An Introduction to NLP-based Textual Anonymisation

Ben Medlock

Cambridge University Computer Laboratory
William Gates Building
JJ Thomson Avenue
Cambridge, CB3 OFD
bwm23@cam.ac.uk

Abstract

We introduce the problem of *automatic textual anonymisation* and present a new publicly-available, pseudonymised benchmark corpus of personal email text for the task, dubbed ITAC (Informal Text Anonymisation Corpus). We discuss the method by which the corpus was constructed, and consider some important issues related to the evaluation of textual anonymisation systems. We also present some initial baseline results on the new corpus using a state of the art HMM-based tagger.

1. Introduction

Statistical NLP requires training data from which to derive model parameters, and test data on which to execute and evaluate techniques. If such data is shared between researchers, comparisons of different approaches may be reliably drawn. However, this can be problematic if the data involved is sensitive in nature (eg. personal email). In such cases, an *anonymisation* procedure must be used to obscure the identities of actual entities revealed in some way by the text. In some cases, entities will be referenced directly, while in others, indirect but related references may betray their identity. The nature of these references will vary depending on the characteristics of the text, and whether a given reference is sensitive clearly requires a measure of subjective judgement. In some cases, the identity of the author of a piece of text may be revealed through the pragmatics of his/her use of language. This presents a somewhat different problem and is not addressed in this study.

The problem of *textual anonymisation* (the anonymisation of written language; henceform 'anonymisation' for brevity) applies not only in the case of data for NLP research, but also more widely in any area where textual data sharing is of benefit. For example, in the medical domain, information about the diagnosis and treatment of past patients can be used to inform current procedures and to establish statistical trends; however, such data often contains references to actual patients and must therefore be anonymised before it can be shared.

The cost of anonymising large data sets by hand is often prohibitively high. Consequently, data that could be widely beneficial for research purposes may be withheld to protect its authors against undesirable legal and personal repercussions. A potentially viable alternative to manual anonymisation is automatic, or semi-automatic anonymisation through the use of NLP technology, if the effectiveness of such a procedure can be reliably established.

This study offers three main contributions: firstly we present a description of the textual anonymisation problem and consider how the characteristics of the task affect the manner in which it is approached. Secondly we present a new corpus of personal email text as a benchmark for evaluating and comparing anonymisation techniques, with particular attention given to the semi-automated *pseudonymisation* procedure used to prepare the corpus for public release and the two annotation schemes used to represent different levels of sensitivity. Finally we discuss evaluation strategies and report some initial results using a state-of-the-art HMM-based tagger.

2. Related Work

There is little in the way of published literature on the topic of anonymisation in general, and no detailed studies of anonymisation methods using NLP technology. A number of articles have been written on privacy issues as they relate to the ethical storage and use of data (Clarke, 1997; Corti et al., 2000). Additionally, some researchers have highlighted the need for anonymisation in the area of automatic data mining and knowledge discovery (Wahlstrom and Roddick, 2001). Roddick and Fule (2003) propose a method for automatically assessing the sensitivity of mining rules which bears some relation to the task considered in this paper, though is not of direct relevance. Anonymisation has also been discussed in the context of the electronic storage of medical records (Lovis and Baud, 1999) and in relation to various other public data repositories, eg. (ESDS, 2004). Perhaps the most comprehensive study of anonymisation is carried out by Frances Rock (2001). She considers many aspects of the problem, highlighting both the reasons why corpora anonymisation is important and the particular nuances of the task. The following issues (amongst others) are addressed:

- Prevalent attitudes towards anonymisation amongst linguistic researchers
- Potential personal and legal implications of publicly available unanonymised corpora
- Which references should be anonymised
- Options for replacing sensitive references

3. The Anonymisation Task

We define the anonymisation task in terms of the following concepts:

- *token*: a whitespace-separated unit of text
- *document*: an ordered collection of tokens

removal:	*Joe Bloggs works at Somerton Bank* $\longrightarrow$ <REF> *works at* <REF>
categorisation:	*Joe Bloggs works at Somerton Bank* $\longrightarrow$ <PER> *works at* <ORG>
pseudonymisation:	*Joe Bloggs works at Somerton Bank* $\longrightarrow$ *Phil Day works at Higgins Bank*

Figure 1: Example of anonymisation processes

- *reference*: a span of one or more tokens used by the author to refer to a concept outside of the language
- *sensitivity*: a binary measure determining whether or not a particular reference, if publicly disclosed, might potentially cause harm or offence and thus engender undesirable personal or legal repercussions

Given these premises, we present the following definition:

> ***Anonymisation* is the task of identifying and neutralising sensitive references within a given document or set of documents.**

The task of anonymisation can be seen as a two-stage process. Firstly, sensitive references must be identified, and secondly they must be neutralised. In this context, neutralisation means obscuring the link provided by a given reference to an actual entity by means of:

- *removal*: replacing a reference with a 'blank' placeholder
- *categorisation*: replacing a reference with a label in some way representing its type or category.
- *pseudonymisation*: replacing a reference with a variant of the same type

Figure 1 gives an example of each of these techniques. Note that the identification phase involves an implicit sensitive/non-sensitive classification as well as detection of the reference boundaries.

Because sensitive references are usually those that refer directly to real-world entities, anonymisation is quite similar in nature to the task of Named Entity Recognition (NER) which has received significant attention in recent years, and we would expect similar ideas to find application in both areas. It might be appealing to consider anonymisation as a special variant of NER; however, the tasks are not strictly subsumptive:

- Sensitive references are not necessarily named entities. For instance consider the following sentence:

  > *John Brown, the long jump record holder, retired yesterday.*

  The constituent phrase *long jump record holder* betrays the identity of the named entity *John Brown* and is therefore a sensitive reference, though it is not a named entity itself.
- NER operates on the basis of objective judgements about the nature of referent entities (*Cambridge* is a place) whereas anonymisation relies on subjective judgements about referential sensitivity (*Cambridge* may or may not be a sensitive reference).
- NER is the process of identifying and classifying entity references, whereas anonymisation can include removal or pseudonymisation.

The inherent subjectivity of anonymisation means that different instances of the task may exhibit different characteristics even within the same domain. In light of this, it is probably impractical to deploy a solution requiring a large amount of annotated training data, bearing in mind that such training data may not generalise within the same domain, let alone across domains. In reality, application of an NLP-based anonymisation procedure would probably be carried out on an instance-by-instance basis, with rapid adaptation to the characteristics of the required solution through the use of interactive, weakly-supervised machine learning techniques.

Another important factor when considering the application of previous research into NER to the anonymisation problem is that NER has traditionally been carried out in the newswire domain where quite strict grammatical and orthographic conventions are observed and where the range of entity references tends to be quite limited. Conversely, the data that we present as a testbed for anonymisation is informal email text, where the use of grammar and orthography is highly colloquial in nature and there is a wider range of entity references (see 4.3.).

4. Corpus

We have assembled a publicly-available[1] data set, dubbed ITAC (Informal Text Anonymisation Corpus), as a testbed for the anonymisation task.

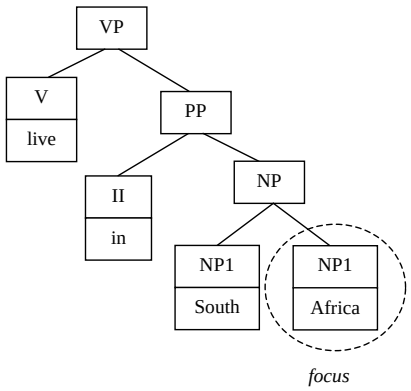
4.1. Corpus Construction

The corpus is comprised of approximately 2500 personal email messages collected by the author over a seven-year period divided as follows:

- Training set: 666,138 tokens, pseudonymised, unannotated
- Test set: 31,926 tokens, pseudonymised, annotated
- Development set: 6,026 tokens, pseudonymised, annotated

The authorship of the text is highly varied, with both private and corporate communication represented, and the language and orthography consequently exhibits much variability. Capitalization and punctuation are often used inconsistently and in many cases are entirely absent, making reliable sentence boundary detection difficult. Though some automatic sentence boundary detection techniques were investiaged and a significant amount of time was spent manually delimiting sentences, the final data set (especially

[1] http://www.cl.cam.ac.uk/users/bwm23/



orthography:	First letter capitalised
part of speech:	"NP1"
parent constituent label:	"NP"
inner left constituent label:	"NP1"
2nd inner left constituent label:	∅
inner right constituent label:	∅
outer left constituent label:	"II"
outer left constituent token:	"in"

Figure 2: Feature set example

the training data) still contains many spurious sentence boundaries. Additionally, the text contains many spelling errors and inconsistencies, as well as a great deal of pragmatic grammar usage. Whilst such issues increase the difficulty of the task, they are to be expected when working with informal text.

4.2. Pseudonymisation

Releasing data for the anonymisation task introduces an interesting conundrum: a realistic anonymisation testbed relies on sensitive experimental text with references preserved to facilitate the task, yet such text, in its original form, requires anonymisation before it can be publicly released. We overcome this problem by using a hybrid semi-supervised and manual *pseudonymisation* procedure to anonymise sensitive references without changing their nature. The procedure uses syntactic and orthographic features to cluster more obviously sensitive terms (such as person names) into semantically coherent groups and then randomly chooses replacement pseudonyms appropriate to the semantic category of the cluster, as specified by human input. The text is then scanned manually to identify and pseudonymise more complex sensitive references.

We use the RASP parser (Briscoe and Carroll, 2002) to generate the feature set for term clustering. The following syntactic and orthographic features are used:

- *part of speech*: a token's part of speech, as assigned by the RASP PoS tagger
- *inner left constituent label*: the label of the focus constituent's left sister
- *2nd inner left constituent label*: the label of the focus constituent's left-but-one sister
- *inner right constituent label*: the label of the focus constituent's right sister
- *outer left constituent label*: the label of the terminal constituent directly preceding the scope of the focus constituent's immediate ancestor
- *outer left constituent token*: the surface form of the terminal constituent directly preceding the scope of the focus constituent's immediate ancestor
- *orthography*: set of nine non-mutually exclusive orthographic features:
  - First letter capitalised (eg. *Mary*)
  - All letters capitalised (eg. *BARGAIN*)
  - Single capital letter (eg. *I*)
  - Integer-like number (eg. *01985*, token length also part of the feature)
  - Float-like number (eg. *12.75*)
  - Contains non-alphanum char (eg. *Yahoo!*)
  - Contains period (eg. *S.W.A.T.*)
  - Contains hyphen (eg. *26-year-old*)
  - Contains an upper/lower case or alphanumeric mix (eg. *BigSplash*, *win2000*)

Figure 2 illustrates the use of these features via an arbitrary syntax tree fragment. Potentially sensitive terms with identical features are clustered, and each resulting cluster is presented to a human annotator, who classifies the whole cluster as either sensitive or non-sensitive and labels it with a semantic category if appropriate. An example of such a cluster is given in Figure 3.

Because many of the more obvious sensitive references appear in similar contexts, labeling an entire cluster saves much time over annotating individual examples. When a ceiling on annotation cost has been reached, the data is scanned using the information acquired through the annotation process and pseudonyms are automatically generated (by random selection from previously compiled gazateers) for all references that have been identified as sensitive and labeled with a semantic category. Pseudonyms are chosen under the constraint that a given term is always replaced by the same pseudonym. This preserves the distribution of sensitive terms across the corpus, an important characteristic of the data. The automatically generated pseudonyms are then propagated through the text to minimise the number of cases missed due to sparse feature sets.

Because of the nature of the text, only firstname, surname and certain location names can be safely pseudonymised by automatically generated replacements. Names of organisations, for instance, often contain terms that cannot be automatically pseudonymised without changing the concept conveyed. For example, *The Financial Times* must be replaced with a phrase that carries a similar conceptual idea, while obscuring the identity of the actual organisation. This is a subtly difficult task and cannot reliably be carried out automatically. Consequently we spent a number of days manually generating pseudonyms for such instances and scanning the entire corpus for other references that might betray the identity of actual entities.

Cambridge	NP	the city of Cambridge hencewith on	Bath / York / Cambridge

Figure 3: Annotation example

An overview of the process is as follows:

- Parse text with RASP (Briscoe and Carroll, 2002).
- Generate feature sets for potentially sensitive terms.
- Cluster terms by feature equivalence.
- Present clusters to human for sensitivity and type classification.
- Generate pseudonyms of same entity type for specified references.
- Propagate pseudonyms throughout text.
- Examine text for missed sensitive references and manually generate replacement pseudonyms.

Note that due to the predictable coding of email addresses, URLs and date/time references, we do not consider them as part of the anonymisation process for the purposes of this study; rather we identify and anonymise them beforehand using regular expression matching.

4.3. Annotation

In light of the subjectivity of the sensitivity measure, we use two annnotation schemes to represent different views of what constitutes a sensitive reference. In the first, which we will call *blanket* anonymisation, we label as sensitive every reference that could potentially be used to trace the identity of a person or organisation, even if the chance of undesirable personal or legal repercussions is small. References of the following nature are included:

- Person, organization and location names and descriptors
- Postal addresses and telephone/fax numbers
- Commercial titles (*Yahoo!*)
- Film, TV and book titles (*Star Wars*)
- Job titles (*Director of Graduate Studies*)
- Geographic/ethnic terms (*S. African, Hebrew*)
- Titles of academic papers (*A study of text classification methods*)
- Course titles (*Computer Science*)
- Conference titles (*5th Conference on Gene Identification*)
- Usernames/passwords
- Transactional identification/reference codes

This is not a definitive list, but covers most of the types of reference found in our corpus.

The second annotation scheme, which we will call *selective* anonymisation, involves labelling only those references which relate directly to a person or organisation and thus consitutes a minimum level of anonymisation. These include:

- Person and organization names and descriptors
- Postal addresses and telephone/fax numbers
- Commercial product names
- Usernames/passwords
- Transactional identification/reference codes

Whilst the risk of traceability may be increased under this scheme, reduced intrusion means advantageously less distortion of the data.

Manually annotated versions of the development and test data sets are provided using both schemes, while the training data is supplied only in unannotated form. Whilst this may appear to be a limitation on the usefulness of the corpus, as explained in Section 3 we do not expect solutions to the anonymisation task to rely on large amounts of annotated training data, thus the ITAC corpus lends itself to experiments with weakly-supervised machine learning techniques where guided learning strategies are used to choose the most informative training samples for annotation from the unlabeled training pool.

The current annotation schemes contain no entity class information, thus limiting experiments to the identification/removal variant of the anonymisation task. Class information could be added to the existing sensitivity annotation schemes, either by ourselves or others, and this would facilitate experimentation into the identification/classification variant of the task.

4.4. Format

The corpus is formatted on a one-sentence-per-line basis (though due to boundary detection errors, sentences are sometimes split over multiple lines). The data is tokenised using the RASP tokeniser, which is based on a small number of regular expressions compiled using *flex*[2]. Orthography and punctuation are preserved as far as possible and codified references (such as email addresses) are represented by &*REF_TYPE* (eg. &EMAIL). Annotation is added in the form of <ANON> ... </ANON> tags that delimit sensitive references. Figure 4 shows a small sample from the blanket annotated version of the test data set.

```
From : " <ANON> Lance Malone </ANON> " ( &EMAIL )
To : " <ANON> tabitha ropp </ANON> " ( &EMAIL )
Subject : An email
Date : &DATE &TIME +0100
<ANON> Tabitha </ANON> ,
I can see absolutely no reason for your blank emails .
Can you see this one ?
I suppose you can because you 're reading this .
I 'VE FINISHED WORK ! ! ! ! !
I had a pretty hectic day today .
There was really too much to finish .
Still .
Have a relaxing weekend .
Doing anything interesting ?
<ANON> O </ANON>
```

Figure 4: Sample ITAC representation

5. Possible Approaches

HMM-based techniques have recently proven successful for NER, eg. (Zhou and Su, 2002), though they usually rely on a reasonably sized, previously annotated training corpus, and while informed data selection techniques such

[2] http://dinosaur.compilertools.net

	Blanket			Selective		
Eval scheme	Recall	Precision	F_1	Recall	Precision	F_1
Token	70.78	77.28	73.87	62.17	65.92	63.99
MUC	65.56	77.02	70.83	56.51	64.18	60.10
CoNLL	60.89	71.47	65.76	50.99	58.34	54.42

Table 1: Results

as *active learning* (Cohn et al., 1995) can potentially reduce the amount of requisite annotation, eg. (Scheffer et al., 2001), further research is required to demonstrate how much reduction can actually be achieved in terms of the true cost of annotation.

We suspect that an iterative, weakly-supervised learning procedure (such as an active learning variant) which makes use of available unannotated training data in an annotation cost-efficient manner is the most sensible way to approach the anonymisation problem, though further investigation is left for future research.

5.1. Lingpipe

To provide some initial experimental results on the new corpus, we train an HMM-based tagger on the labeled ITAC development data and evalute it on the test data. We use the Alias-i *Lingpipe*[3] tagger which has achieved state of the art results on a number of well-known test corpora for the NER task in both the newswire and biomedical domain. We expect the type of features chosen for NER to work reasonably well for anonymisation, and present these results as a baseline for further study.

6. Results and Analysis

6.1. Evaluation Measures

Evaluating the anonymisation task raises issues similar to those found in NER evaluation. Complications arise due to the comparison of boundaries and partial matches. Arguably the simplest strategy is to evaluate the sensitivity of each token on an individual basis, with recall, precision and F_1 defined in the usual manner:

$$\mathrm{r} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad \mathrm{p} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \qquad \mathrm{F}_1 = \frac{2\mathrm{pr}}{\mathrm{p}+\mathrm{r}}$$

where

TP = count of sensitive tokens correctly identified
FN = count of sensitive tokens missed
FP = count of non-sensitive tokens spuriously identified as sensitive

In one sense, this is a well-motivated approach, bearing in mind that a partially-anonymised reference is increasingly hard to identify as more of its constituent terms are anonymised, eg. *Lee ... ...* is preferable to *Lee ... Oswald*.

However, the anonymisation task is actually defined in terms of discrete references, not individual tokens, so arguably it is better to evaluate each referential span as a single item. This raises the question of what to do if a reference is only partially identified (eg. *Smith* instead of *Will Smith*) or if the boundaries are too wide, or crossing (eg. *Jack 's house* instead of just *Jack*).

One approach that attempts to take some of these complications into account is the scheme specified in the MUC guidelines for evaluating information extraction tasks where multiple word spans can represent single items of information.[4]

Another currently popular approach is that used by the CoNLL community for the NER task, where no credit is given unless an entity reference is fully and correctly identified. In this scheme a partially identified reference is counted both as a false negative and positive. Consequently, overall scores tend to be significantly lower than in either the token-level or MUC evaluation schemes.[5]

To facilitate as broad a range of comparison as possible, we report recall, precision and F_1 under all three evaluation schemes. We have developed our own evaluation scripts for the token-based and MUC schemes and use the CoNLL evaluation script available from the CoNLL website.

6.2. Results

Table 1 shows the results for the blanket and selective anonymisation tasks. The selective variant is harder due to the fact that there are often only fine-grained contextual distinctions between sensitive and non-sensitive references. For instance, if the term *Munich* appears as part of an address it is considered sensitive, whereas if it appears in the context of free text, eg. *I was in Munich last Friday*, it is considered non-sensitive. Conversely, in the blanket case, it would be considered sensitive in both these contexts. Instances such as these must be differentiated via their context, which significantly increases the sparsity problem, especially in the presence of limited training data.

The tagger is trained only on the development set, and would no doubt perform significantly better had it access to more training data. Adding samples from the training set (unutilised in the experiments reported here) in a cost efficient manner is an obvious subsequent experimental step to be explored in the future.

6.3. Error Analysis

As might be expected, most errors are caused by terms that are orthographically misleading. Some commonly problematic instances include:

- Complex, sensitive references containing many commonly non-sensitive terms, eg. *the Royal Society for the Protection of Birds*. Some of the terms in this reference are clearly non-sensitive in general (*the*) and capitalisation of common nouns in the email domain (*Pro-*

[3] www.alias-i.com/lingpipe/

[4] www.itl.nist.gov/iaui/894.02/related_projects/muc

[5] www.cnts.ua.ac.be/conll/

tection, *Birds*) is not particularly suggestive of sensitivity as it is often used simply for emphasis (*Get the New Version of Messenger!*).

- Uncapitalised sensitive terms that look like common non-sensitive terms, of which there are numerous instances in informal text (*penny*, *windows*, *west road*)
- Capitalised references to non-sensitive entities (*New Year*, *God*, *Hemisphere*)
- Non-sensitive references and turns of phrase that do not refer to real world entities yet are functionally and orthographically indistinct. (*the Bogey Man*, *Bill No Mates*)

7. Discussion

Anonymisation is a complex issue. Any distortion and ensuing loss of information is likely to have some impact on the usefulness of a given dataset, and in each case a decision must be made as to whether any form of anonymisation is feasible. For instance, removing brand-names from email text for spam filtering research might be considered an unacceptable distortion of the nature of unsolicited email, and could thus be seen to jeopardise the validity of research into the problem.

Bearing in mind the current state of NLP technology, it is clear that automatic textual anonymisation must realistically be viewed as an aid to, rather than a replacement for manual anonymisation of sensitive data. An automatic procedure cannot guarantee 100% reliability, even if the parameters of the task can be clearly defined (which is not always the case for anonymisation), and some form of manual checking will need to be carried out to validate the results of the procedure, most importantly to neutralise sensitive references that have evaded detection.

If a probabilistic model is employed (either native or derived) it would be helpful if the final model parameters could be used to point the validator toward uncertain instances, as these represent the boundary cases where misidentification is most likely to have occured. It would then be up to the validator to decide whether or not he/she can 'safely' ignore instances lying further from the decision boundary. In light of this, when evaluating a probabilistic anonymisation procedure it would be informative to know what percentage of misidentified instances lie near the decision boundary, and also the concentration of misidentified instances in this area (for in the limit *all* remaining instances might be located near the decision boundary, in which case such information is meaningless to the validator). In reality an approach in which misidentified instances occur in high concentration around the decision boundary is likely to be more useful than an approach that achieves greater accuracy but cannot reliably point the validator toward potential misidentifications.

8. Conclusions

We have presented a formal description of the automatic textual anonymisation problem and considered how the characteristics of the task affect the manner in which it is approached. We have presented a new corpus of personal email text as a benchmark for evaluating and comparing anonymisation techniques within this domain, and outlined the semi-automated *pseudonymisation* procedure used to prepare the corpus for public release. Finally we have reported initial results for the task using the *lingpipe* HMM tagger. We hope that this study will raise awareness of the issue of anonymisation and spur further research into methods for tackling the task and discussion into alternative perspectives on the nature of the problem.

9. References

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1499–1504.

Roger Clarke. 1997. Privacy and dataveillance, and organisational strategy. *Proceedings of the Region 8 EDPAC'96 Information Systems Audit and Control Association Conference*.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1995. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press.

Louise Corti, Annette Day, and Gill Backhouse. 2000. Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. *Forum for Qualitative Social Research*, 1(3).

ESDS. 2004. The economic and social data service: Identifiers and anonymisation: dealing with confidentiality. `http://www.esds.ac.uk/aandp/create/identguideline.asp`.

Christian Lovis and Robert Baud. 1999. Electronic patient record: dealing with numbers or with words?

Frances Rock. 2001. Policy and practice in the anonymisation of linguistic data. In *International Journal of Corpus Linguistics*, volume 6. John Benjamins Publishing Company.

John Roddick and Peter Fule. 2003. A system for detecting privacy and ethical sensitivity in data mining results.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden Markov models for information extraction. *Lecture Notes in Computer Science*, 2189:309+.

Kirsten Wahlstrom and John F. Roddick. 2001. On the impact of knowledge discovery and data mining. In John Weckert, editor, *Selected papers from the 2nd Australian Institute of Computer Ethics Conference (AICE2000)*, pages 22–27, Canberra. ACS.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *ACL*, pages 473–480.