Authorship Attribution of E-Mail: Comparing Classifiers Over a New Corpus for Evaluation

Ben Allison & Louise Guthrie

Department of Computer Science, University of Sheffield
Sheffield, UK, S1 4DP,
b.allison@dcs.shef.ac.uk, l.guthrie@dcs.shef.ac.uk

Abstract

Until very recently, the email collections that have been available for research have been rather artificially created and consist of emails that contributors have chosen to make available. These collections serve very well for certain applications, but are certainly not representative of a person's email habits; thus, they have not been realistic resources for testing automatic techniques for detecting the author of an email. In contrast, the majority of collections available for detection of author make extensive use of out–of–copyright literary texts freely available on the Internet, and conclusions drawn from experimentation on these corpora are not necessarily applicable to email. The release of the Enron corpus provided a unique resource because it is largely unfiltered, and therefore presents a relatively complete collection of emails for a reasonably large number of correspondents. This paper describes a newly created sub-corpus of the Enron emails, which we suggest can be used to test techniques for authorship attribution, and further shows the application of three different classification methods to this task to present baseline results. Two of the classifiers used are are standard, and have been shown to perform well in the literature, and one of the classifiers is novel and based on concurrent work that proposes a Bayesian hierarchical distribution for word counts in documents. For each of the classifiers, we present results using six text representations, including use of linguistic structures derived from a parser as well as lexical information. The corpus is available at `www.dcs.shef.ac.uk/~ben`.

1. Introduction

Authorship attribution is an area of study which has attracted great interest, and for many different reasons: in some cases, scholars have sought identification of historic texts such as Shakespeare's plays (Merriam, 1998), the works of John Milton (Tweedie et al., 1998), and the Federalist papers (Mosteller and Wallace, 1964). In other cases, efforts have concentrated on identifying the authors of modern media, such as forum postings (Abbasi and Chen, 2005), newspaper articles (Diederich et al., 2000) and newsgroup postings.

Because of email's ubiquity as a communication tool, attribution of the authorship of email messages is also a critical problem. However, substantial collections of email have been extremely difficult to come by, as people are extremely protective of their personal communications; the release of the Enron collection of emails after the company's investigation by the FERC thus provided an invaluable tool for research.

The purpose of this paper is two–fold: firstly, we present a corpus of emails derived from this collection, split by author, for the purposes of evaluation of methods of authorship attribution. Secondly, we provide baseline results for various representations of the text using several different classifiers. Two of these classifiers are relatively standard: an SVM, which has been shown to be highly effective as a tool for authorship attribution and document classification more generally (Dumais et al., 1998; Abbasi and Chen, 2005), and a multinomial probabilistic classifier similar to those used in (McCallum and Nigam, 1998; Guthrie et al., 1994). The third classifier is based upon ongoing work to refine the multinomial classifier by hypothesising an alternate (and more expressive) distribution of word counts in documents (Allison, 2008), in a similar vein to (Madsen et al., 2005).

The rest of this paper is organised as follows: Section 2. describes the original corpus and the manner in which we create our corpus for email attribution; Section 3. presents the representations of the text with which we experiment; Section 4. describes the classifiers in some detail; Section 5. presents results of the experiments, and finally Section 6. ends with some brief concluding remarks.

2. The Corpus

The corpus we create for the purposes of this work is derived from the Enron corpus. The Enron email collection was seized during the FERC investigation into Enron's business practices, and contains email from approximately 160 employees. The data were originally released by the FERC, and this was followed by substantial work correcting integrity issues in the data, described in detail in (Klimt and Yang, 2004). The full dataset is available at `http://www.cs.cmu.edu/~enron/`.

From this raw dataset, we create a corpus using nine employees' mail, and select those nine on the basis of the largest outgoing mailboxes (that is, the "Sent Items" folders, since there is considerable duplication between several folders with "Sent" in the title, some of which are machine-generated). Box 1 gives an example of the format of messages in the corpus before preprocessing.

The resulting corpus is a nine–way authorship attribution problem, with 9688 messages in total. We remove message headers and signatures, and also discount all emails with fewer than twenty words (the result is that only the body text in the example would be kept). Preprocessing consists of selecting contiguous alpha-numeric strings as "words", normalising case, but no stemming or stoplisting (except in the experiments which explicitly use stems). The resulting corpus has 4071 emails, with between 174 and 706 emails per author. The corpus is available for download

Email Example 1 An Example of Email Before Preprocessing

```
Message-ID: <33403699.1075852498628.JavaMail.evans@thyme>
Date: Wed, 29 Aug 2001 15:00:11 -0700 (PDT)
From: d..steffes@enron.com
To: angela.schwarz@enron.com
Subject: RE: Direct Access
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Steffes, James D. </O=ENRON/OU=NA/CN=RECIPIENTS/CN=JSTEFFE>
X-To: Schwarz, Angela
X-cc:
X-bcc:
X-Folder: \JSTEFFE (Non-Privileged)\Steffes, James D.\Sent Items
X-Origin: Steffes-J
X-FileName: JSTEFFE (Non-Privileged).pst

Thanks for the info.  While there may be some differences in presentation,
there don't appear to be any basic differences in the analysis.  All of
the discussions of the legislation need to be aware that the passage is
highly uncertain.

I agree with the analysis that if you have never been DA and sign up today
there is a very large possibility that you will pay a surcharge
(assuming the legislation passes).

Let me know if there is anything else.

Jim
```

from www.dcs.shef.ac.uk/~ben.

After preprocessing, the average length of messages in the corpus is approximately 75 words, although we note that the distribution of message length is far from normal – the majority of messages are short, with the median message length being 44 words, and the mode length 22. The total size of the corpus is approximately 305,000 words.

3. The Text Representations

Many authors have noted that the best representation of an author's style is not the full bag–of–words that is so often used for other text classification problems. Generally, it is considered that an author's use of his full vocabulary is highly topic–dependent, and the stylistic signature derived from a full word set is not appropriate for identifying the author across topics.

To explore this issue with the corpus described above, we use several representations of text: the first is the bag–of–words representation most commonly used. Because we believe that longer–length n-grams have the potential to capture more style than single words, we also consider a bag–of–bigrams and a bag–of–trigrams. We also consider a bag of stemmed words, where the stemming is done using an implementation of the popular Porter stemmer (Porter, 1980). From a more stylistically enlightened perspective, and following on from the work initial work of (Burrows, 1987) and many subsequent studies, we also consider using purely closed–class words to represent a text, since the use of these words is believed to be far more invariant across topic.

Finally, we consider an ambitious stylistic representation derived from the output of RASP (Briscoe and Watson, 2006). One possible output of the parser, given a document, is the list of grammar rules from which the text was derived. For example, the parse of the sentence "John hit the ball" is shown in Figure 1, which corresponds to the following rules:

$$\begin{aligned} S &\rightarrow NPVP \\ NP &\rightarrow N \\ NP &\rightarrow Det\,N \\ VP &\rightarrow V\,NP \end{aligned}$$

Rules of the form above clearly indicate sentence structure, and can be viewed in the same way as words (and thus as "features" for a classifier) by representing a text with counts of the number of times each rule is used in the generation of the text. Such a scheme is proposed in (Baayen et al., 1996), but in conjunction with a corpus of hand-parsed text; in contrast, in this work all parsing is performed automatically, using the latest release of the RASP parser.

4. Classifiers

This work examines the effectiveness of three different classifiers on this resource, each of which uses all of the

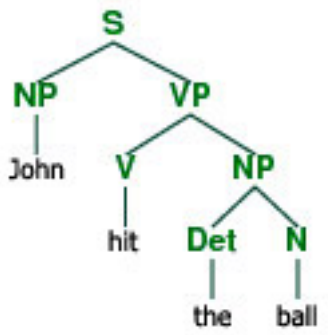

Figure 1: A parse tree for the sentence "John hit the ball"

six representations discussed in the previous section. Two of the three classifiers are probabilistic, that is they derive an explicit estimate for the probability that a new document belongs to each of the possible classes, and the third is a linear SVM.

The problem is presented as a mutually exclusive multi-class problem; that is, the task is to determine which of a possible nine authors wrote a new email. For the probabilistic classifiers, the extension to cases where there are more than two possible classes is trivial: for the SVM, the extension is a little more involved, but we use the most standard method for this purpose.

4.1. Multinomial Probabilistic Classifier

The multinomial probabilistic classifier has been widely used, and in many cases has been shown to perform robustly (see (Lewis, 1998) for an overview of early probabilistic classifiers, and (Guthrie et al., 1994; McCallum and Nigam, 1998; McCallum et al., 1998) for examples of the multinomial classifier applied to real problems).

In terms of notation, we use $\tilde{c}$ to represent a random variable and c to represent an outcome. We use roman letters for observed or observable quantities and Greek letters for unobservables (i.e. parameters). We write $\tilde{c} \sim \varphi(c)$ to mean that $\tilde{c}$ has probability density (discrete or continuous) $\varphi(c)$, and write $p(c)$ as shorthand for $p(\tilde{c} = c)$. Finally, we make no explicit distinction in notation between univariate and multivariate quantities; however, we use θ_j to refer to the j-th component of the vector θ.

We consider documents to be represented as vectors of count–valued random variables such that $d = \{d_1 ... d_v\}$. As with most other work, we further assume that words in a document are exchangeable and thus a document can be represented simply by the number of times each word occurs.

In classification, interest centres on the conditional distribution of the class variable, given a document. Where documents are to be assigned to one class only (as in the case of this paper), this class is judged to be the most probable class. Classifiers such as the probabilistic classifiers considered here model the posterior distribution of interest from the joint distribution of class and document; thus if $\tilde{c}$ is a variable representing class and $\tilde{d}$ is a vector of word counts, then:

$$p(c|d) \propto p(c) \cdot p(d|c) \tag{1}$$

For the purposes of this work we also assume a uniform prior on $\tilde{c}$, meaning the ultimate decision is on the basis of the document alone.

Note that the classification rule depends upon an estimate $p(d|c)$ for each possible value of $\tilde{c}$, which can be treated independently of one another. A natural way to model the distribution of word counts (rather than the presence or absence of words) for a fixed class is to let $p(d|c)$ be distributed multinomially; the multinomial model assumes that documents are the result of repeated trials, where on each trial a word is selected at random, and the probability of selecting the j-th word is θ_j.

Under multinomial sampling, the term $p(d|c)$ has distribution:

$$p_{multinomial}(d|\theta) = \frac{\left(\sum_j d_j\right)!}{\prod_j (d_j!)} \prod_j \theta_j^{d_j} \tag{2}$$

As is usual, given training data we estimate the vector θ as its posterior mean assuming a uniform Dirichlet prior. Thus if $\theta \sim Dirichlet(1...1)$ *a priori*, the posterior mean for the j-th component is:

$$\hat{\theta_j} = \mathrm{E}[\theta_j|\mathcal{D}] = \frac{1 + n_j}{v + \sum_j n_j} \tag{3}$$

where the n_j are the sufficient statistics from training documents, that is the total count of the j-th word in all documents from the class in question, and there are v words in the vocabulary.

4.2. Hierarchical Probabilistic Classifier

Following the recent successes of (Madsen et al., 2005), we also provide results using a hierarchical probabilistic classifier, but make several notable modifications to the method in that work—these modifications constitute ongoing work, and a more complete description can be found in (Allison, 2008). The hierarchical model we propose still assumes that all documents are samples from *some* multinomial, but rather than all being samples from the same multinomial

	Words	Bigrams	Trigrams	Stems	Closed-Class	Rules
Multinomial	74.31	75.83	61.51	78.46	43.06	44.19
Hierarchical	83.62	80.13	71.28	**87.05**	46.75	43.60
SVM	80.37	79.05	66.42	86.74	45.39	49.84

Table 1: Results of the three classifiers using different representations. Columns denote different representations, and rows different classifiers. All figures are percentage accuracy over 10 folds. The highest score is highlighted in bold.

(as above), we hypothesise that each is sampled from a different multinomial, and that each of the multinomials are themselves sampled from some other distribution.

(Madsen et al., 2005) suggest modelling the multinomials as having been sampled from a Dirichlet distribution, but such a modelling assumption has several drawbacks: chiefly amongst those is that under the Dirichlet, the expected value of θ_j and the variance of θ_j is functionally linked. Thus it is impossible to model the differences between words which are *on average* used the same amount, but where in some cases the use of the word varies significantly between documents, and in others it varies very little. This is a particularly important distinction for authorship attribution, where topical words have large variance in their underlying parameter (and thus are not so characteristic of an author) while truly stylistic words should have low variance in their parameter and be reliable indicators of authorship.

To overcome these problems, and following on from (Jansche, 2003; Lowe, 1999) who use similar models for individual words, we use a classifier which decomposes the term $p(d|c)$ into a sequence of independent terms of the form $p(d_j|c)$, and hypothesises that conditional on known class (i.e. c) $\tilde{d}_j \sim Binomial(\theta_j, n)$. However, unlike before, we also assume that $\tilde{\theta}_j \sim Beta(\alpha_j, \beta_j)$, that is $\tilde{\theta}_j$ is allowed to vary between documents subject only to the restriction that $\tilde{\theta}_j \sim Beta(\alpha_j, \beta_j)$. Integrating over the unknown θ_j in the new document gives the distribution of d_j as:

$$p_{bb}(d_j|\alpha_j, \beta_j) = \frac{n!}{d_j!(n-d_j)!} \times \frac{\mathrm{B}(d_j + \alpha_j,\, n - d_j + \beta_j)}{\mathrm{B}(\alpha_j, \beta_j)} \tag{4}$$

Where $B(\bullet, \bullet)$ is the beta function:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{5}$$

and $\Gamma(\bullet)$ is the Gamma function.

Then, the whole term $p(d|c)$ has distribution:

$$p_{beta-binomial}(d|\alpha, \beta) = \prod_j p(d_j|\alpha_j, \beta_j) \tag{6}$$

Given training documents from fixed class, we look to estimate the parameters of interest, in this case the $(\alpha_j,\ \beta_j)$ since the θ_j are integrated out, using a method–of–moments estimate similar to that in (Jansche, 2003).

4.3. Support Vector Machine Classifier

Our final classifier is a linear Support Vector Machine, shown in several comparative studies to be the best performing classifier for document categorization (Dumais et al., 1998; Yang and Liu, 1999).

Briefly, the support vector machine seeks the hyperplane which maximises the separation between two classes while minimising the magnitude of errors committed by this hyperplane. The preceding goal is posed as an optimisation problem, evaluated purely in terms of dot products between the vectors representing individual instances. The flexibility of the machine arises from the possibility to use a whole range of kernel functions, $\phi(x_1, x_2)$ which is the dot product between instance vectors x_1 and x_2 in some transformed space.

Despite the apparent flexibility, the majority of NLP work uses the linear kernel such that $\phi(x_1, x_2) = x_1 \cdot x_2$. Nevertheless, the linear SVM has been shown to perform extremely well, and so we present results using the the linear kernel from the SVM^{light} toolkit (Joachims, 1999) (we note that experimentation with non–linear kernels made little difference, with no consistent trends in performance). We use default parameter values for the SVM, and the most typical method for transforming the SVM into a multi-class classifier, the One-Vs-All method, shown to perform extremely competitively (Rennie and Rifkin, 2001). All vectors are also normed to unit length.

5. Experiments

We prepare the corpus as for experiments as follows: we define contiguous alpha-numeric strings to be words and case is normalised. For all representations, a document is represented as a vector of count–valued features, which for the case of the SVM is then normalised unit length. In all cases, we use ten–fold cross validation, where folds are assigned at random but the distribution of classes is kept. We report the most simple performance measure, accuracy, which is simply the total number of correct decisions divided by the corpus size (since all documents are test documents using ten–fold cross–validation).

The results of the experiments for each of the representations are shown in Table 1. The table shows that the hierarchical classifier generally outperforms the other two methods, and the SVM generally outperforms the multinomial classifier.

Perhaps more surprisingly still, the least stylistically enlightened approaches have far more discriminatory power than those which have been suggested as more suitable for capturing style. There are several possible reasons for this, and the overall effect is likely a combination of all of these: firstly, because the emails are short compared to many document classification scenarios, reliable statistics are difficult to obtain, and the fewer base units there are to model, the more accurately this can be achieved. Secondly,

the closed–class words and rewrite rules provide surprising amounts of information beyond the use of vocabulary, but clearly not sufficient in isolation. Finally, in order to achieve a relatively substantial corpus, it has not been possible to pick a collection which is entirely topic–invariant; thus there may be some element of topical association influencing the results.

6. Conclusion

Throughout this paper, we have described a new corpus for evaluating authorship attribution methods, and have also described baseline experiments to determine how well this task can be achieved. Despite the brevity of the messages (even when discounting extremely short emails), for certain representations the task can be performed with considerable accuracy.

The results of the preliminary experiments demonstrate some general trends: firstly, unigram representations appear to be generally superior to bigrams, and always to trigrams: thus while certain stylistic aspects are doubtless captured by the longer n–grams (i.e. use of phrases, an elementary representation of sentence structure) the extra and in many cases nonsensical grouping of words into phrases generally damages performance. Even more surprisingly, using word stems allows even better performance than single words: this goes somewhat against intuition, since one would expect that tendencies for different verb tenses, pluralization and so on would set authors apart. Finally, more complex linguistic features do not allow for such successful discrimination; we suggest that this is perhaps due to the extremely short and often informal nature of the communication, and suggest that perhaps methods designed particularly for short messages, and using resources customised to work with this particular form of writing, will allow even greater success.

7. References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

Ben Allison. 2008. *A Bayesian Model of Word Counts for Document Classification (in preparation)*. Ph.D. thesis, Department of Computer Science, University of Sheffield.

H Baayen, H van Halteren, and F Tweedie. 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Lit Linguist Computing*, 11(3):121–132.

J. Carroll Briscoe, E. and R. Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia*.

J. Burrows. 1987. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2(2):61–70.

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2000. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2):109–123.

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *CIKM '98*, pages 148–155.

Louise Guthrie, Elbert Walker, and Joe Guthrie. 1994. Document classification by machine: theory and practice. In *Proceedings COLING '94*, pages 1059–1063.

Martin Jansche. 2003. Parametric models of linguistic count data. In *ACL '03*, pages 288–295.

Thorsten Joachims. 1999. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning*.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Proceedings of ECML 2004*, pages 217–226.

David D. Lewis. 1998. Naïve (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98*, pages 4–15.

S. Lowe. 1999. The beta-binomial mixture model and its application to tdt tracking and detection. In *Proceedings of the DARPA Broadcast News Workshop*.

Rasmus E. Madsen, David Kauchak, and Charles Elkan. 2005. Modeling word burstiness using the Dirichlet distribution. In *ICML '05*, pages 545–552.

A. McCallum and K. Nigam. 1998. A comparison of event models for naïve bayes text classification. In *Proceedings AAAI-98 Workshop on Learning for Text Categorization*.

Andrew K. McCallum, Ronald Rosenfeld, Tom M. Mitchell, and Andrew Y. Ng. 1998. Improving text classification by shrinkage in a hierarchy of classes. In *ICML-98*, pages 359–367.

Thomas Merriam. 1998. Heterogeneous Authorship in Early Shakespeare and the Problem of Henry V. *Lit Linguist Computing*, 13(1):15–28.

F. Mosteller and D. L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist Papers*. Addison-Wesley.

Martin Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Jason D. M. Rennie and Ryan Rifkin. 2001. Improving multiclass text classification with the Support Vector Machine. Technical report, Massachusetts Insititute of Technology, Artificial Intelligence Laboratory.

Fiona J. Tweedie, David I. Holmes, and Thomas N. Corns. 1998. The Provenance of De Doctrina Christiana, attributed to John Milton: A Statistical Investigation. *Lit Linguist Computing*, 13(2):77–87.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In *22nd Annual International SIGIR*, pages 42–49, Berkley, August.