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

MICHAEL CARNE, & SHUNICHI SHIMARA

Forensics Streams of Speech and Language Lab

Institute of Linguistics, University of Aizu

michael.carne@u-auizu.ac.jp
shunichi.shimara@u-auizu.ac.jp

Introduction

There are now two main methods for estimating a forensic likelihood ratio (LR) quantifying the strength of evidence: forensic- and feature-based approaches. In forensic-based approaches, models are used to estimate the evidence (e.g., the suspect and offender samples, distance measures), and the models are compared (e.g., by assuming a model for LR estimation). However, vertical data often violates the statistical assumptions underlying distance measures. Frequently occurring words, such as “r-rated” (figure 2), tend to be normally distributed, but the distribution starts deviating for less frequently occurring words, such as “not” (figure 1) and “they” (figure 1). This behavior is further described for score-based LR estimation models.

Score and feature-based LR estimation

The standardized forensic-likelihood-ratio (LR) estimation framework is a means of quantifying the weight of evidence for a variety of forensic evidence e.g. DNA (Scott & Wake, 1996), voice (Morrison et al., 2010; Rose, 2008) fingerprinting (Warne et al., 2007), MDAA tables (Bolck, 2007). A forensic-based ratio quantifies the strength of evidence with respect to two competing hypotheses (H₀: the claim; H₁: the alternative) and these are expressed as a ratio of conditional probabilities:

\[ LR = \frac{P(D|H₀)}{P(D|H₁)} \]

Where A and B 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 LR from different A values is compared (figure 1). The greater support for either the H₀ or H₁ the greater LR.

Score-based methods project the complex, multivariate feature vector into a univariate score vector (Morrison & Ekenizov, 2018). This approach avoids complexity but can decrease computational efficiency. Feature-based methods allow the typology, not only the certainty of forensic data to be assessed. In order to study a Poisson-based model was used to construct the LR model:

\[ LR = \exp \left[ \frac{1}{\lambda} \sum_i \log (i+1) \right] \]

Where \( \lambda \) is the number of unigram features selected (e.g., looking at the unigram frequencies in the document), the \( \lambda \) is the overall mean of the background database.

Data

- Cards were obtained from the Amazon Product Database (Accessed 10/10/18)
- Total number of cards was more than six times 62000 cards were selected, the database is for browsing the unigram frequencies in the document, the \( \lambda \) is the overall mean of the background database.

Feature-based models: feature probabilities from the database. This has the potential to present inconsistent bias but allows the ease of added model complexity and reduced computational efficiency. Feature-based models allow typology, not only the certainty of forensic man data to be assessed. In order to study a Poisson-based model was used to construct the LR model:

\[ LR = \exp \left[ \frac{1}{\lambda} \sum_i \log (i+1) \right] \]

Where \( \lambda \) is the count of given feature word (e.g., “the”) appearing in the suspect document, \( \lambda \) is the count of feature word (e.g., “the” appearing in the offender document, and the \( \lambda \) is the overall mean of the background database.

Development database: in a feature-based method, a feature set is ranked based on the LR. Score-based features were found to be already well-calibrated, so calibration/fusion weights were only derived for the feature-based method.

Results

To investigate the reasons for the deterioration in the performance of the feature-based LR models we examined other performance characteristics: discrimination (discrimination) and calibration loss (calibration).

Results Accuracy (LR)

- On average the feature-based Poisson model yields better accuracy (on average lower LLRs) than the score-based model.
- 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 more than 180 features are included.

Results Discrimination (discrimination) and Calibration (calibration)

- The \( \lambda \) feature-based LR models decrease discrimination as the number of features increases. The lower \( \lambda \) for feature-based LR models is relative stable as the number of features included increases, while it deteriorates for the feature-based model when more than 180 features are included.

Conclusions

The feature-based FTC system outperformed the score-based FTC system with Cosine distance.

In conclusion, the performance of the feature-based system can be further improved by selecting the set of all is fixed according to the \( \lambda \) and LLRs.

Discrimination loss is the base of feature predictors; the reduction of the base increases, but becomes less well-calibrated with a larger feature space. While a simple one-plot Poisson LR model shows good performance, alternatives such as the negative binomial and the zero-inflated Poisson model may be better options (Evett and Weir, 2001; Law, 2007; and two-level Poisson models may also be considered (Mohan and Gold, 2014; Blank and Graefes, 2017).

Only a limited set of features used (word counts), a richer feature set could be used in future work.

Discussion

- It was observed that the performance of a given feature (e.g., word) did not always correspond to the frequency of its occurrence. It is illustrated in Table 2, which lists the ten most frequent occurring words and the two words with the highest discriminability (e.g., \( \lambda \)).

- In a second set of experiments, words were first sorted according to their discrimination LLR (SALRs) values, then were first sorted/Calibrated based on this basis.

- The optimal \( \lambda \) for the Poisson model is lower (0.11) with less features \( N = 1016 \) (discarded, filled circle) compared to the results with the uncalibrated values (full circle).

- The performance of the Poisson based model can be also appreciated visually in the Togit plots in figure 5, which show the cumulative proportion of LLRs from the A and B samples, which are plottings arising from the left, as well as the LLRs of the A and B samples (discarded), plotted long from the right. For all Togit plots, the cumulative proportion of LLRs at the pavement increases the log_1, k, on at the e.

References