Annotating Subsets of the Enron Email Corpus

Jade Goldstein[1], Andres Kwasinski[2], Paul Kingsbury[3], Roberta Evans Sabin[4], Albert McDowell[1]

[1]US Department of
Defense

jadeg@acm.org

[2]Institute for Systems
Research
University of Maryland
College Park, MD 20742

ak@umd.edu

[3]Northrup Grumman TASC
2701 Technology Drive.
Suite 100
Linthicum, MD 20701

paul.kingsbury@ngc.com

[4]Computer Science
Department
Loyola College in MD
Baltimore, MD 21210

res@loyola.edu

ABSTRACT

We present an annotation project for two subsets of the Enron email corpus. The first is a subset of the UC Berkeley Enron Email Analysis Project and the second consists of a portion of emails from the Voice Transcripts Email Correlated Corpora. Parts of the automatic content extraction (ACE) annotation guidelines, extended for the email domain are used for annotation. We also categorize the emails with email speech acts, mark whether the text contains discussions of meetings/conversations, and determine the degree of correlation of the subject line with the text body.

1. CORPUS CREATION

The purpose of this project was to create an annotated corpus that could be used for further email research. In 2003, the Federal Energy Regulatory Commission (FERC) as a result of its investigation of Enron's energy trading practices [3] made available to the public the Enron email corpus. We chose to use two subsets of this corpus. The first is a subset of the UC Berkeley Enron Email Analysis Project (BEEAP) (http://bailando.sims.berkeley.edu/enron_email.html), which consists of approximately 1900 genre-labeled emails from the Enron corpus. Selected emails are primarily related to business and specifically to the California Energy Crisis; each is labeled with a category. The second was a subset of emails from the Voice Transcripts Email Correlated Corpora. We created this data set by selecting emails from authors for whom there are also available audio files and corresponding voice transcripts.

1.1 BEEAP Subset

We filtered the BEEAP Enron subset, removing emails in which the forwarded information would not be interesting to annotate. This resulted in the removal of emails containing forwarded news articles, government and academic reports, press releases, pointers to urls, newsletters, and jokes (see Table 1).

1.2 VTECC Subset

The VTECC Enron subset represents a collection of different forms of communication for an individual: written content in emails, spoken content from recorded phone calls and audio transcripts for the calls. The phone calls dataset (http://www.enrontapes.com/files.html) is formed by 93 audio files in wave format, which were submitted as evidence in June, 2004 and January, 2005 There are 93 transcribed audio files, (a subset of a collection of 52 DVDs) spanning 88 days between August, 2000 and January, 2001.

Table 1. BEEAP: Enron Coarse Genre Email Statistics

Coarse Genre	BEAAP	Our Corpus
Company Business	855	304
Purely Personal	49	33
Personal Professional	165	104
Logistics	533	355
Employment	96	68
Document	176	111
Missing Attachment	25	21
No Sender Text Body	26	17
Total Emails	1925	1013

To create the joint corpus we identified Enron employees in the phone calls, and then selected all available emails for that employee. All the phone calls were saved by a system that recorded Enron's Western electric traders' operations. To identify the Enron employees in the calls we used a combination of heuristics and inference from objective evidence. In some cases it was possible to identify a party from the call's conversation flow. In other cases, we took advantage of the fact that calls from the same trader tend to appear in the same channel of the recording system (FERC Exhibit SNO-161) [3]. Using all this information we classified participants as “certain”, “probable”, or “unknown”. The “certain” category was used for those Enron employees (15) who are identified without doubt as participating in at least one call. The “probable” category was used for those Enron employees (14) who may be participating in at least one call. This category occurred when we identified the first name of one of the call parties and there was a match between this name and the name of an employee who used the recording channel corresponding to the call. A person identified as both “certain” and “probable” was assigned the overall category “certain”.

The subset of emails for this corpus consists of all emails sent or received by the persons in our list from the Enron corpus (http://www-2.cs.cmu.edu/~enron) [5]. Duplicates were eliminated. We organized the selected emails into 4 directories: “from certain”, “to certain”, to store the emails received and sent, respectively, to those persons who belonged to the “certain” class and “from probable”, “to probable”, to store the emails received and sent, respectively, to those persons who belonged to the “probable” class. We considered that an email was sent to a person if the email address of that person is the “To”, “cc”, or “Bcc" fields in the email header. Statistics for the corpus are shown in Table 2. Identifier codes are M = male, F = female and C = certain and P = probable. Overlap is marked “Y” if the voice transcripts overlapped with the email date range for the emails

CEAS 2006 – Third Conference on Email and Anti-Spam, July 27-28,
2006, Mountain View, California USA

from that individual and "+/- one day" was marked "Y" if there was an email from the individual within one day of a voice transcript.

Table 2. VTECC: Enron Email + Voice Transcripts Statistics

ID	#From	#To	#V. Trans.	Overlap	+/-1Day
MC1	2	461	4	N	
MC2	429	420	5	N	
MC3	36	394	5	Y	Y
MC4	21	1091	5	Y	
MC5	9	879	3	N	
MC6	0	1	3	N	
MC7	13	1280	5	N	
MC8	37	453	3	Y	
MC9	514	62823	1	Y	
MC10	89	1191	4	Y	Y
MC11	20	887	4	N	
MC12	4	98	3	N	
FC1	137	763	3	Y	Y
FC2	6	221	3	Y	
FC3	32	317	1	Y	
MP1	581	4275	9	N	
MP2	3	283	2	N	
MP3	44	303	5	Y	Y
MP4	83	4207	6	Y	
MP5	6	1868	3	N	
MP6	10	184	2	N	
MP7	154	1149	2	Y	
MP8	60	4187	3	N	
MP9	53	376	2	N	
MP10	222	3030	5	Y	Y
MP11	74	934	1	N	
MP12	8	884	2	N	
MP13	62	237	3	N	
FP1	283	1032	1	N	

2. ANNOTATIONS

We developed annotation guidelines for the Enron corpus based on two frameworks. The first was a subset and extension [2] of the Automatic Content Extraction (ACE) guidelines [1] for email. The second was a top level (overall) annotation about the email. After an annotator had reached a certain level of accuracy (80%) on training data, each email was singly annotated. Annotators had continued internal consistency checks and inter-annotator consistency checks to ensure high quality of the annotated corpus. Inter-annotator agreement for three annotators on a set of 57 documents was 91.5%. All emails in Table 1 and a portion of the emails in Table 2 were annotated to create a total corpus of 2000 annotated emails.

2.1 ACE Guidelines Subset

The corpus was annotated according to ACE guidelines [1]; thus entities were co-referenced with each other. Accordingly, a name of a person in an email header was co-referenced with any names in the text, such as in the phrase "Hi Bob", or a name in a signature block. Only ACE relations that had to do with a person were marked, such as social relations, location information and organization affiliation: PER-PER relations (business, familial), PER-GRP (is-a-member-of) PER-SOC, PER-ORG (belongs-to) and PER-LOC/GPE-LOC (is-located).

2.2 ACE Guidelines Extensions

To address the peculiar characteristics of email, the ACE Guidelines were extended [2]. This included the introduction of new entity types (SIGNATURE BLOCK, BOILERPLATE and ZIPCODE) and the upgrading of some attribute information, previously marked as values, to entities. The latter category includes email address, phone number, phone extension, fax number, and url. These new entity types often have subtypes: phone number has subtypes: work, home, and unspecified. In signature blocks, unspecified phone numbers default to work and addresses are broken down to components, such as FAC for building locations, GPE for cities and states, and ZIPCODE. Addresses mentioned in the text body were annotated using the ACE protocol, therefore they were not broken down into their subcomponents. Boilerplates include information such as company disclaimers or quotes by the author. An attachment event (using ACE event criteria) was introduced to mark cases where there was a description of an email attachment, e.g., "Attached please find the new annotation guidelines."

We included four top level annotations. (1) Subject line alignment, i.e., whether the subject line accurately reflects the content of the email. There were three possible values: (a) *content summary* – the subject line accurately describes the main purpose of the email: the reader can correctly surmise the intent of the sender without reading the email, (b) *connected* – may describe the purpose of the email, but provides little content, for example, "a question", "status", or "tablet pc again," or (c) *unconnected* – the subject is not relevant to the email topic, such as may result from topic drift. (2) Email speech act of the sender (annotated using 30 subcategories [4]). (3) Mention of a face-to-face meeting in the past, present or future, marked yes, no, or unclear. The unclear case covers both possible future meetings and the case where there is ambiguity as to whether the meeting was face-to-face or via telephone. (4) Mention of a telephone conversation in the past, present or future. These were marked in the same manner as face-to-face meetings.

We hope that the availability of these corpora (http://jikd-email.umiacs.umd.edu/corpus) will prove useful in future email research efforts. In the future, we hope to annotate additional email data as well as the voice transcripts.

3. ACKNOWLEDGMENTS

Thanks to JK Davis, Jason Duncan, Alan Goldschen, and Kyle Haycock for helpful discussions pertaining to the guidelines. Thanks also to Colin Ferguson and Matt Seymour for programming support.

4. REFERENCES

[1] ACE 2005 Annotation Guidelines. *http://projects.ldc.upenn.edu/ace/annotation/2005Tasks.html*

[2] Sabin R., Goldstein, J., Haycock K., and McConkie V. *Email Annotation Guidelines*. Loyola CS Tech Report 003, 2006.

[3] FERC online elibrary, docket e103-180. *http://elibrary.ferc.gov/idmws/search/fercgensearch.asp*

[4] Goldstein, J and Sabin, R. Using Speech Acts to Categorize Email and Identify Email Genres. In *Hawaii International Conference on Systems Sciences, HICSS*, January, 2006.

[5] Klimt B. and Yang, Y. Introducing the Enron corpus. In *Conference on Email and Anti-Spam CEAS*, July, 2004.