The Best Kept Secrets with Corpus Linguistics

Neil Cooke,[1] Lee Gillam[2] and
Ahmet Kondoz[1]

Abstract

This paper presents the use of corpus linguistics techniques on supposedly "clean" corpora and identifies potential pitfalls. Our work relates to the task of filtering sensitive content, in which data security is strategically important for the protection of government and military information, and of growing importance in combating identity fraud. A naïve keyword filtering approach produces a large proportion of false positives, and the need for more fine-grained approaches, suggests the consideration of using corpus linguistics for such content filtering.

We present work undertaken on the Enron corpus, a collection of emails that has had various tasks undertaken on various portions and versions of it by other researchers. We have made some efforts to reconcile differences between the different versions by considering what impact some of these versions have on our results. Our anticipated efforts in using automatic ontology learning [Gillam, Tariq and Ahmad 2005; Gillam and Ahmad 2005], and local grammars [Ahmad, Gillam and Cheng 2006] to discover points of interest within the Enron corpus have been re-oriented to the problem of discovering "confidentiality banners" with a view to removing them from the collection. Our work to date makes strong use of collocation patterns [Smada 1993] to identify the signatures of these banners, and to use the banners themselves to better understand the different versions of the Enron corpus. We further consider the use of extended collocation patterns to identify text "zones", following [Teufel and Moens 2000], and the subsequent potential for sentiment analysis [Klimt and Yang 2004]; [Pang, Lee and Vaithyanathan 2002]

1 Introduction

To obtain ever-more robust statistical results, scientists look towards ever-larger datasets. Exemplar initiatives using increasingly large datasets are now to be found in a variety of so-called e-Science initiatives hoping to discover, for example, origins of diseases, species or the universe itself. In Corpus Linguistics, the trend towards larger datasets has been helped by the emergence of a number of technologies, not least of which is the Web. The one hundred million tokens of the British National Corpus have proved to be a beneficial testing ground, and have led even to the development of an American counterpart. Large corpora, typically news-wire, have been used in a variety of US-based competitions sponsored by the Defence Advanced Research Projects Agency (DARPA). These large corpora, such as the English Gigaword

[1] CCSR, University of Surrey
e-mail: n.cooke@surrey.ac.uk, a.kondoz@surrey.ac.uk
[2] Computer Science Department, University of Surrey
e-mail: l.gillam@surrey.ac.uk

corpus, have been used as the basis for investigating and benchmarking of summarization, information retrieval and information extraction techniques. Some corpus linguists can currently be found investigating the properties of language as it exists in the wild, at varying degrees of "quality", on blogs, in wikis, in syndicated news feeds, in the restricted form provided for by SMS, and on the web in general. Topics such as language coverage and representativeness are oft-discussed. Part of the challenge, indeed the first necessary step, for "web-based" corpus linguists is to ensure that analysis of these larger datasets will produce reliable results. Because of this, a recent challenge in the field orients researchers towards the cleaning-up of untamed, typically web, datasets. Data cleansing is considered a common activity within data mining tasks, wherein the removal of outliers and consideration of key variables and their dependencies and correlates are made. For web-based text collections, such data cleansing can involve the removal of advertisements and decisions over the appropriate handling of text embedded within client-side programming statements. The handling of information presented for visual interpretation, within images and tables, presents further challenges and areas of investigation for a variety of researchers of different hues.

One reasonably large dataset for the corpus linguist is the Enron email collection. The Enron corpus was released into the public domain by the Federal Energy Regulatory Commission (FERC) during the investigation of the collapse of the company. FERC released this dataset with the intention that those not involved in the case could make their own minds up about what happened. The original corpus was around 1.5 million emails: experts and non-experts could take some time to process this. The version of the Enron corpus most readily available[3] comprises 517,431 files having undergone some data cleansing, and with the removal of some emails to protect the privacy of the individuals. These files reportedly do not contain attachments, but this claim is demonstrably false. There are other features of these emails that may or may not be needed, depending on the nature of the analysis: for the sake of accuracy, greater efforts at cleansing are required. Yet other researchers have presented subsets of this dataset, claiming to have undertaken removal of duplicates and other features of the data, but without access to the requisite tools and rules it is difficult to relate these new datasets to the originals and to validate the veracity of claims of cleanliness.

In this paper, we present analysis of current versions of the Enron corpus, and demonstrate that assumptions of having a clean collection can lead to false conclusions. During our investigations we found the readily-available Enron corpus [Enron-Raw] to be polluted by the presence of email headers and corporate disclaimers. We encountered further reduced, apparently more clean, collections, and began efforts to assess these in relation to the original data. Our investigations identified yet other issues with the processes undertaken in cleaning. Unless careful and robust efforts have been made to remove the pollution from the data, results obtained could be leading towards false conclusions. Our analysis, in particular, considers a *noisy* feature of modern email – the confidentiality banner – and how the use of this banner, in various forms, throughout the corpus will skew results even following removal of other forms of pollution. We consider some supposedly cleaner versions of the Enron corpus, patterns that emerge, what these patterns may predict, and differences in patterns within different samples of the collection. We then

[3] Available at: http://www.cs.cmu.edu/~enron/

propose a technique to further clean the corpus by zoning out the banners. We note note that corpus linguists and corpus linguistics may be able to benefit from further exploration of analytical techniques typically used in data mining and information retrieval to result in a good quality corpus.

Our domain of application is information assurance, particularly in terms of assurance at an organizational level, and our intention with addressing confidentiality banners is to prototype a system that includes automatic email corpus cleansing, making use of certain corpus linguistics techniques. The final system may provide useful insights into language use, but is initially focussed on the construction of a particular algorithm and application: filtering potentially damaging propagation in running text of sensitive or confidential content. High profile breaches of data security have included the sale of memory sticks containing US military secrets in a bazaar in Afghanistan, high street banks encouraging customers to shred bank statements while leaving un-shredded account details in rubbish bags, and Nigerian fraudsters recovering bank details from PCs sent to Africa for recycling. Contemporary technologies such as email enable secrets of governments and industrials to be rapidly propagated in unencrypted form to a worldwide audience, and governments and industrials may prefer people not to know if such breaches are occurring. Our efforts, therefore, aim at a software solution to a ‘wet-ware’ problem: often, the ingenuity of humans in bypassing techniques aimed at avoiding such breaches can only be marvelled at.

The paper is organized as follows: we identify the challenge in the application domain (Section 2), then discuss our finding on different varieties of the Enron corpus (Section 3) and how this leads towards a filter for identification and subsequent removal of some of the information pollution. We discuss initial findings in relation to sentiment analysis in the Enron corpus using SentiWordNet (Section 4), and conclude by considering the future work to emerge from our findings.

2 Application domain: Can you keep a secret?

Data security is strategically important for the protection of government and military information and personnel, and is of growing importance in combating identity-based crimes such as fraud and cyber-stalking. It is usually the high profile breaches of data security that become newsworthy: the lost government laptop; the US military secrets on a memory sticks for sale at a bazaar in Afghanistan; the high street banks encouraging customers to shred bank statements while leaving un-shredded account details in rubbish bags, and the fraudsters recovering bank details from PCs sent to Africa for recycling. Of course, those volunteering their personal information to a world-wide audience, or to disreputable companies, may find a range of problems also (Jewkes 2003).

Emails have become a primary mode for asynchronous communication in modern business life. In the same way that an organization’s website, allied to effective use of search engines, provides a substantial market presence, emails can represent the organization in other ways. The benefits of effective use of email systems can be in carrying out and gaining trade through discussion, exchange, learning, contacts, contracts etc. Yet there is also a severe risk of loss of reputation, breach of confidence, loss of intellectual property, and loss of tactical and strategic business information. Aspects of human behaviour also present the risk of loss of reputation: organisations need to maintain corporate professionalism within emails leaving their organisation, and a careless or unguarded reply can rapidly bring embarrassment to individual and business alike. Machine-based monitoring of email communications, in a fair and timely manner, can *help* to avoid such lapses. And yet the technology to support this vital activity is extremely limited. Such a system should check outgoing emails for:

- Sensitive subject areas discussed in the body text,
- The wrong kind of sentiment,
- Sensitive attachments,
- Inappropriate addressees, (competitors, reporters etc.)
- Authors out of context,

and cope with all the vagaries of large numbers of unique sparse emails.

A capable system needs to be developed on the basis of a benchmark data collection. For us, this entails a good, freely available, corpus which preferably contains all the vagaries of human behaviour in a business context. Previous research has been undertaken on email corpora, primarily to detect and remove spam. Spam filtering aims at preventing the receipt of propagated emails. Many small corpora, and related publications, exist for such tasks including:

Table 1: Corpora typically used for the detection of Email spam

Email corpus	**Number of emails**	
	SPAM	**Non-SPAM**
SpamAssassin[4]	1897	4150
Synthetic (Annexia/Xpert) Corpus [Trudgian (2004)[5]]	10,025	22,813
LingSPAM [Androutsopoulos, et al (2000)[6]]	481	2412
GenSpam anonymised email/SPAM corpora [Medlock (2006)[7]]	32332	9072

Our efforts differ from those involved in spam filtering in that the action required for detection of an offending item requires more granular identification. For the spam detection task, an email is either allowed or blocked based on scoring mechanisms which usually include some form of naïve keyword filtering. For our task, an approach based on naïve keyword filtering could produce a torrent of false positives: for example, the simple expedient of including the keyword “confidential” would block all email responses with quoted content containing confidentiality banners. This would necessitate extensive human intervention, probably unnecessary from a technical perspective. This one keyword, amongst many, could be strongly indicative of a potential breach of confidence, so such intervention becomes highly necessary from an organisational perspective. Allied, also, to intentional breaches, the

[4] [http://spamassassin.apache.org/ and http://spamassassin.apache.org/publiccorpus/]
[5] [http://www.trudgian.net/spamkann/synthetic_corpus.php]
[6] [http://www.iit.demokritos.gr/skel/i-config/downloads/lingspam_public.tar.gz]
[7] [http://www.cl.cam.ac.uk/~bwm23]

significant potential for human lapses of judgement could result in incorrectly propagated, and hence harmful, messages. We are aiming, therefore, at preventing propagation or somehow intervening at point of production rather than at the point of distribution, i.e. before the email reaches a mailserver. It is clear that the context of the keyword is vital in automating judgements. The availability of Business Email collections on which to base such analysis is somewhat more limited; unsurprisingly given the potential for loss of competitive advantage.

The one widely available collection for such an effort is the Enron corpus. The Enron corpus contains a reflection of the day-to-day business, and sometimes a trace of the personal activities of the employees, for a large corporate. In its original form, 619,446 emails were reportedly available in folders of 158 users. A database comprising 92% of Enron's staff emails is supposedly available at the Federal Energy Regulatory Commission[8] along with a vast array of other documents relating to the investigations into Enron. It is not clear on what basis the 92% is calculated. Other researchers have asked questions about the integrity of the datasets, given the removal of some email account folders and some removal of duplicate records. The Enron corpus most readily available[9] [Enron-Raw], comprising 517,431 emails (approx 84% on number of emails), would still appear to be a useful collection for such analysis. The size, number of files per directory, duplications, attachments, odd character codes and rawness of the data within this corpus has caused difficulties for others wishing to perform analysis of this corpus. The Enron corpus has been used in related work, much of which has been concerned with data cleansing or classification. These researchers use varying numbers of emails or produce a new number of emails as a result of some cleansing activities, and subsequently these cleansed versions are used in yet other work. A sample of these studies is provided in Table 2, below, with brief details of the number of emails analysed.

[8] http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp

[9] http://www.cs.cmu.edu/~enron/

Table 2: Related work on the Enron corpus

Application	Corpus size (# emails)	Further Description
Automatic classification	12,500	Determining whether emails are for "Business" or "Personal" uses [Jabbari et al 2006], University of Sheffield, UK
Data cleansing, Preliminary analysis	200,399	Analysis of email threads and message distribution. Some folders removed. [Klimt and Yang 2004], Carnegie Mellon
Annotation; vizualization	1,700	Manual annotation of email categories. http://bailando.sims.berkeley.edu/enron_email.html University of California, Berkeley
	255,636	Visualisation and clustering. Use of database structure separating bodies, headers and other elements. http://bailando.sims.berkeley.edu/enron_email.html University of California, Berkeley /
Automatic classification	20,581	Automatic approach to building email folders [Bekkerman, McCallum and Huang 2004], Massachusetts Amherst
Data de-duplication	250,485	MD5 Hashes on body text to identify duplicates, resulting in 250,485 emails[10]. http://ciir.cs.umass.edu/~corrada/enron/ Massachusetts,Amherst
Social Network Analysis.	Not available	Link Discovery for Counter terrorism & Fraud. http://sgi.nu/enron/use.php?s=usc Southern California
Deception Theory	289,695	[Keila and Skillcorn, 2005b]. Queens, Canada.

The approach to manually annotated emails of Jabbari et al, building on prior work by Marti Hearst, is interesting for us since the 94% inter-annotator agreement suggests a large degree of differentiation is possible. Of the remainder, the objective of the email, as the authors identify, comes into question: business vs personal travel; purpose of inter-employee meetings, and so forth. As with much of the work on samples of the corpus, the basis for selection of these emails is not known – hence, the extent to which the sample is representative of the corpus is unclear. Furthermore, the automatic classifier appears to have been run on a smaller sample of 5000 emails, and using what appears to be the two extreme classes identified; this may contribute significantly to high performance figures. This work is interesting for us since we may be able to identify breaches of email policies where personal emails are forbidden, or where policies allow, to identify those unintentionally providing confidential information about themselves to wider audiences – at potential loss only to the sender.

Work on deception theory [Keila and Skillcorn 2005a], suggests that those intending to deceive using text as the only medium leave a particular linguistic trace. According to this theory, authors attempt to tell a simpler story and to disassociate from the story. For these researchers, such deception is traceable in text by fewer instances of first-person pronouns, fewer low-frequency words and increased frequency of both verbs and negative sentiments. The latter, which is of some interest for our work, would suggest that sentiment analysis and deception theory overlap – perhaps the analysis of negative movie reviews would assist in determining the overlap? The researchers investigate word frequencies with respect to the BNC but

[10] Large numbers of low entropy responses, for example, "yes" "no" "proceed" "thanks for that" and "see attached" may result in duplicate codes, and such work suggests further investigation is needed.

do not appear to have constructed a set of expected values for such items in general communication, or within given contexts and especially within email corpora. Without such expected values, it is difficult to know whether claimed results would be robust. One question, for us, would be whether it is even possible to differentiate between lexical cohesion due to repetition in well-formed and focussed arguments, or whether the increased frequency of certain words is an indicator of deception.

Our work as currently formulated is specifically aimed at providing a general model for avoiding the confidentiality banners – or more generally, corporate disclaimers. As work progresses, it will be interesting to investigate further the ability to classify and to discover deception, and this may stem from directly from our efforts. Other work, also, will become of greater interest. Currently, however, we are focused on "confidentiality banners" causing false positives for outgoing email filtering systems, with emphasis on protective markings as used, for example, by UK government departments. The distinctions between business and personal emails may inform, and be informed by, these efforts, and finer-grained deception analysis may help such efforts - or with the removal of these objects provide for a better input set for deception analysis.

3 Enron: initial analysis

Our experiments with the Enron corpus are aimed at characterizing the email collection [Enron-Raw]. We have undertaken analysis similar to that presented for using automatic ontology learning techniques [Gillam, Tariq and Ahmad 2005; Gillam and Ahmad 2005], and the use of local grammar discovery [Ahmad, Gillam and Cheng 2006]. Our experiments on the Enron corpus were originally intended to discover a basis for the Enron *ontology*, and subsequently to identify concepts of corporate importance within the ontology, as well as to detect sentiments about these concepts. We are now using these techniques, allied to others, in a slightly different orientation: to attempt to automate the cleansing of the data.

We have used Surrey's in-house Unix version of System Quirk, a package of software for tasks such as text analysis, ontology learning, and terminology and text management[11]. System Quirk implements simple frequency counts, keyword-in-context (KWIC) analysis, indexing and document frequency analysis, contrastive analysis with reference corpora producing smoothed "weirdness" values and the subsequent statistical generation of collocational patterns [Gillam, 2004]. Use of these complementary techniques has been demonstrated across a range of domains from nanotechnology to automotive engineering to financial trading [Gillam, Tariq and Ahmad 2005; Gillam and Ahmad 2005]. We augment these techniques with others developed in the course of our work and specific to the task at hand, in the expectation that these developed techniques will have broad utility subsequently.

According to our analysis, the Enron corpus [Enron-Raw] comprises some 209,204,013 tokens - excluding punctuation. This is twice the size of the British and Americal Nationals. This count is prior to any additional cleansing activities. In Table 3, below, we present the top twenty most frequent words, their frequency of occurrence as an absolute value and as a proportion of the corpus, and smoothed weirdness values in relation to the British National Corpus.

[11] A subset of these applications is available, though less powerful, for Windows at http://www.computing.surrey.ac.uk/SystemQ.

Table 3: Enron-Raw Frequency and Weirdness values

Word	Frequency	Relative Frequency	Weirdness
enron	7,555,888	0.0361	157198.32
com	6,881,814	0.0329	20710.76
the	5,684,275	0.0272	0.44
to	5,072,137	0.0242	0.95
x	3,654,791	0.0175	259.40
and	2,593,183	0.0124	0.46
of	2,391,399	0.0114	0.39
cn	2,332,235	0.0111	74399.72
a	2,146,189	0.0103	0.48
from	1,798,262	0.0086	2.08
in	1,759,898	0.0084	0.45
for	1,487,268	0.0071	0.84
on	1,268,134	0.0061	0.84
s	1,253,059	0.0060	56.62
o	1,243,397	0.0059	97.17
na	1,238,285	0.0059	37.66
is	1,225,139	0.0059	0.59
ect	1,201,951	0.0057	12503.14
you	1,192,850	0.0057	0.82
i	1,166,769	0.0056	0.62

Based on prior work on a number of corpora of varied sizes, we expect the weirdness values of grammatical words to be approximately unity, and in particular "the" to make up around 6% of the total: these results underperform expectations. The distribution of tokens covers 618,761 words, a similar distribution to frequency lists for the BNC. This underperformance is a concern, and the impacts of such low frequencies on related tools such as part-of-speech taggers could be significant. This shortage could, however, be explained in part by the dominance of emails headers.

For brevity and ease of presentation, we include two "tag clouds" (Figures 1 and 2). The first presents the top 100 most frequent words in the Enron corpus; the second presents the top 100 most weird. Both clouds are presented in alphabetic order with font size representing relative frequency and font colour representing weirdness. Low frequency words appear smaller and high frequency words appear larger, with low weirdness towards the blue and high weirdness towards red.

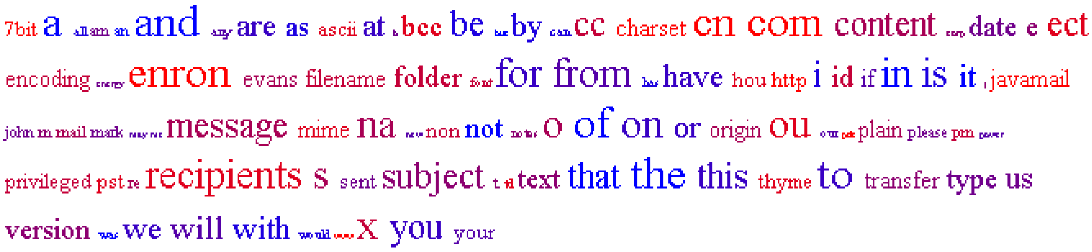


Figure 1: Tag Cloud of the Top 100 most frequent words

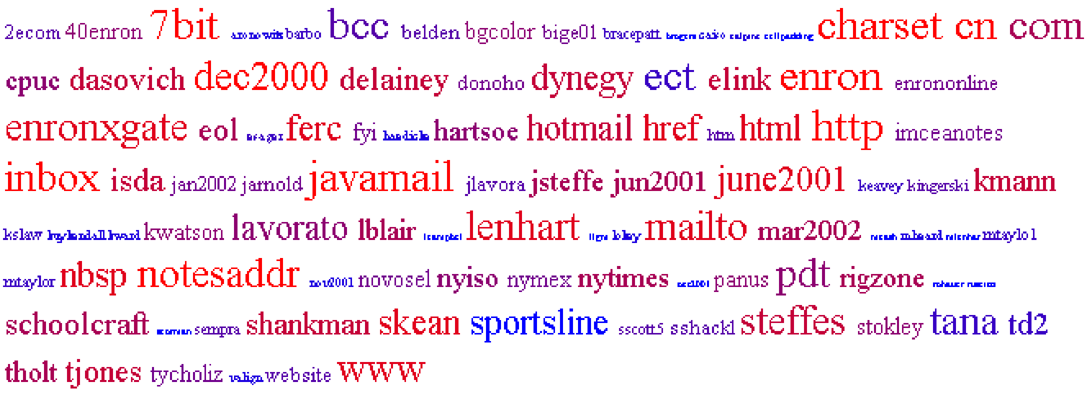


Figure 2: Tag Cloud of the Top 100 most weird words

It appears at first glance that the results for both sets are skewed towards noise, although without further analysis it is difficult to make conclusions over the highly frequent and highly weird occurrences of "enron" in the collection. High frequency grammatical words, as seen in the frequency cloud, are suppressed in the weirdness cloud. A number of employee names, evident in the folder names being analysed, can be seen in the weirdness cloud, including *delainey* (frequency: 14968; weirdness: 24501), *dasovich* (13651, 22345), *lenhart* (26565, 43485) and *shankman* (12752, 20874). These names could come from the email headers, or from the email body text.

We obtained the database from the Berkeley Enron Email Analysis Project and extracted the text of the emails from the database records [Enron-CleanUCB]. Analysis of this set shows that the total token count has reduced substantially from our original 209,204,013 tokens, to a mere 72,998,376; around 35 percent. The number of emails in this collection has dropped from the 517,431 to 255,636, suggesting that the creators believed over 50 percent of the emails to be duplicates. The frequency and weirdness values for the bodies of the emails [Enron-CleanUCB], presented alongside that for raw corpus [Enron-Raw] shows impacts of some of this cleaning activity: "enron" has dropped from 1st position to 8th , with an overall reduction in frequency of 70 percent; grammatical words have been boosted, but still underperform in contrast to our expectations, and this suggests the need for subsequent investigation.

Table 4: Enron-Raw and Enron-CleanUCB Frequency and Weirdness values

Enron-Raw				Enron-CleanUCB			
Word	***Frequency***	***Relative Frequency***	***Weirdness***	***Word***	***Frequency***	***Relative Frequency***	***Weirdness***
enron	7,555,888	0.0361	157198.32	**the**	2,806,643	0.0384	0.62
com	6,881,814	0.0329	20710.76	**to**	2,025,907	0.0278	1.09
the	5,684,275	0.0272	0.44	**r**	1,896,214	0.0260	383.93
to	5,072,137	0.0242	0.95	**and**	1,286,641	0.0176	0.66
x	3,654,791	0.0175	259.40	**of**	1,180,840	0.0162	0.55
and	2,593,183	0.0124	0.46	**a**	1,077,861	0.0148	0.69
of	2,391,399	0.0114	0.39	**in**	862,776	0.0118	0.63
cn	2,332,235	0.0111	74399.72	**enron**	766,304	0.0105	45689.91
a	2,146,189	0.0103	0.48	**for**	726,210	0.0099	1.17
from	1,798,262	0.0086	2.08	**com**	721,493	0.0099	6222.75
in	1,759,898	0.0084	0.45	**you**	624,933	0.0086	1.23
for	1,487,268	0.0071	0.84	**is**	613,752	0.0084	0.84
on	1,268,134	0.0061	0.84	**on**	597,406	0.0082	1.13
s	1,253,059	0.0060	56.62	**i**	593,368	0.0081	0.90
o	1,243,397	0.0059	97.17	**s**	553,435	0.0076	71.67
na	1,238,285	0.0059	37.66	**that**	541,559	0.0074	0.67
is	1,225,139	0.0059	0.59	**this**	458,320	0.0063	1.36
ect	1,201,951	0.0057	12503.14	**from**	424,235	0.0058	1.41
you	1,192,850	0.0057	0.82	**ect**	414,488	0.0057	12356.66
i	1,166,769	0.0056	0.62	**be**	413,709	0.0057	0.85

Again, we characterize this information as Tag Clouds to obtain a better overall view of Enron-CleanUCB. On the basis of frequency, we begin to see words such as "energy" and "power" attaining greater importance or creeping into the top 100. By weirdness, however, we see that that further noise, not least from what could be HTML, appears to remain.

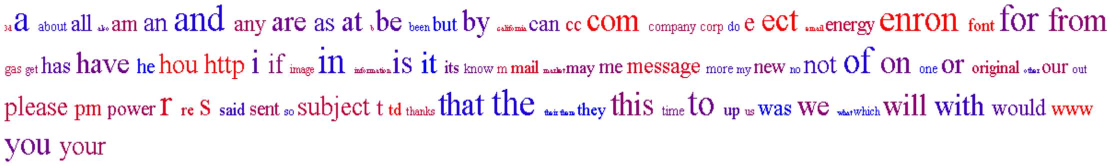


Figure 3: Tag Cloud of the Top 100 most frequent words

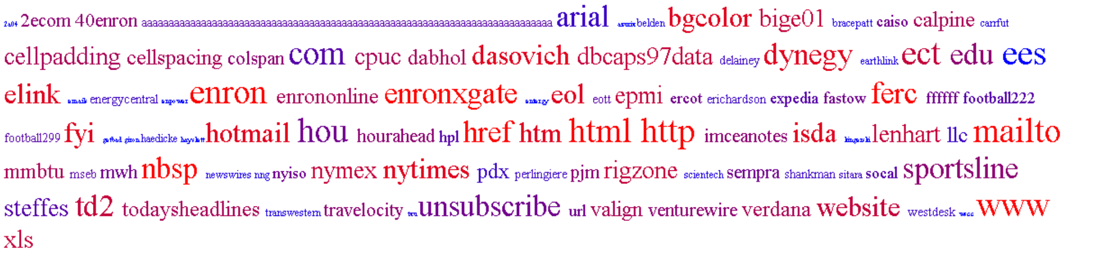


Figure 4: Tag Cloud of the Top 100 most weird words

The Enron-CleanUCB analysis demonstrates that further work is likely to be needed before the corpus could be clean enough to usefully extract context, subject matter and meaning, for example, for ontology learning. Content in HTML, for example, would need to be dealt with appropriately as suggested in the introduction to this paper. The Enron-CleanUCB collection can help us deal with email headers, even within the duplicated emails that have been removed from Enron-Raw, but we have already demonstrated the existence of further Email-specific elements that are also not useful. What we have not assessed, fully, is the rigour with which the removal of duplicated emails has been undertaken: we may already have lost a certain amount of useful content before we do more cleaning. Work on cleaning the corpus is, for us, ongoing. However, for our application domain we are concerned specifically with the pollution caused by the confidentiality banners. We investigate this using both datasets, in part to determine what information we could be losing if we were to use Enron-CleanUCB rather than Enron-Raw.

3.1 Confidentiality Banners

As noted above, the keyword “confidential” could be used to indicate material of a sensitive nature, but is now prevalent in email privacy banners. As such, even following the removal of email headers there will remain some proportion of content that is not interesting for analytical purposes – beyond, perhaps, understanding the structure of such banners. Besides, the act of removal suggests the need to understand their structure.

Enron-Raw contains 35,621 instances of the word confidential, and previous analysis [Cooke, Gillam and Kondoz (2007)] suggests that more than fifty percent of these are from banners. An example of such a banner is included below. Note the length of the banner could make up a large proportion of the text in brief messages and contribute to corpus pollution. The banners will also act to conceal the behaviour in free-running text, of a variety of other contained words.

```
++++++CONFIDENTIALITY NOTICE+++++
The information in this email may be confidential and/or privileged.
This email is intended to be reviewed by only the individual or
organization named above.  If you are not the intended recipient or
an authorized representative of the intended recipient, you are
hereby notified that any review, dissemination or copying of this
email and its attachments, if any, or the information contained
herein is prohibited.  If you have received this email in error,
please immediately notify the sender by return email and delete this
email from your system.  Thank You
```

Figure 5: Example Banner

Using “confidential” as the nucleate of our collocations, frequencies of collocating words in a 5-word window (L5-R5, with adjacent frequencies at L1 and R1) in Enron-Raw are as show in Table 5 (ordering in the Table is based on further analysis with these values, according to the work of Smadja 1993, but not presented here):

Table 5: Enron-Raw “confidential” collocations

Collocate	**Total**	**L5**	**L4**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**	**R4**	**R5**
and	30584	623	300	7512	1961	3310	15262	768	230	372	246
may	15264	307	1144	161	10290	5	74	2815	249	190	29
contain	11004	1613	205	428	7	8630	0	0	38	83	0
the	13886	459	1485	681	160	473	151	551	966	667	8293
for	10129	493	377	149	88	116	105	296	629	7228	648
privileged	19390	68	13	1375	4780	1647	71	6599	2593	1398	846
material	6367	2	0	30	0	0	21	22	5122	1153	17
relevant	4863	4856	7	0	0	0	0	0	0	0	0
information	11143	688	1013	704	271	123	5379	338	1111	715	801
affiliate	5051	185	4855	7	0	0	0	0	0	0	4

Collocations with “confidential” appear to suggest a pattern similar to that of the example confidentiality notice. A clearer pattern emerges with the simple removal of the 2000 most frequent words of the BNC (Table 6).

Table 6: Enron-Raw “confidential” collocations: BNC top 2000 removed

Word	**Total**	**L5**	**L4**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**	**R4**	**R5**
contain	11004	1613	205	428	7	8630	0	0	38	83	0
privileged	**19390**	**68**	**13**	**1375**	**4780**	**1647**	**71**	**6599**	**2593**	**1398**	**846**
affiliate	5051	185	4855	7	0	0	0	0	0	0	4
legally	4724	3	1117	342	71	0	70	139	2475	499	8
intended	3990	69	17	4	0	0	10	2535	516	570	269
exempt	2480	0	0	0	0	0	0	430	218	1832	0
proprietary	3097	147	107	1726	70	7	649	258	129	3	1
unauthorized	1415	0	0	0	0	0	0	0	8	0	1407
Solely	1399	0	0	1	0	0	0	9	1275	98	16
Email	2864	63	510	932	506	1	5	4	21	5	817

Significant peaks for “privileged” can be seen at L2 and R2. We cannot yet discount the possibility that there are substantial contributions to these values from body text.

We next investigate impacts on the collocation patterns following removal of email headers and some of the duplicates [Enron-CleanUCB] (Tables 7 and 8). The cleaning process has reduced the frequency of “confidential” from 35,621 in all of Enron-Raw to 19,297 in body text of Enron-CleanUCB – reduction of about 46%. In Enron-Raw, “privileged” co-occurs at R2 6599 times; around 19%. In Enron-CleanUCB, “privileged” now co-occurs at R2 5632 times; around 29%. We could speculate that a greater proportion of the remainder is now banners. However, a substantial reduction is now seen at L2

Table 7: Enron- CleanUCB "confidential" collocations

Keyword	**Total**	**L5**	**L4**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**	**R4**	**R5**
And	19931	377	188	5990	957	1628	9959	461	91	157	123
May	10051	166	642	88	7361	1	55	1547	119	63	9
Contain	7861	741	129	298	14	6617	0	0	23	39	0
The	9617	191	1001	337	83	199	77	257	516	296	6660
For	7651	323	216	53	27	44	46	121	413	6113	295
Privileged	**11747**	**35**	**41**	**579**	**2075**	**868**	**43**	**5632**	**1242**	**691**	**541**
Material	5468	3	0	15	0	0	12	8	4855	561	14
Relevant	4610	4603	7	0	0	0	0	0	0	0	0
Affiliate	4769	158	4602	7	0	0	0	0	0	0	2
Information	5867	471	497	425	84	58	2859	144	587	367	375

Table 8: Enron-CleanUCB "confidential" collocations: BNC top 2000 removed

Keyword	**Total**	**L5**	**L4**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**	**R4**	**R5**
Contain	7861	741	129	298	14	6617	0	0	23	39	0
Privileged	**11747**	**35**	**41**	**579**	**2075**	**868**	**43**	**5632**	**1242**	**691**	**541**
Affiliate	4769	158	4602	7	0	0	0	0	0	0	2
Intended	2383	36	10	2	0	0	5	1587	217	389	137
Legally	2131	4	472	216	36	0	29	85	923	364	2
Solely	928	0	0	1	0	0	0	5	867	48	7
Unauthorized	718	0	0	0	0	0	0	0	4	0	714
Corporation	621	0	0	1	620	0	0	0	0	0	0
Email	1583	34	314	464	369	1	1	2	17	3	378
Exempt	931	0	0	0	0	0	0	228	109	594	0

In comparison to Table 6, frequency of collocations with "affiliate" have dropped by a rather small margin of around 6%, while those for "legally", "email" and " have dropped by around 50%, with "exempt" dropping around 63%.

To understand the extent of impact of removing email headers, we analysed the header records in Enron-CleanUCB. Header records produce 646 instances of "confidential", and only 143 of these actually collocate with "privileged". Even considering that the original collection may contain double or treble this amount, a substantial proportion appears to be due to body text.

The values above appeared to be increasingly indicative of banners. To attempt to discover robust statistics for banner keywords, for about 40% of the raw corpus we obtained collocation statistics for use of the word "confidential", not considering the 2000 most frequent words of the BNC. In this subset, 14,384 instances of confidential were found. We further split this subset into four, on the basis of folder names alone, and looked at the proportions of collocates with "privileged".

Table 9: Enron-Raw 4 subset comparison, collocations centred on “confidential”

	Count	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
1	918	4	1	17	132	85	2	539	87	41	10
		0.44%	0.11%	1.85%	14.38%	9.26%	0.22%	**58.71%**	9.48%	4.47%	1.09%
2	1737	2	1	116	234	91	6	831	239	92	125
		0.12%	0.06%	6.68%	13.47%	5.24%	0.35%	**47.84%**	13.76%	5.30%	7.20%
3	3465	1	1	308	1377	270	13	824	465	107	99
		0.03%	0.03%	8.89%	**39.74%**	7.79%	0.38%	23.78%	13.42%	3.09%	2.86%
4	2002	12	0	128	364	125	8	922	212	83	148
		0.60%	0.00%	6.39%	18.18%	6.24%	0.40%	**46.05%**	10.59%	4.15%	7.39%
	8122	19	3	569	2107	571	29	3116	1003	323	382

The pattern “confidential X privileged” accounts for 39% overall. The four subsets tend towards slightly different patterns: for three of these, this pattern is rather higher, while for the fourth, the pattern “privileged X confidential” shows a peak.

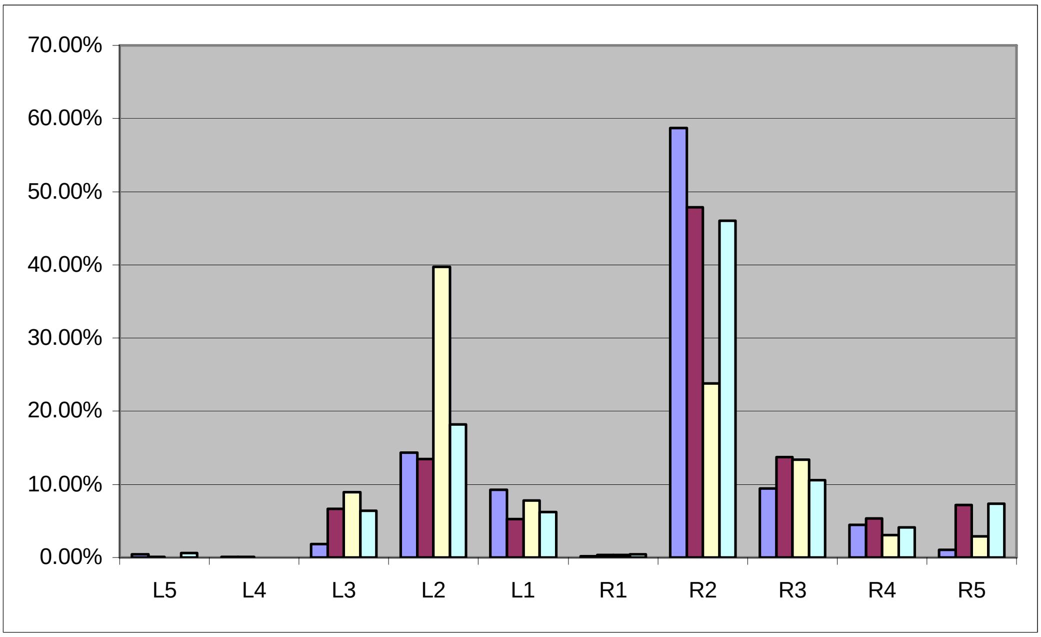


Figure 6: Enron-Raw comparison of collocation patterns for “confidential” and “privileged” across four subsets of a proportion of the corpus.

The intention here is to ascertain probabilities for “confidential” and “privileged” co-occurring at location *n*, and to determine the extent to which this holds across the corpus. Extending this further, to consider the probability of “prohibited” co-occurring with this pattern at location *m*, would improve confidence in detection. We would be attempting to discover a set of optimum values, or ranges, that can be used for confidence that the item we are dealing with is a banner. Given the variation seen above, we plan to continue this work to make determinations over the whole corpus and in contrast with other corpora.

3.2 Banner "Zoning"

Cleanly and efficiently extracting, delimiting, or otherwise removing banners is not a simple mechanistic process: according to our investigations, the banners appear to have some comparable structure, but do not follow a strict format according merely to email system protocols, as would be expected for email headers. Banners may be very different for each originating organisation. One may also have an expectation that email headers appear at the top of emails, and banners appear as footers, however the reality of quoted emails means that we could be searching for multiple instances of both throughout a given email. The challenge, then, is to identify the "zone" of each banner within an email and to successfully delimit it. This notion of zoning is inspired, in part, by Teufel's work on attribution of scientific text (Teufel and Moens 2000), and may be helpful in dealing with quoted responses.

In [Cooke, Gillam and Kondoz (2007)] we demonstrated the results of frequency analysis on 100 manually extracted instances of "confidential", comprising 50 unique banners and 50 non-banner paragraphs. Results of the analysis were compared to the BNC and to a subset of the Enron corpus. The manual extraction step demonstrated that banners consist of a large, but relatively limited set, of words, and in some instances account for a large proportion of the email body. Using a fixed-distance window and simple summation produced good discrimination for banner and body: 85.4 percent of banner instances were correctly identified and 0.37 percent of body instances were incorrectly identified as banners. In the application domain, body instances should be presented to a human for inspection, while banners incorrectly presented are of less importance than body instances not being presented (missed messages).

We expand on this work, analysing 3226 manually extracted confidentiality banners. We considered an expansion to distances of 120 words either side of our selected keyword as a means to detect the extent of the banner. This contrasts with traditional analysis of collocations, presented above, although we are using the same keyword set. On the basis of this analysis we can identify a clustering effect, with certain words dominant in particular positions, and suggest that banners are, on average, around 80 words in length (
Figure 7, Table 10). We can see two interesting peaks closely centred on "confidential":

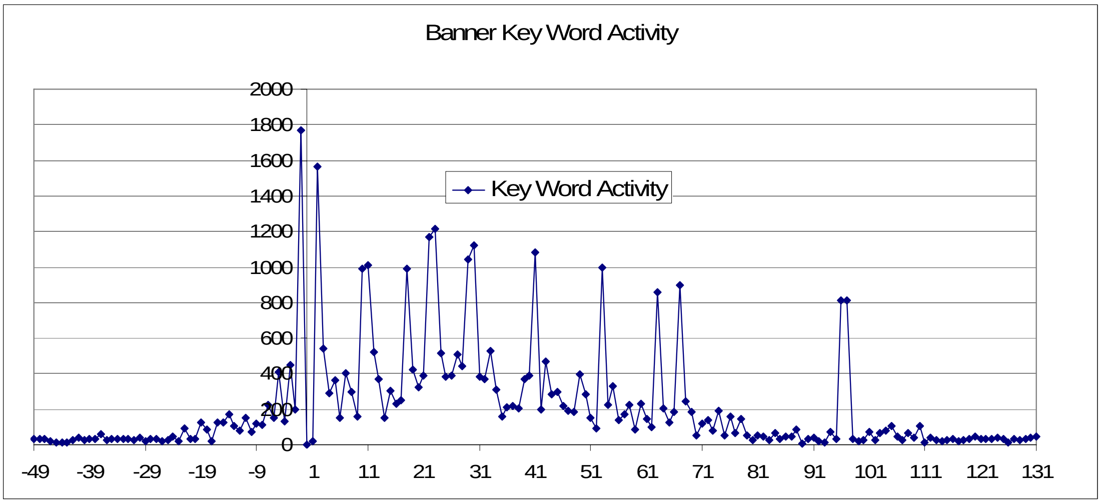


Figure 7: Key word activity surrounding "confidential"

Table 10: Key word activity peaks

Word	Distance
Contain	-1
Privileged	+2
Intended	+10, +29, +67, +96
Recipient	+11; +30, +97
Disclosure	+18
Strictly	+22
Prohibited	+23
Sender	+41
Delete	+53
Attachments	+63

There are two potential conclusions from the peaks identified above: (i) there are a lot of identical or very similar banners within the corpus; (ii) banners are large constructs with a predictable structure.

Analysis of the same keywords as above, centred on "confidential", for body text produces a substantially different result (Figure 8). The results show one peak, and further investigations have shown that the source of this is email correspondence with lawyers involved with litigation actions. The peak does not coincide with the banner instances for "privileged", and may partly explain the results in subset 3 of Table 9 at position L2.

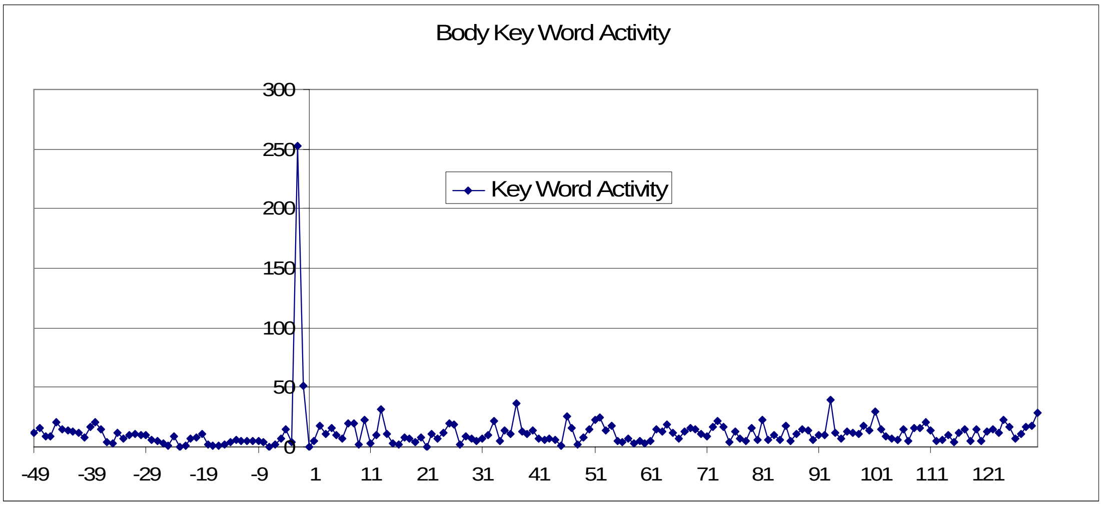


Figure 8: Key word activity surrounding "confidential" in ordinary body text

On the basis of this evidence, we have developed and evaluated an algorithm for discriminating banner and body text use of "confidential". The original algorithm described in our recent paper [Cooke, Gillam and Kondoz (2007)] has been modified for a window of –25 to +115 words, and scores according to keyword use with particular weight given to the use of "privileged" at L2 and R2. With the trigger point set to 2.75, this resulted in 91.7 percent of banner instances correctly identified and only 0.3 percent of body instances incorrectly identified (Figure 9), with minimal improvements at higher values.

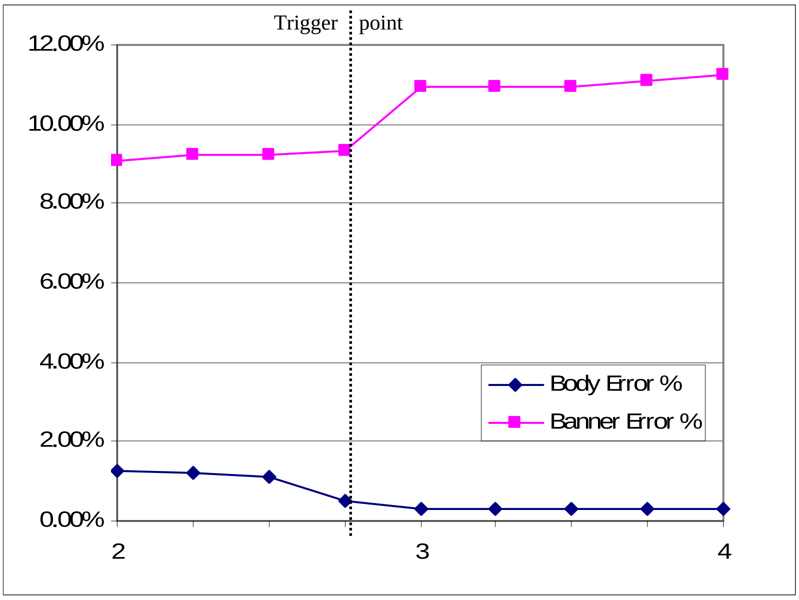


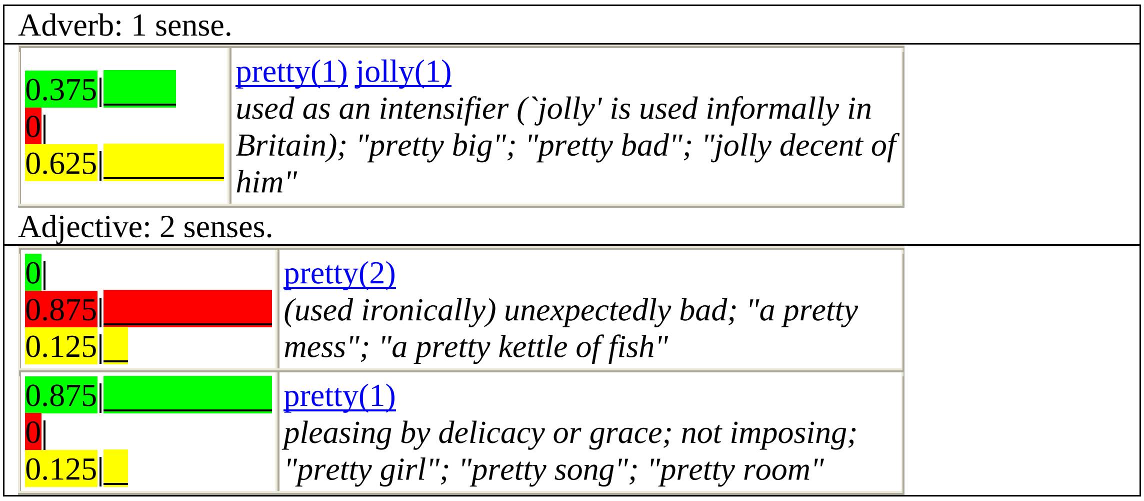
Figure 9: Misidentification percent against trigger point (total weight)

Assuming that our discoveries for collocation between "confidential" and "privileged" hold across the corpus, which we will be investigating, we should be able to remove a very high proportion (91.7 percent) of the 20,000 estimated banners from the corpus. Further evaluation, both manual and automatic, is planned, and the results will be published in due course.

4 Sentiment analysis

Having found that the Enron corpus was perhaps more polluted than hoped, results to date of our sentiment analysis are limited. Sentiment analysis, variously-titled opinion analysis, or affect analysis [Grefenstette et al., 2004], attempts to separate the subjective from the objective, and to produce values for the polarity of subjective statements associated to given "objects". In part, we were hoping to undertake further efforts in automatic classification in relation to the Enron corpus [Klimt and Yang 2004], and lessons learnt from the analysis of corpora of movie reviews [Pang, Lee and Vaithyanathan 2002]. The motivation for this work was to improve differentiation between appropriate and inappropriate business email, and personal email. Such analysis would be concerned not only with the identification of a personal email, but potentially the identification of workplace harassment and other inappropriate behaviours.

Our initial approach to sentiment analysis is coarse-grained: we obtained the SentiWordNet data [Esuli, A and Sebastiani, F (2006)] and produced average values across the senses to gain an overall average for each word. The web interface to SentiWordNet presents values for each of "positive", "negative", and "objective" where the sum of the three is 1 (specifically, "positive"+"negative"=1-"objective"). An example is given below for the word "pretty", with three senses. To the left are three values: for each sense, the first row (green) shows the value for positive sentiment, the second row (red) shows the value for negative sentiment and the third row (yellow) in each box is objective.



Adverb: 1 sense.	
0.375|🟩 0| 0.625|🟨	pretty(1) jolly(1) *used as an intensifier (`jolly' is used informally in Britain); "pretty big"; "pretty bad"; "jolly decent of him"*
Adjective: 2 senses.	
0| 0.875|🟥 0.125|🟨	pretty(2) *(used ironically) unexpectedly bad; "a pretty mess"; "a pretty kettle of fish"*
0.875|🟩 0| 0.125|🟨	pretty(1) *pleasing by delicacy or grace; not imposing; "pretty girl"; "pretty song"; "pretty room"*

Figure 10: Example entry for "pretty" from SentiWordNet

We produced overall values for sentiment in individual texts by multiplying values for average positive/negative/objective by the frequency of the words and creating totals for all three, ignoring those for which SentiWordNet does not have values. From this analysis, we further investigate texts that are highly positive, highly negative or highly objective. We expected, at minimum, to be able to use sentiment differences to help confirm the identification of banner text. Sentiment was calculated for 50 banners (containing "confidential") from the Enron corpus, with results in the ratio positive: negative : objective as 0.0953 : 0.0528 : 0.8519 (largely objective, slightly positive). Similar analysis on 50 body text items (containing "confidential") gave the ratio 0.0508 : 0.0407 : 0.9085 indicating even greater objectivity. By way of comparison, we also analysed contexts from the BNC where "confidential" occurred, with a ratio 0.0645 : 0.0460 : 0.8895 – still largely objective.

Low differentiation made us further investigate the SentiWordNet data. Aside from a variety of missing words, accounting for up to 50% of the vocabulary of our emails, a large number of words are annotated as completely objective (0 : 0 :1). Indeed, for the available words (203145-word sense pairs[12] which we have averaged out values for 144317 words), 76% are fully objective (73% in averaging). Positive and negative values relate only to around one word in four, and we have considered the frequencies with which positives (+), negatives (-) and objectives (o) occur. As plotted on a single graph (Figure 11), to see cumulative values between 0 and 1, the potential for strong negative sentiments seems limited. Relative scoring may be worth considering on such a basis.

[12] based on total count from WordNet 2.0, SentiWordNet's base data

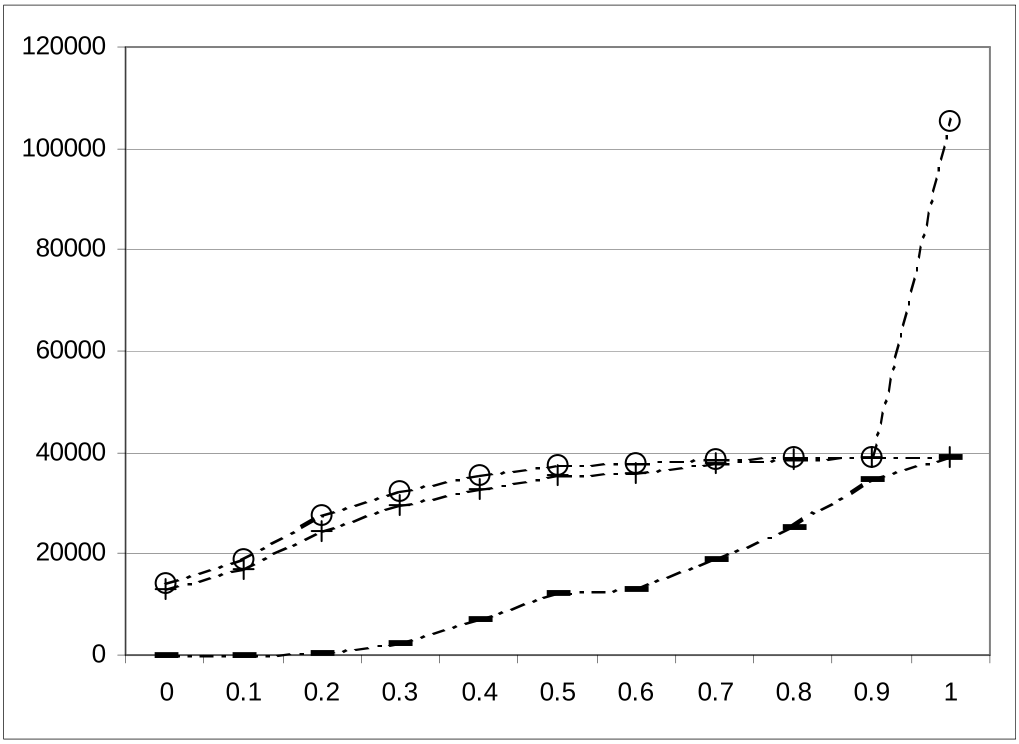
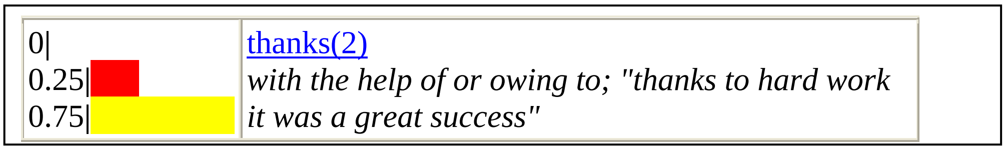


Figure 11: Frequencies of occurrence of sentiments within SentiWordNet

Furthermore, SentiWordNet values for some items may be confusing: a particular sense of "thanks", as below, attains negative scoring, potentially due to the phrase "hard work" in the definition.



0|
0.25| 🟥
0.75| 🟨

thanks(2)
with the help of or owing to; "thanks to hard work it was a great success"

Figure 12: Example entry for "thanks" from SentiWordNet

Understanding the distribution of proportions of sentiment words in SentiWordNet, and subsequently in resources such as the British National Corpus may help to gain a sense of comparability. Further efforts are needed on the Enron corpus and in relation to SentiWordNet.

5 Further work and conclusions

We have explored work undertaken on the Enron corpus, and one of the difficulties we have found is comparability. Our efforts to date have shown a number of differences between these collections on the basis of collocation patterns alone. These differences will have some impacts on subsequent research undertaken with supposedly cleaner versions of the corpus – and it is interesting to note that various cleaner versions exist, of varying numbers of emails. The challenge for any large-scale analysis is to fix the dataset and apply different analytical techniques. Applying different techniques to different datasets represents a challenge for those wishing to compare their work with that of others. Although the corpus linguistics community has come towards the idea of benchmarking data cleansing, the trace of cleansing activity is not currently available for the Enron corpus.

We are now preparing our analysis of the full corpus, or at least Enron-Raw. We hope to obtain the FERC version and make further comparisons between the different versions so that we understand the extent to which further results can be extrapolated. The intention of deriving an Enron ontology through work on collocation patterns, and of understanding sentiment in relation to the ontology has been necessarily reformulated to consider the pollution of the corpus, and to work towards a set of techniques for data cleansing. The ideal would be a configurable system in which researchers can selectively clean the collection and produce results that are immediately comparable with those who used the same cleansing routines, putting such work on par with other e-Sciences. Our attempts to identify confidentiality banners, deal with email headers, and subsequently to deal with other vagaries of email systems are steps towards this. Correct delimitation of the zone or zones occupied by banners within emails will help to ensure that we are dealing, more or less, with email *content*. Manual verification at various stages of the automation will be required, but with the intention of moving towards greater levels of automation. Work to date has demonstrated that automatic identification of banners in the "full" Enron corpus is highly possible, but will have to be provably accurate for use in mission-critical enterprises.

Aside from the continuation of work to date, identified in earlier sections, future directions for this work are many and various, and could incorporate aspects of word sense disambiguation to better identify sentiment and provide a sentiment ranking engine, and a variety of classification tasks are under consideration. We may pay greater heed, also, to the potential for deception theory to play a role in this research.

Results and achievements to date have been highly encouraging.

Acknowledgements

This work has been sponsored, in part; by the EU eContent project LIRICS (22236). The authors are grateful to Neil Newbold for the generation of the Tag Clouds.

References

Ahmad, K., Gillam, L. and Cheng, D. (2006) Sentiments on a Grid: Analysis of Streaming News and Views. Proc. of 5th Intl. Conf. on Language Resources and Evaluation (LREC).

Androutsopoulos, I, Georgios Paliouras, Vangelis Karkaletsis, Georgios Sakkis, Constantine D. Spyropoulos and Panagiotis Stamatopoulos (2000). Learning to filter spam email: A comparison of a naive bayesian and a memorybased approach. Workshop on Machine Learning and Textual Information Access 4.

Bekkerman, R., McCallum, A., Huang, G. (2004) Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora (Massachusetts)

Cooke, N, Gillam, L, and Kondoz, A. (2007), IP protection: Detecting Email based breaches of confidence IAS2007 Manchester.

Esuli, A, Sebastiani, F. (2006), Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation, Genowa, IT, 2006, pp. 417–22.

Gillam, L . (2004) "Systems of concepts and their extraction from text". Unpublished PhD thesis, University of Surrey.

Gillam, L., Tariq, M. and Ahmad, K. (2005) Terminology and the Construction of Ontology. Terminology 11(1), pp55-81. John Benjamins Publishing Company. ISSN 0929-9971; E–ISSN 1569–9994.

Gillam, L. and Ahmad, K. (2005). Pattern mining across domain-specific text collections. LNAI 3587, pp. 570–79.

Grefenstette, G., Qu, Y., Shanahan, J.G., Evans, D.A. (2004) "Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application". Proc. of Recherche d'Information Assistée par Ordinateur (RIAO).

Jabbari, S, Allison, B, Guthrie, D, Guthrie, L, (*2006)* Towards the Orwellian Nightmare Separation of Business and Personal Emails *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 407–411

Jewkes Yvonne Editor *(2003)* Dot.cons: Crime, Deviance and Identity on the Internet, *Cullompton:Willan Press, 256 pp., ISBN 1843920018.*

Keila, P.S, and Skillcorn , D.B. (2005a). Detecting Unusual and Deceptive Communication in Email. Queen's University, CA 411.

Keila P.S. and D.B. Skillicorn (2005b) Structure in the Enron Email Dataset. SIAM 2005 pp. 55–64.

Klimt, B. and Yang, Y. (2004) The Enron Corpus: A New Dataset for Email Classification Research. ECML 2004: 217–26

Medlock, B. (2006) An Adaptive, Semi-Structured Language Model Approach to Spam Filtering on a New Corpus, CEAS 2006.

Pang, B., Lee, L., Vaithyanathan, S.. "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79–86.

Smadja, F. (1993). "Retrieving collocations from text: Xtract". Computational Linguistics, 19(1) pp. 143–78. Oxford University Press.

Teufel, S and Moens, M. (2000) "What's yours and what's mine: Determining Intellectual Attribution in Scientific Text". Proc. of 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong.

Trudgian, D. (2004), Spam Classification Using Nearest Neighbour Techniques**.** In: Proceedings of Fifth International Conference on Intelligent Data Engineering and Automated Learning, IDEAL04, Exeter, UK, *Lecture Notes In Computer Science*, vol. 3177, pp. 578–85.