Feature-Based Forensic Text Comparison Using a Poisson Model for Likelihood Ratio Estimation

MICHAEL CARNE[1] & SHUNICHI ISHIHARA[1,2]

1Forensics Stream of Speech and Language Lab

2Linguistics Program, Australian National University

michael.carne@anu.edu.au shunichi.ishihara@anu.edu.au

Introduction

There are two main methods for estimating a forensic likelihood ratio (LR) quantifying the strength of forensic evidence: score- and feature-based. In score-based methods, the evidence consists of scores, $\Delta(x, y)$, which are often measured as the distance between the suspect and offender samples. Distance measures (e.g. Burrows' Delta, Cosine distance) are a standard tool in authorship attribution studies (Burrows, 2002; Argamon, 2008) , and a natural first step in the estimation of an LR in forensic text comparison (FTC). However, textual data often violates the statistical assumptions underlying distance measures. Frequently-occurring words, such as 'a' (Figure 1a), tend to be normally distributed. However, the distribution starts skewing positively for less-frequently occurring words, such as 'not' (Figure 1b) and 'they' (Figure 1c).

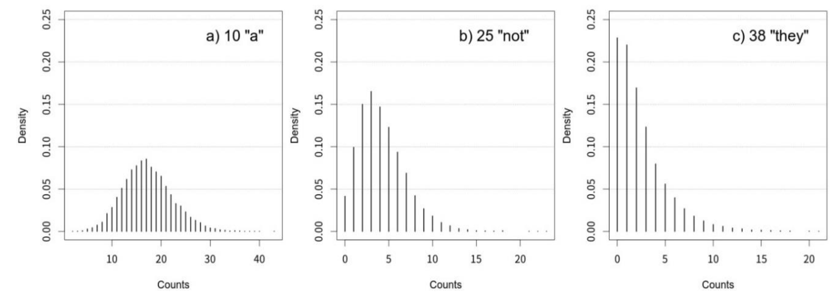

Figure 1: Histograms showing the distributional patterns of the counts of three words from the database; 'a', 'not' and 'they' for Panel a), b) and c), respectively. They are the 10th, 25th and 38th most frequently-occurring words in the database used for the current study.

Further, score-based distance models only assess the similarity, not the typicality, of the objects (i.e. documents) under comparison. A Poisson model is theoretically more appropriate than distance-based measures for authorship attribution, but it has never been tested with linguistic text evidence within the LR framework. In this study, a score-based method using the Cosine distance is compared with a feature-based method built on a Poisson model with texts collected from 2,157 authors.

Score and feature-based LR estimation

The Likelihood Ratio framework is a means of quantifying the weight of evidence for a variety of forensic evidence e.g. DNA (Evett and Weir, 1998), voice (Morrison et al., 2018; Rose, 2002), fingerprints (Neumann et al., 2007), MDMA tablets (Bolck et al., 2009). A likelihood ratio quantifies the strength of evidence with respect to two completing hypotheses: (H_p) specifies the prosecution (or the same–author), hypothesis (H_d) the defence (or the different-author) and these are expressed as a ratio of conditional probabilities.

$$LR = \frac{f(x, y|H_p)}{f(x, y|H_d)}$$

Where x and y are feature values obtained from the known-source and questioned-source respectively. The relative strength of the evidence with respect to the competing hypotheses is reflected in the magnitude of the LR: the more the LR deviates from unity (LR = 1), the greater support for either the H_p (LR > 1) or the H_d (LR < 1).

Score-based methods project the complex, multivariate feature vector into a univariate score space (Morrison and Enzinger, 2018: 47) and estimate the probabilities densities from those scores.

$$LR = \frac{f(x, y|H_p)}{f(x, y|H_d)} = \frac{f(\Delta(x, y)|H_p)}{f(\Delta(x, y)|H_d)}$$

Where $\Delta(x, y)$ the distances between the suspect and offender samples. The robustness and ease of implementation for various types of forensic evidence have been reported as benefits of score-based methods (Bolck et al., 2015).

Feature-based models estimate probabilities directly from the feature values. This has the potential to prevent information loss but comes at the cost of added model complexity and reduced computational efficiency. Feature-based methods allow the typicality, not only the similarity, of forensic data to be assessed. In this study a Poisson distribution was used to construct the LR model.

$$LR = \frac{f(x, y|H_p)}{f(x, y|H_d)} = \frac{e^{-\lambda_x}\frac{\lambda_x^y}{y!}}{e^{-\lambda_B}\frac{\lambda_B^y}{y!}}$$

Where λ_x is the count of a given feature word (e.g. w_1^x) appearing in the suspect document, y is the count of feature word (e.g. w_1^y) appearing in the offender document, and the λ_B is the overall mean λ of the background database

Data

- Data was obtained Amazon Product Data Authorship Verification Corpus (Halvani et al., 2017)
- From the corpus, authors (= reviewers) who contributed more than six reviews longer than 700 words, were selected as the database for simulating offender vs. suspect comparisons, resulting in 2,157 reviewers
- Data was partitioned into three separate databases, each containing 719 authors:

Test database. Used for assessing the FTC system performance by simulating same-author (SA) and different-author (DA) comparisons. 719 same-authour (SA) comparisons and 516,242 (= $_{719}C_2 \times 2$) different-authour (DA) comparison were possible.

Development database. In LR-based FTC a development database is used fuse and calibrate the raw LRs. Score-based LRs were found to be already well calibrated, so calibration/fusion weights were only derived for the feature-based method.

Background database:

- score-based method: used to train the score-to-LR conversion model.
- feature-based: to assess the typicality of the documents under comparison.

Tokenisation and Bag of Words Model

- The `tokens()` function from the `quanteda` library (Benoit et al., 2018) in R (R Core Team, 2017) was used to tokenise document texts.
- All characters were converted to lower case without punctuation marks being removed; punctuation marks were treated as single word tokens.
- The 400 most frequent occurring words in the entire dataset were selected as components for a bag-of-words model.
- The documents (*x*, *y*) under comparison were modelled as the vectors ($x = \{w_1^x, w_2^x \cdots w_N^x\}$ and $y = \{w_1^y, w_2^y \cdots w_N^y\}$) with the word counts ($w_i^j$, $i \in \{1 \cdots N\}$, $j \in \{x, y\}$).
- The size (*N*) of the bag-of-words vector was incremented by 5 from *N* = 5 to *N* = 20, and then by 20 until *N* = 400. The 400 most frequent words are sorted according to their frequencies in a descending order. *N* = 400 was chosen as the cap of the experiments because the experimental results showed the performance ceiling before *N* = 400

Evaluation of performance

- Performance was assessed using the log-likelihood-ratio cost (C_{llr}) (Brümmer & du Preez, 2006).
- C_{llr} is a gradient measure of the validity (accuracy) of the system.

$$C_{llr} = \frac{1}{2}\left(\left[\frac{1}{N_{SA}}\sum_i^{N_{SA}} log_2\left(1 + \frac{1}{LR_i}\right)\right] + \left[\frac{1}{N_{DA}}\sum_j^{N_{DA}} log_2(1 + LR_j)\right]\right)$$

- Where, N_{SA} and N_{DA} are the number of SA and DA comparisons, and LR_i and LR_j are the LRs for the SA and DA comparisons, respectively.
- Optimum performance (accuracy) is achieved when a C_{llr} = 0 and degrades as C_{llr} approaches and exceeds 1.
- C_{llr} can be decomposed into additional performance metrics: $C_{llr} = C_{llr}^{min}$ (discrimination loss) + C_{llr}^{cal} (calibration loss)

Results: Accuracy (C_{llr})

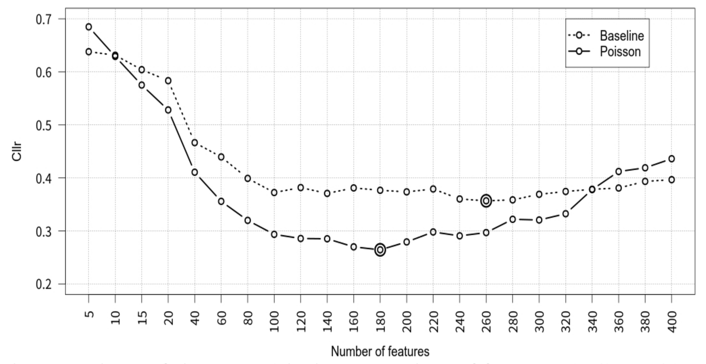

Figure 2: The C_{llr} values of the LRs with the *N* number of features indicated in the Y-axis are plotted separately for the Baseline and the Poisson models. The features are sorted according to the frequencies of the words. The large circles indicate the best C_{llr} values for the models

- On average the feature-based Poisson model yields better accuracy (on average lower C_{llr} values) relative to the score-based model
- Optimum performance is achieved with 180 for the Poisson LR model (C_{llr} = 0.26) and 260 for the score-based cosine LR model (C_{llr} = 0.36)
- The performance of the score-based model is relative stable as the number of features included increases, while it deteriorates for the feature-based model when > 180 features are included.

Results: Discrimination (C_{llr}^{min}) and Calibration (C_{llr}^{cal}).

To investigate the reasons for the deterioration in the performance of the feature-based LR models we examined other performance characteristics: discrimination (C_{llr}^{min}) and calibration loss (C_{llr}^{cal}).

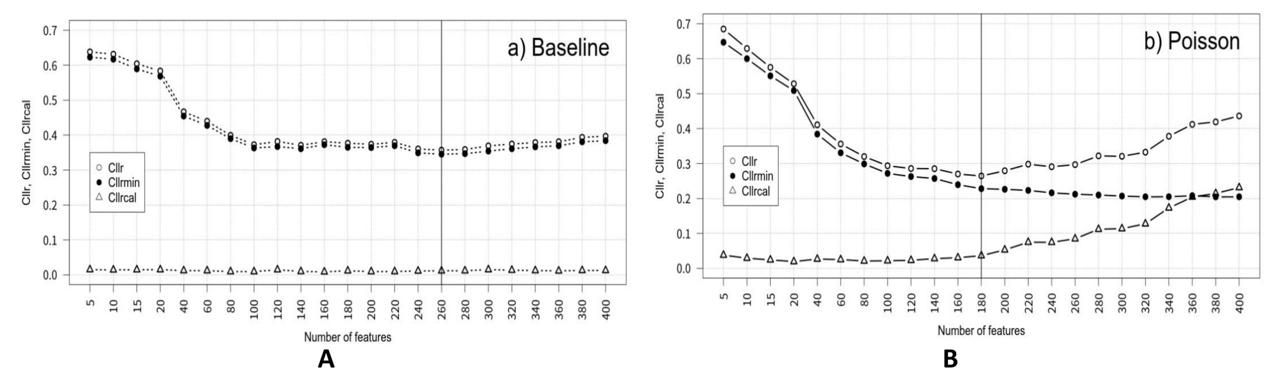

Figure 3: The C_{llr}, C_{llr}^{min} *and* C_{llr}^{cal} values of the LRs, with the *N* number of features indicated in the *y*-axis, are plotted separately for the Baseline (Panel a) and the Poisson (Panel b) models. The features are sorted according to word frequency. The vertical solid line indicates where the best C_{llr} value was obtained

- Discrimination loss (filled circles) in the feature-based model decreases as the number of features increases, and is lower relative to the score-based model.
- Calibration (triangles): Worsens after 180 features for the Poisson model (Panel B), whereas the baseline shows good calibration, which remains stable as the number of features increases (Panel A)
- The deterioration in the C_{llr} value for the Poisson model (filled circles, Panel B) after 180 features coincides with worsening calibration (triangles). It is likely therefore that reduced accuracy is a function of poor calibration in larger feature spaces, rather than discrimination performance which is seen to improve.

Feature selection using C_{llr}^{min}

It was observed that the performance of a given feature (i.e. word) did not always correspond to the frequency of its occurrence. This is illustrated in Table 1, which lists the ten most frequently occurring words and the ten words with the highest discriminability (i.e. C_{llr}^{min}).

By word frequency		By C_{llr}^{min}	
Frequency	Words	Frequency	Words
1	'.'	3	','
2	'the'	1	'.'
3	','	41	'it's'
4	'and'	35	'!'
5	'i'	31	'-'
6	'a'	28	'('
7	'to'	27	')'
8	'it'	5	'i'
9	'of'	84	'i'm'
10	'is'	4	'and'
Table 1: Ten most-frequent (left) and lowest-C_{llr}^{min} (right) words			

Table 1: Ten most-frequent (left) and lowest-C_{llr}^{min} (right) words

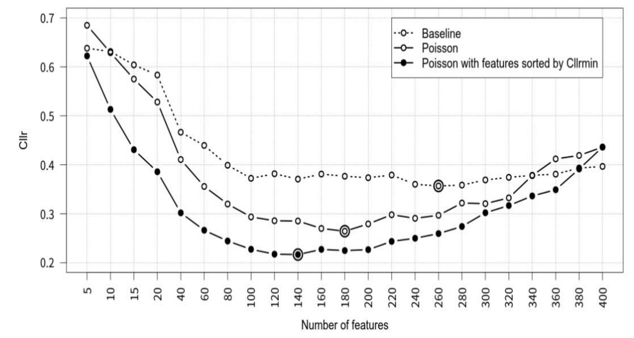

Figure 4: The C_{llr} values of the (fused) LRs with the *N* number of C_{llr}^{min}-sorted features indicated in the *y*-axis are plotted together with the result presented in Figure 2 for comparison. The large circles indicate the best C_{llr} values for the models.

- In a second set of experiments, words were first sorted according to their discrimination loss (C_{llr}^{min} values), LRs were then fused/calibrated based on this basis.
- Figure 4. shows feature selection based C_{llr}^{min} values contributes to an overall improvement in performance for the Poisson model.
- The optimum C_{llr} for the Poisson model is lower (0.217) with less features (*N* = 140) (solid lines, filled circles) compared to the results with the unsorted values (unfilled circles).
- The superior performance of the Poisson based model can also be appreciated visually in the Tippet plots in Figure 5, which show the cumulative proportion of LRs from the SA comparisons (SALRs), which are plotted rising from the left, as well as of the LRs of the DA comparisons (DALRs), plotted rising from the right. For all Tippet plots, the cumulative proportion of trails is plotted on the *y*-axis against the log_{10} LRs on the *x*-axis.

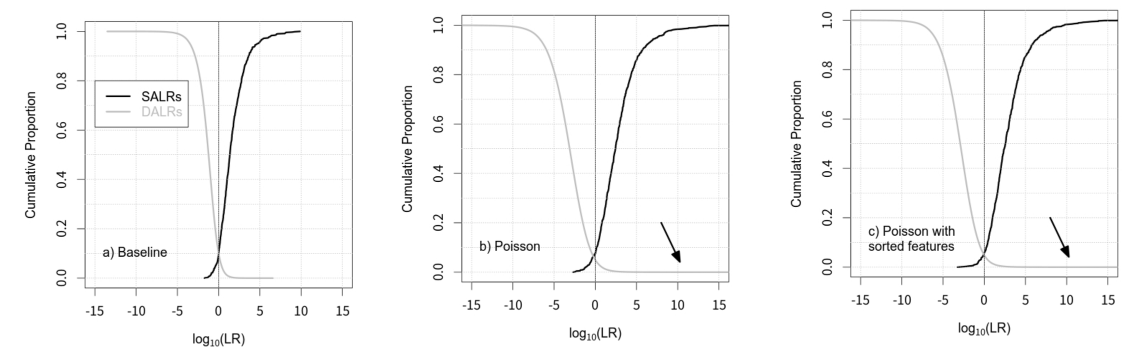

Figure 5: Tippett plots showing the magnitude of the derived LRs. Panel a) = Best-performing Baseline model; Panel b) = Best-performing original Poisson model; Panel c) = Best-performing Poisson model with sorted features according to their C_{llr}^{min} values. Note that some LRs extend beyond ±15 of the *y*-axis. Arrows indicate very strong contrary-to-fact DALRs.

- Although the overall magnitude of LRs is greater for the Poisson models (Panels B, C), relative to the Baseline model (Panel B), they evince strong contrary-to-fact DALRs (which are indicated by arrows in Figure 5).

Conclusions & Limitations

- The feature-based FTC system outperformed the score-based FTC system with Cosine distance.
- It was demonstrated that the performance of the feature-based system can be further improved by selecting the sets of LRs to be fused according to their C_{llr}^{min} values.
- Discrimination loss in the feature-based FTC system reduces as the number of features increases, but becomes less well calibrated with a larger feature space.
- While a simple one-level Poisson LR model shows good performance, alternatives such as the negative Binomial and the zero-inflated Poisson may be better motivated (Jansche, 2003; Pawitan, 2001) and two-level Poisson model might also considered (Aitken and Gold, 2013; Bolck and Stamouli, 2017).
- Only a limited set of features used (word counts), a richer feature set could be used in future work.

References

Aitken, C. G. G. and Gold, E. (2013) Evidence evaluation for discrete data. *Forensic Science International 230*(1-3): 147-155. https://dx.doi.org/10.1016/j.forsciint.2013.02.042

Argamon, S. (2008) Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing 23*(2): 131-147. https://dx.doi.org/10.1093/llc/fqn003

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. and Matsuo, A. (2018) quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software 3*(30): 774-776. https://doi.org/10.21105/joss.00774

Bolck, A. and Stamouli, A. (2017) Likelihood ratios for categorical evidence; Comparison of LR models applied to gunshot residue data. *Law, Probability and Risk 16*(2-3): 71-90. https://dx.doi.org/10.1093/lpr/mgx005

Bolck, A., Ni, H. F. and Lopatka, M. (2015) Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: Applied to forensic MDMA comparison. *Law, Probability and Risk 14*(3): 243-266. https://dx.doi.org/10.1093/lpr/mgv009

Bolck, A., Weyermann, C., Dujourdy, L., Esseiva, P. and van den Berg, J. (2009) Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International 191*(1-3): 42-51. https://dx.doi.org/10.1016/j.forsciint.2009.06.006

Brümmer, N. and du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech and Language 20*(2-3): 230-275. https://dx.doi.org/10.1016/j.csl.2005.08.001

Evett, I. W. and Weir, B. S. (1998) *Interpreting DNA Evidence : Statistical Genetics for Forensic Scientists*. Sunderland, Mass.: Sinauer Associates

Halvani, O., Winter, C. and Graner, L. (2017). Authorship verification based on compression-models. *arXiv preprint arXiv:1706.00516*. Retrieved on 25 June 2020 from http://arxiv.org/abs/1706.00516

Jansche, M. (2003) Parametric models of linguistic count data. *Proceedings of Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*: 288-295.

Morrison, G. S. and Enzinger, E. (2018) Score based procedures for the calculation of forensic likelihood ratios - Scores should take account of both similarity and typicality. *Science & Justice 58*(1): 47-58. https://dx.doi.org/10.1016/j.scijus.2017.06.005

Morrison, G. S., Enzinger, E. and Zhang, C. (2018) Forensic speech science. In I. Freckelton and H. Selby (eds.), *Expert Evidence*. Sydney, Australia: Thomson Reuters.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A. and Bromage-Griffiths, A. (2007) Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences 52*(1): 54-64. https://dx.doi.org/10.1111/j.1556-4029.2006.00327.x

Pawitan, Y. (2001) *In All Likelihood : Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Rose, P. (2002) *Forensic Speaker Identification*. London: Taylor & Francis.