need to read through all the remaining messages to de-identify data. That said, any future requests for redaction of the corpus or removal of messages will be considered and likely implemented.

2.3 Formalizing the corpus

Finally, to make the corpus more accessible, the nearly 85 MB plaintext file was split up into 44 plaintext files of ~50,000 lines each, and a README file was created.

2.4 EnronSubjects

During the cleaning process, all subject lines from human-created messages were removed with other message headers, but rather than discarding them, they were instead put aside into a subcorpus, the EnronSubjects corpus. Although some mild cleaning was done to the subject lines (removing “Subject:" markup and extra tabs and spaces), this subcorpus contains many near- and total repetitions and was not subjected to the same scrutiny was the primary corpus.

The EnronSubjects subcorpus is a single, plaintext file containing 370,537 lines, each a different subject from each email considered. Note that the number of messages whose subjects are displayed and the number included in the corpus are different. This is mostly due to the presence of HTML mail which was stripped out of the dataset and forwards/replies which contained no new human-readable text. It is available alongside the main corpus with hope that it may prove useful to some researcher, but is not formally supported.

3 The EnronSent Corpus

3.1 Conclusion and Disclaimers

Even following all the steps described above to "clean" the corpus, the corpus is far from pristine. There is no shortage of forwarded articles, automated messages, and probably errant headers that escaped the scrubber, but compared to the original Enron Email Corpus, the EnronSent corpus is a far more accessible and usable corpus for basic linguistic research, and the four basic goals (usability, removal of non-human-generated text, optimal size and ease of access and search) have been met.

The corpus has since been used in several informal studies, and the author is in communication with several researchers in various stages of projects which incorporate this dataset, and the author sincerely hopes that future researchers in Linguistics and beyond will find this particular preparation of a unique and valuable dataset useful.

3.2 Final Corpus Statistics

In its final, posted form, the EnronSent corpus contains 96106 messages, 220500 lines, 13810266 words and 88171505 characters, all spread across 44 plaintext files, each containing ~50,000 lines.

3.3 Availability

The EnronSent corpus has been released into the public domain, and is available for free download from http://www.verbs.colorado.edu/enronsent. You may use and redistribute the data as you wish, but we ask that a citation to this paper be included along with the corpus README file.

4 Acknowledgements

The author wishes to thank the following people and entities for their assistance in preparing the corpus.

- Martha Palmer, University of Colorado; for her help and guidance through this project and the eventual deployment of the corpus into the world.
- Mark Dredze, University of Pennsylvania; for his help, insights and inspiration.
- Travis Millburn, University of Colorado; for his help and suggestions with Python coding and his constant willingness to help.
- Dmitriy Dligach, University of Colorado; for his help with Python coding.
- The good folks at the United States Federal Energy Regulatory Commission, for releasing this data into the public domain.