Forensic Science International 278 (2017) 184–197

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Strength of linguistic text evidence: A fused forensic text comparison system

Shunichi Ishihara*

Department of Linguistics, The Australian National University, Canberra, Australia

ARTICLE INFO

Article history:
Received 27 March 2017
Received in revised form 5 June 2017
Accepted 30 June 2017
Available online 8 July 2017

Keywords:
Forensic text comparison
Likelihood ratio
Logistic-regression fusion
Multivariate kernel density
N-grams
Authorship attribution features

ABSTRACT

Compared to other forensic comparative sciences, studies of the efficacy of the likelihood ratio (LR) framework in forensic authorship analysis are lagging. An experiment is described concerning the estimation of strength of linguistic text evidence within that framework. The LRs were estimated by trialling three different procedures: one is based on the multivariate kernel density (MVKD) formula, with each group of messages being modelled as a vector of authorship attribution features; the other two involve *N*-grams based on word tokens and characters, respectively. The LRs that were separately estimated from the three different procedures are logistic-regression-fused to obtain a single LR for each author comparison. This study used predatory chatlog messages sampled from 115 authors. To see how the number of word tokens affects the performance of a forensic text comparison (FTC) system, token numbers used for modelling each group of messages were progressively increased: 500, 1000, 1500 and 2500 tokens. The performance of the FTC system is assessed using the log-likelihood-ratio cost (C_{llr}), which is a gradient metric for the quality of LRs, and the strength of the derived LRs is charted as Tippett plots. It is demonstrated in this study that (i) out of the three procedures, the MVKD procedure with authorship attribution features performed best in terms of C_{llr}, and that (ii) the fused system outperformed all three of the single procedures. When the token length is 1500, for example, the fused system achieved a C_{llr} value of 0.15. Some unrealistically strong LRs were observed in the results. Reasons for these are discussed, and a possible solution to the problem, namely the empirical lower and upper bound LR (ELUB) method is trialled and applied to the LRs of the best-achieving fusion system.

1. Introduction

The history of authorship attribution study is long. Mendenhall's [1] study on Shakespeare's plays is often quoted as the first authorship attribution study based on a statistical/computational method. It was followed by many influential studies in the first half of the 20th century [2–5]. Since the end of the 20th century, due to the change in communication medium, the focus of authorship attribution has started shifting from literary texts to electronically-generated texts (e.g. emails, chatlogs, SMS), with some studies focusing on the domain of forensics [6–17].

However, forensic authorship attribution has considerably fallen behind in comparison to other forensic comparative sciences in that the above forensic authorship attribution studies were not conducted in the likelihood ratio (LR) framework, which is increasingly held to be the logically and legally correct framework of evaluating forensic evidence (cf. [55,18]). In many branches of forensic sciences, including fingerprint [19], handwriting [20], voice [21], DNA [22], glass fragments [23], earmarks [24] and footwear marks [25] analysis, the LR framework has been or has started being accepted as the standard framework for the evaluation of forensic evidence. The spotlight on the LR framework is, needless to say, largely attributable to the success of DNA profiling [26,27], as well as to some rulings (*Daubert v. Merrell Dow Pharmaceuticals Inc, 1993; Kuhmo Tire Co. v. Carmichael, 1999*) and reports (*Strengthening Forensic Science in the United States: A Path Forward* (2009)) [28] regarding the rules of evidence in the United States. As a matter of fact, the use of the LR framework has been advocated for quite some time in the main textbooks on the evaluation of forensic evidence [29] and by forensic statisticians [30,31].

In this study, therefore, the LR framework is implemented for authorship attribution. First of all, three different procedures are trialled to estimate LRs for predatory chatlog messages—one based

* Corresponding author.Present address: Department of Linguistics, School of Culture, History and Language, College of Asia and the Pacific, The Australian National University, Building #110, ACTON, Canberra, ACT 2601, Australia.
E-mail address: shunichi.ishihara@anu.edu.au (S. Ishihara).

http://dx.doi.org/10.1016/j.forsciint.2017.06.040
0379-0738/

on authorship attribution features with the multivariate kernel density (MVKD) LR formula (the MVKD procedure); one with word token-based *N*-grams (the token *N*-grams procedure) and one with character-based *N*-grams (the character *N*-grams procedure). In the MVKD procedure, each message group (e.g. a set of messages written by a suspect or an offender) is modelled as a vector of authorship attribution features, such as the vocabulary richness feature, the average token number per message line, upper case character ratio, etc. (refer to Section 3.3.1 for further details on authorship attribution features). In the token and character *N*-grams procedures, each message group is modelled by token- and character-based *N*-grams, respectively. The performances of the three different procedures are compared.

In a second step, the LRs that were separately derived by the three different procedures are fused into a single LR for each comparison, representing the combined evidence. This allows us to investigate the extent to which fusion improves (or deteriorates) the performance of the forensic text comparison (FTC) system. The current study employs logistic-regression fusion [32] as it is a robust technique and has been applied to some LR-based forensic comparison systems (some examples are given in Ref. [33]). The performance of each FTC system is assessed by means of the log likelihood ratio cost (C_{llr}) [32,34], which is a gradient metric for assessing the quality of LRs. The strength of the derived LRs is visually displayed as Tippett plots. Detailed explanations of logistic-regression fusion, C_{llr} and Tippett plots are given in Sections 3.4, 3.5 and 4, respectively.

2. Likelihood ratio and Bayesian theorem

Many forensic scientists and statisticians [30,31,29] explicitly state that the role of the forensic scientist is to estimate the strength of evidence, which can be quantified by the LR. The LR is the ratio of the probability that the evidence (E) would occur if one hypothesis (e.g. the prosecution hypothesis—H_p) is true and the probability that the same evidence would occur if the alternative hypothesis (e.g. the defence hypothesis—H_d) is true [29]. Thus, the LR can be expressed as in Eq. (1).

$$LR = \frac{p(E|H_p)}{p(E|H_d)} \quad (1)$$

Consider a typical FTC scenario where the forensic scientist is required to compare a set of messages written by the offender and another set of messages written by the suspect, and the following hypotheses are of interest:

Hp: the two sets of messages were written by the same author.

H_d: the two sets of messages were written by different authors.

In FTC, the evidence (E) consists of the measured properties of the messages written by the suspect and the messages written by the offender. The numerator of Eq. (1) determines the probability of the evidence assuming the same author hypothesis (H_p). Likewise, the denominator of Eq. (1) determines the probability of the same evidence assuming the different author hypothesis (H_d). The LR is the ratio of those two probabilities under the competing hypotheses. If the evidence is more likely to be observed under the same author hypothesis than under the different author hypothesis, then the LR will be greater than one. On the contrary, if the evidence is more likely to be observed under the different author hypothesis than under the same author hypothesis, then the LR will be smaller than one. That is to say, the relative strength of the given evidence with respect to the competing hypotheses (H_p vs. H_d) is reflected in the magnitude of the LR. The more the LR deviates from one, the greater support there is considered to be for either the prosecution hypothesis or the defence hypothesis.

In an LR estimation, the equation given in (1) is interpreted in terms of similarity (numerator) and typicality (denominator). The numerator quantifies the degree of similarity between the two samples (e.g. messages) in the comparison, and the denominator quantifies the significance of such similarity. Even if the samples of the offender and the suspect are found to be very similar, the similarity is less significant if the measured properties of the samples are very typical against the relevant population, as there would be many other individuals in the population who could present the same measured properties. Therefore, at least three different sets of data are required for estimating LRs: offender samples, suspect samples and samples from the population relevant to the case (background reference data).

An LR is a relative strength of evidence. It *indicates* whether the evidence supports the prosecution or defence hypothesis. To quantify the amount of support or obtain a probability score for the offender and suspect being the same person or otherwise, given the evidence (i.e. the probability of the hypotheses in light of the evidence; namely posterior odds or strength of hypothesis), the LR needs to be combined with the prior odds of the hypotheses via Bayes' theorem. The prior odds is the trier-of-fact's belief in relative favour of the two competing hypotheses, which is a result of initial assumptions and changes in belief after the presentation of all the relevant evidence. Such trier-of-fact's belief is not knowledgeable to the forensic scientist; thus the latter cannot logically calculate the posterior odds [35]. In addition, they must not calculate the posterior odds for legal reasons: referring to the posterior odds is equivalent to referring to the suspect as being guilty or not guilty, which is not the role of the forensic expert but of the fact finder; the forensic expert should not be usurping the role of the trier-of-fact [31,p. 4], [36].

3. Experiments

3.1. Database

Real pieces of chatlog communication between later-sentenced paedophiles and undercover police officers in the US, drawn from an archive of chatlog messages (http://pjfi.org/) were used for the research reported on in this paper. However, as the archive had not been designed as a database for authorship analysis studies, the messages written by each author had to be manually checked and transformed to a computer-readable format prior to the commencement of the current study. In total, the messages written by 383 authors between 2007 and 2011 were processed as described. Out of the 383 authors, only those who enabled us to create two groups of messages that do not chronologically overlap and that each consist of 2500 tokens were further selected to meet the experimental specifications detailed later in this subsection. This resulted in 115 authors and their messages being selected for the FTC experiments that were carried out.

The 115 authors were separated into three mutually exclusive sub-databases: the test database (39 authors); the background database (38 authors); and the development database (38 authors). The test database was used to simulate the various offender-suspect comparisons by means of which the performance of the FTC system was assessed. The background database was used as a reference to determine typicality for calculating LRs. The development database was used to calculate weights for calibrating the derived LRs of the SA and DA comparisons generated from the test database. As the test database contained material by 39 authors, 39 same author (SA) and 1482 independent different author (DA) comparisons (=741 author pairs ($=_{39}C_2$) $\times$ 2 comparisons for each author pair) were possible. Given their identical origins, the LRs estimated for the 39 SA comparisons were anticipated to be greater than LR = 1, to the extent that the