Towards the Orwellian Nightmare

Separation of Business and Personal Emails

Sanaz Jabbari, Ben Allison, David Guthrie, Louise Guthrie
Department of Computer Science
University of Sheffield
211 Portobello St.
Sheffield
S1 4DP
{s.jabbari, b.allison, d.guthrie, l.guthrie}@dcs.shef.ac.uk

Abstract

This paper describes the largest scale annotation project involving the Enron email corpus to date. Over 12,500 emails were classified, by humans, into the categories "Business" and "Personal", and then subcategorised by type within these categories. The paper quantifies how well humans perform on this task (evaluated by inter-annotator agreement). It presents the problems experienced with the separation of these language types. As a final section, the paper presents preliminary results using a machine to perform this classification task.

1 Introduction

Almost since it became a global phenomenon, computers have been examining and reasoning about our email. For the most part, this intervention has been well natured and helpful – computers have been trying to protect us from attacks of unscrupulous blanket advertising mail shots. However, the use of computers for more nefarious surveillance of email has so far been limited. The sheer volume of email sent means even government agencies (who can legally intercept all mail) must either filter email by some preconceived notion of what is interesting, or they must employ teams of people to manually sift through the volumes of data. For example, the NSA has had massive parallel machines filtering e-mail traffic for at least ten years.

The task of developing such automatic filters at research institutions has been almost impossible, but for the opposite reason. There is no shortage of willing researchers, but progress has been hampered by the lack of any data – one's email is often hugely private, and the prospect of surrendering it, in its entirety, for research purposes is somewhat unsavoury.

Recently, a data resource has become available where exactly this condition (several hundred people's entire email archive) has been satisfied – the Enron dataset. During the legal investigation of the collapse of Enron, the FERC (Federal Energy Regulatory Commission) seized the emails of every employee in that company. As part of the process, the collection of emails was made public and subsequently prepared for research use by researchers at Carnegie Melon University (Klimt and Yang, 2004).Such a corpus of authentic data, on such a large scale, is unique, and an invaluable research tool. It then falls to the prospective researcher to decide which divisions in the language of email are interesting, which are possible, and how the new resource might best be used.

Businesses which offer employees an email system at work (and there are few who do not) have always known that they possess an invaluable resource for monitoring their employees' work habits. During the 1990s, UK courts decided that that an employee's email is not private – in fact, companies can read them at will. However, for exactly the reasons described above, automatic monitoring has been impossible, and few businesses have ever considered it sufficiently important to employ staff to monitor the email use of other staff. However, in monitoring staff productivity, few companies would decline the use of a system which could analyse the email habits of its employees, and report the percentage of time which each employee was spending engaged in non-work related email activities.

The first step in understanding how this problem might be tackled by a computer, and if it is even feasible for this to happen, is to have humans perform the task. This paper describes the process of having humans annotate a corpus of emails, classifying each as to whether they are business or personal, and then attempting to classify the type of business or personal mail being considered.

A resource has been created to develop a system able to make these distinctions automatically. Furthermore, the process of subcategorising types of business and personal has allowed invaluable insights into the areas

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 407–411,
Sydney, July 2006. ©2006 Association for Computational Linguistics

where confusion can occur, and how these confusions might be overcome.

The paper presents an evolution of appropriate subcategories, combined with analysis of performance (measured by inter-annotator agreement) and reasons for any alterations. It addresses previous work done with the Enron dataset, focusing particularly on the work of Marti Hearst at Berkeley who attempted a smaller-scale annotation project of the Enron corpus, albeit with a different focus. It concludes by suggesting that in the main part (with a few exceptions) the task is possible for human annotators. The project has produced a set of labeled messages (around 14,000, plus double annotations for approximately 2,500) with arguably sufficiently high business-personal agreement that machine learning algorithms will have sufficient material to attempt the task automatically.

2 Introduction to the Corpus

Enron's email was made public on the Web by FERC (Federal Energy Regulatory Commission), during a legal investigation on Enron Corporation. The emails cover 92 percent of the staff's emails, because some messages have been deleted "as part of a redaction effort due to requests from affected employees". The dataset was comprised of 619,446 messages from 158 users in 3,500 folders. However, it turned out that the raw data set was suffering from various data integrity problems. Various attempts were made to clean and prepare the dataset for research purposes. The dataset used in this project was the March 2, 2004 version prepared at Carnegie Mellon University, acquired from http://www.cs.cmu.edu/~enron/. This version of the dataset was reduced to 200,399 emails by removing some folders from each user. Folders like "discussion threads" and "all documents", which were machine generated and contained duplicate emails, were removed in this version.

There were on average 757 emails per each of the 158 users. However, there are between one and 100,000 emails per user. There are 30,091 threads present in 123,091 emails. The dataset does not include attachments. Invalid email addresses were replaced with "user@enron.com". When no recipient was specified the address was replaced with "no_address@enron.com" (Klimt and Yang, 2005).

3 Previous Work with the Dataset

The most relevant piece of work to this paper was performed at Berkeley. Marti Hearst ran a small-scale annotation project to classify emails in the corpus by their type and purpose (Email annotation at Berkely). In total, approximately 1,700 messages were annotated by two distinct annotators. Annotation categories captured four dimensions, but broadly speaking they reflected the following qualities of the email: coarse genre, the topic of the email if business was selected, information about any forwarded or included text and the emotional tone of the email. However, the categories used at the Berkeley project were incompatible with our requirements for several reasons: that project allowed multiple labels to be assigned to each email; the categories were not designed to facilitate discrimination between business and personal emails; distinctions between topic, genre, source and purpose were present in each of the dimensions; and no effort was made to analyse the inter-annotator agreement (Email annotation at Berkely).

User-defined folders are preserved in the Enron data, and some research efforts have used these folders to develop and evaluate machine-learning algorithms for automatically sorting emails (Klimt and Yang, 2004). However, as users are often inconsistent in organising their emails, so the training and testing data in these cases are questionable. For example, many users have folders marked "Personal", and one might think these could be used as a basis for the characterisation of personal emails. However, upon closer inspection it becomes clear that only a tiny percentage of an individual's personal emails are in these folders. Similarly, many users have folders containing exclusively personal content, but without any obvious folder name to reveal this. All of these problems dictate that for an effective system to be produced, large-scale manual annotation will be necessary.

Researchers at Queen's University, Canada (Keila, 2005) recently attempted to categorise and identify deceptive messages in the Enron corpus. Their method used a hypothesis from deception theory (e.g., deceptive writing contains cues such as reduced frequency of first-person pronouns and increased frequency of "negative emotion" words) and as to what constitutes deceptive language. Single value decomposition (SVD) was applied to separate the emails, and a manual survey of the results allowed them to conclude that this classification method for detecting deception in email was promising.

Other researchers have attempted to analyse the Enron emails from a network analytic perspective (Deisner, 2005). Their goal was to analyse the flow of communication between employees at times of crisis, and develop a characterisation for the state of a communication network in such difficult times, in order to identify looming crises in other companies from the state of their communication networks. They compared the network flow of email in October 2000 and October 2001.

4 Annotation Categories for this Project

Because in many cases there is no definite line between business emails and personal emails, it was decided to mark emails with finer categories than

Business and Personal. This subcategorising not only helped us to analyse the different types of email within business and personal emails, but it helped us to find the nature of the disagreements that occurred later on, in inter-annotation. In other words, this process allowed us to observe patterns in disagreement.

Obviously, the process of deciding categories in any annotation project is a fraught and contentious one. The process necessarily involves repeated cycles of category design, annotation, inter-annotation, analysis of disagreement, category refinement. While the process described above could continue ad infinitum, the sensible project manager must identify were this process is beginning to converge on a set of well-defined but nonetheless intuitive categories, and finalise them.

Likewise, the annotation project described here went through several evolutions of categories, mediated by input from annotators and other researchers. The final categories chosen were:

Business: Core Business, Routine Admin, Inter-Employee Relations, Solicited/soliciting mailing, Image.

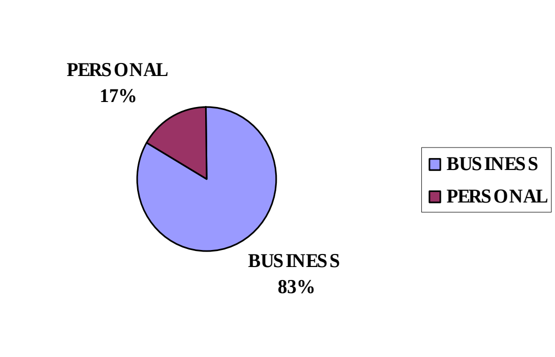
Personal: Close Personal, Forwarded, Auto generated emails.

5 Annotation and Inter-Annotation

Based on the categories above, approximately 12,500 emails were single-annotated by a total of four annotators.

The results showed that around 83% of the emails were business related, while 17% were personal. The company received one personal email for every five business emails.

Fig 1: Distribution of Emails in the Corpus

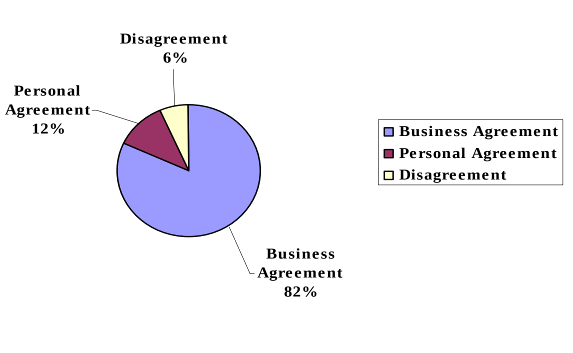


A third of the received emails were "Core Business" and a third were "Routine Admin". All other categories comprised the remaining third of the emails. One could conclude that approximately one third of emails received at Enron were discussions of policy, strategy, legislation, regulations, trading, and other high-level business matters. The next third of received emails were about the peripheral, routine matters of the company. These are emails related to HR, IT administration, meeting scheduling, etc. which can be regarded as part of the common infrastructure of any large scale corporation.

The rest of the emails were distributed among personal emails, emails to colleagues, company news letters, and emails received due to subscription. The biggest portion of the last third, are emails received due to subscription, whether the subscription be business or personal in nature.

In any annotation project consistency should be measured. To this end 2,200 emails were double annotated between four annotators. As Figure 2 below shows, for 82% of the emails both annotators agreed that the email was business email and in 12% of the emails, both agreed on them being personal. Six percent of the emails were disagreed upon.

Fig 2: Agreements and Disagreements in Inter-Annotation



By analysing the disagreed categories, some patterns of confusion were found.

Around one fourth of the confusions were solicited emails where it was not clear whether the employee was subscribed to a particular newsletter group for his personal interest, private business, or Enron's business. While some subscriptions were clearly personal (e.g. subscription to latest celebrity news) and some were clearly business related (e.g. Daily Energy reports), for some it was hard to identify the intention of the subscription (e.g. New York Times).

Eighteen percent of the confusions were due to emails about travel arrangements, flight and hotel booking confirmations, where it was not clear whether the personal was acting in a business or personal role.

Thirteen percent of the disagreements were upon whether an email is written between two Enron employees as business colleagues or friends. The emails such as “shall we meet for a coffee at 2:00?” If insufficient information exists in the email, it can be hard to draw the line between a personal relationship and a relationship between colleagues. The annotators were advised to pick the category based on the formality of the language used in such emails, and reading between the lines wherever possible.

About eight percent of the disagreements were on emails which were about services that Enron provides for its employees. For example, the Enron’s running club is seeking for runners, and sending an ad to Enron’s employers. Or Enron’s employee’s assistance Program (EAP), sending an email to all employees, letting them know that in case of finding themselves in stressful situations they can use some of the services that Enron provides for them or their families.

One theme was encountered in many types of confusions: namely, whether to decide an e-mail’s category based upon its topic or its form. For example, should an email be categorised because it is scheduling a meeting or because of the subject of the meeting being scheduled? One might consider this a distinction by topic or by genre.

As the result, final categories were created to reflect topic as the only dimension to be considered in the annotation. “Solicited/Soliciting mailing”, “Solicited/Auto generated mailing” and “Forwarded” were removed and “Keeping Current”, “Soliciting” were added as business categories and “Personal Maintenance” and “Personal Circulation” were added as personal categories. The inter-annotation agreement was measured for one hundred and fifty emails, annotated by five annotators. The results confirmed that these changes had a positive effect on the accuracy of annotation.

6 Preliminary Results of Automatic Classification

Some preliminary experiments were performed with an automatic classifier to determine the feasibility of separating business and personal emails by machine. The classifier used was a probabilistic classifier based upon the distribution of distinguishing words. More information can be found in (Guthrie and Walker, 1994).

Two categories from the annotation were chosen which were considered to typify the broad categories – these were Core Business (representing business) and Close Personal (representing personal). The Core Business class contains 4,000 messages (approx 900,000 words), while Close Personal contains approximately 1,000 messages (220,000 words).

The following table summarises the performance of this classifier in terms of Recall, Precision and F-Measure and accuracy:

Class	Recall	Precision	F-Measure	Accuracy
Business	0.99	0.92	0.95	0.99
Personal	0.69	0.95	0.80	0.69
AVERAGE	**0.84**	**0.94**	**0.88**	**0.93**

Based upon the results of this experiment, one can conclude that automatic methods are also suitable for classifying emails as to whether they are business or personal. The results indicate that the business category is well represented by the classifier, and given the disproportionate distribution of emails, the classifier’s tendency towards the business category is understandable.

Given that our inter-annotator agreement statistic tells us that humans only agree on this task 94% of the time, preliminary results with 93% accuracy (the statistic which correlates exactly to agreement) of the automatic method are encouraging. While more work is necessary to fully evaluate the suitability of this task for application to a machine, the seeds of a fully automated system are sown.

7 Conclusion

This paper describes the process of creating an email corpus annotated with business or personal labels. By measuring inter-annotator agreement it shows that this process was successful. Furthermore, by analysing the disagreements in the fine categories, it has allowed us to characterise the areas where the business/personal decisions are difficult.

In general, the separation of business and personal mails is a task that humans can perform. Part of the project has allowed the identification of the areas where humans cannot make this distinction (as demonstrated by inter-annotator agreement scores) and one would not expect machines to perform the task under these conditions either. In all other cases, where the language is not ambiguous as judged by human annotators, the challenge has been made to automatic classifiers to match this performance.

Some initial results were reported where machines attempted exactly this task. They showed that accuracy almost as high as human agreement was achieved by the system. Further work, using much larger sets and incorporating all types of business and personal emails, is the next logical step.

Any annotation project will encounter its problems in deciding appropriate categories. This paper described the various stages of evolving these categories to a stage where they are both intuitive and logical and also, produce respectable inter-annotator agreement scores. The work is still in progress in ensuring maximal consistency within the data set and refining the precise definitions of the categories to avoid possible overlaps.

References

Brian Klimt and Yiming Yang. 2004. *Introducing the Enron Email Corpus,* Carnegie Mellon University.

Brian Klimt and Yiming Yang. 2004. *The Enron Corpus: A New Data Set for Email Classification Research*. Carnegie Mellon University.

Email Annotation at Berkely
http://bailando.sims.berkeley.edu/enron_email.html

Jana Diesner and Kathleen M. Karley. 2005. *Exploration of Communication Networks from the Enron Email Corpus*, Carnegie Mellon University

Louise Guthrie, Elbert Walker and Joe Guthrie. 1994 *Document classification by machine: Theory and practice*. Proc. of COLING'94

Parambir S. Keila and David B. Skillcorn. 2005. *Detecting Unusual and Deceptive Communication in Email.* Queen’s University, CA