An Adaptive, Semi-Structured Language Model Approach to Spam Filtering on a New Corpus

Ben Medlock

Computer Laboratory
University of Cambridge

July 21, 2006

Contents

Motivation

- The development of effective spam filters requires realistic experimental corpora.
- Recent developments are starting to bring this about – TREC 2005, Enron etc. (Cormack and Lynam, Klimt and Yang ...)
- Two spam filtering datasets are better than one: our contribution – *GenSpam*.
- Build classifiers to take advantage of the specific characteristics of the spam filtering task.

Motivation

- The development of effective spam filters requires realistic experimental corpora.
- Recent developments are starting to bring this about – TREC 2005, Enron etc. (Cormack and Lynam, Klimt and Yang . . .)
- Two spam filtering datasets are better than one: our contribution – *GenSpam*.
- Build classifiers to take advantage of the specific characteristics of the spam filtering task.

Motivation

- The development of effective spam filters requires realistic experimental corpora.
- Recent developments are starting to bring this about – TREC 2005, Enron etc. (Cormack and Lynam, Klimt and Yang ...)
- Two spam filtering datasets are better than one: our contribution – *GenSpam*.
- Build classifiers to take advantage of the specific characteristics of the spam filtering task.

Motivation

- The development of effective spam filters requires realistic experimental corpora.
- Recent developments are starting to bring this about – TREC 2005, Enron etc. (Cormack and Lynam, Klimt and Yang ...)
- Two spam filtering datasets are better than one: our contribution – *GenSpam*.
- Build classifiers to take advantage of the specific characteristics of the spam filtering task.

GenSpam Overview

- 9072 genuine, personal email messages sourced from 15 friends and colleagues of the author.
- 32332 spam email messages sourced from sections 10-29 of the *spamarchive* collection, along with a batch collected by the author and colleagues.
- Time period: 2002-2003 (genuine mail more widely time-distributed).

Split

Aim is to facilitate experiments with a large background training set and a smaller, specialised set for adaptation.

- *Training set*: 8018 genuine, 31235 spam
- *Adaptation set*: 300 genuine, 300 spam
- *Test set*: 754 genuine, 797 spam

Adaptation and *Test* sets sourced from two inboxes during Nov 2002 – June 2003

Content and Markup

- Relevant information is extracted from the raw email data and marked up in XML.
- Retained fields include: *Date*, *From*, *To*, *Subject*, *Content-Type* and *Body*.
- Meta-level structure and attachment type preserved but attachment content discarded, except for text and HTML.
- Text embedding preserved.

Anonymisation

- Identity protection is clearly an issue for personal email.
- We use a combination of part-of-speech analysis, pattern matching and manual examination to 'anonymise' the data.
- Only top-level domain (TLD) information is retained in the *From* and *To* fields.
  bwm23@cam.ac.uk → ac.uk
  sam@spamjam.co.uk → co.uk

Anonymisation

- Identity protection is clearly an issue for personal email.
- We use a combination of part-of-speech analysis, pattern matching and manual examination to 'anonymise' the data.
- Only top-level domain (TLD) information is retained in the *From* and *To* fields.
  bwm23@cam.ac.uk → ac.uk
  sam@spamjam.co.uk → co.uk

Anonymisation

- Identity protection is clearly an issue for personal email.
- We use a combination of part-of-speech analysis, pattern matching and manual examination to 'anonymise' the data.
- Only top-level domain (TLD) information is retained in the *From* and *To* fields.
  bwm23@cam.ac.uk $\longrightarrow$ ac.uk
  sam@spamjam.co.uk $\longrightarrow$ co.uk

Anonymisation

The following labels are used as anonymous markers in free text:

- &NAME (proper name)
- &CHAR (individual character)
- &NUM (number)
- &EMAIL (email address)
- &URL (internet URL)

Example

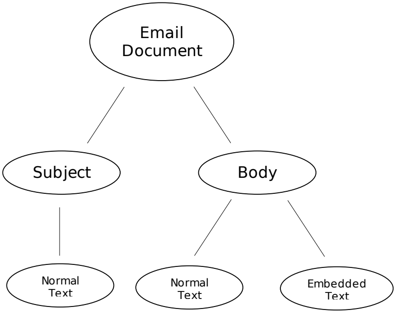
An example of the format of *GenSpam*:

```
<MESSAGE>
<FROM> net </FROM>
<TO> ac.uk </TO>
<SUBJECT>
<TEXT_NORMAL> ^ Re : Hello everybody </TEXT_NORMAL>
</SUBJECT>
<DATE> Tue, 15 Apr 2003 18:40:56 +0100 </DATE>
<CONTENT-TYPE> text/plain; charset="iso-8859-1" </CONTENT-TYPE>
<MESSAGE_BODY>
<TEXT_NORMAL>
^ Dear &NAME ,
^ I am glad to hear you 're safely back in &NAME .
^ All the best
^ &NAME
^ - On &NUM December &NUM : &NUM &NAME ( &EMAIL ) wrote :
...
</TEXT_NORMAL>
</MESSAGE_BODY>
</MESSAGE>
```

A classification model for semi-structured documents (benchmarking *GenSpam*)...

Semi-Structured Document Classification

- A document is viewed as a tree.
- Non-leaf nodes represent meta-level structure
- Leaf nodes represent actual content



Basic Decision Rule

$$Decide(D_i \rightarrow C_j) \text{ where } j = \arg\max_k [P(C_k|D_i)]$$

- Idea: calculate posterior probabilities of individual document nodes and combine using the tree structure.
- Posterior for entire document is posterior for top-level node.

Non-leaf Node Estimation

Non-leaf node posterior is estimated as a weighted interpolation of its subnode posteriors.

$$P(C_j|D_i) = \sum_{n=1}^{N} \lambda_n \left[P(C_j^n|D_i^n) \right]$$

Leaf Node Estimation

Leaf node posterior estimated in standard generative fashion:

$$P(C_j^n|D_i^n) = \frac{P(C_j^n) \cdot P(D_i^n|C_j^n)}{P(D_i^n)}$$

- $P(C_j^n)$ is the class prior
- $P(D_i^n)$ is the document prior and constant with respect to class, though important for normalisation.
  It is calculated by $\sum_{k=1}^{|\mathbf{C}|} P(C_k^n) \cdot P(D_i^n|C_k^n)$
- $P(D_i^n|C_j^n)$ is the language model probability of the field.

LM Construction

We use *n*-gram language models:

$$P_N(t_1, \ldots, t_K) = \prod_{i=1}^{K} P(t_i | t_{i-N+1}, \ldots, t_{i-1})$$

Sparsity handled by Katz back-off:

$$P(t_j|t_i) = \begin{cases} d(f(t_i, t_j))\frac{f(t_i,t_j)}{f(t_i)} & \text{if } f(t_i, t_j) > C \\ \alpha(t_i)P(t_J) & otherwise \end{cases}$$

where f is the frequency-count function
d is the discounting function
α is the back-off weight
C is the *n*-gram cutoff point

Discounting

We use a simple discounting function – *confidence discounting*:

$$d(r) = \frac{r}{R}\omega$$

where R is the number of distinct n-gram frequencies.
ω represents a ceiling on discount mass ($\sim$1).

Idea: confidence in an n-gram estimate is based on the absolute frequency of that n-gram in the training data. Higher confidence results in less discounted probability mass.

Unseen Event Modelling

A small probability must be assigned to events that remain unobserved at the end of the back-off chain. We can use this to model discrepancies between the likelihood of observing previously unseen events in spam/genuine mail.

Adaptivity

Spam filters need to be *adaptive*.

Two forms of adaptivity:

- Adapt to changes in the nature of email over time.
- Fit individual user instances while taking account of evidence of accumulated common knowledge (client-server analogy).

One potential solution is to employ two sets of language models:

- a larger, static background set.
- a smaller, user-specific set to be regularly re-trained with new evidence.

Evidence from both these sets of models would then be combined.

Adaptivity

Spam filters need to be *adaptive*.

Two forms of adaptivity:

- Adapt to changes in the nature of email over time.
- Fit individual user instances while taking account of evidence of accumulated common knowledge (client-server analogy).

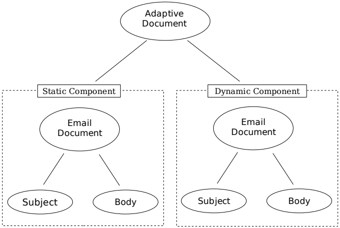
One potential solution is to employ two sets of language models:

- a larger, static background set.
- a smaller, user-specific set to be regularly re-trained with new evidence.

Evidence from both these sets of models would then be combined.

Adaptive Decision Rule

$$Decide(D_i \rightarrow C_j) \quad \ldots \quad j = \arg\max_k [\lambda_s P_s(C_k|D_i) + \lambda_d P_d(C_k|D_i)]$$



Classifiers

For benchmarking the *GenSpam* corpus we use:

- Multinomial Naïve Bayes (MNB)
- Support Vector Machines (SVM) – Vapnik 95, Joachims 98
- Bayesian Logistic Regression (BLR) – Genkin et. al 05
- Interpolated Language Model (ILM) – our classifier

SVM and BLR both state-of-the-art on text categorization.

Hyperparameter Tuning

ILM:

- Interpolation weights
- Unseen event estimates
- *n*-gram cutoff (for higher-order *n*-grams)

SVM:

- Kernel type (linear)
- Regularization parameter

BLR:

- Prior distribution type (Gaussian)
- Prior variance

Asymmetric Classification

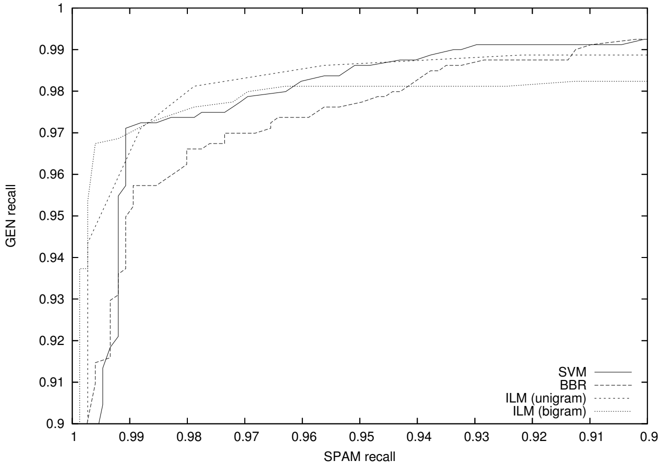
- Spam filtering requires near-perfect recall of genuine mail.
- Evaluate classifiers under genuine recall threshold constraint: recall$\geq$0.995 ($\leq$ 1 in 200 genuine messages missed)
- MNB, SVM, BLR – bias decision boundary
- ILM – bias language models through unseen estimate modification

Results

Training Data	Classifier	GEN recall	SPAM recall	accuracy
Training	MNB	0.9960	0.1556	0.5642
	SVM	0.9960	0.7064	0.8472
	BLR	0.9960	0.8105	0.9007
	ILM Unigram	0.9960	0.7340	0.8614
	ILM Bigram	0.9960	0.8331	**0.9123**
Adaptation	MNB	0.9960	0.4090	0.6944
	SVM	0.9960	0.9147	0.9491
	BLR	0.9960	0.9097	**0.9542**
	ILM Unigram	0.9960	0.8269	0.9091
	ILM Bigram	0.9960	0.8934	0.9433
Combined	MNB	0.9960	0.4103	0.6950
	SVM	0.9960	0.8808	0.9368
	BLR	0.9960	0.9021	0.9478
	ILM Unigram	0.9960	0.9573	0.9761
	ILM Bigram	0.9960	0.9674	**0.9813**

Table: Asymmetric results (best results for each dataset in bold)

ROC Curves



1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
GEN recall
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
SPAM recall
SVM
BBR
ILM (unigram)
ILM (bigram)

Discussion

ILM Advantages:

- Efficient linear ML training of n-gram LMs.
- Efficient combination of distinct distributional evidence.
- Native probabilistic output.
- Effective bias control.

ILM Disadvantages:

- Potentially expensive hyperparameter estimation.
- Sensitivity to domain character adaptation – a relevant issue for spam filtering.

Conclusions

Conclusions:

- Spam filtering research needs realistic corpora – *GenSpam*
- ILM classification model has some useful properties for spam filtering.

Future work:

- Update spam component of *GenSpam*.
- Hyperparameter estimation techniques for ILM.
- Discriminative techniques for semi-structured spam filtering.
- Combine separate distributional evidence in SVM, BLR etc.