D55 Comparing Statistical and Machine-Learning Techniques in Author Identification and Verification

Carole E. Chaski, PhD, ALIAS Technology, LLC, Institute for Linguistic Evidence, 25100 Trinity Drive, Georgetown, DE 19947; Gary Holness, PhD, Delaware State University, Computer Science Dept, 1200 N Dupont Highway, Dover, DE 19901; and Michael J. Harris, MA, University of California Santa Barbara, Dept of Spanish and Portuguese, Phelps Hall 4206, Santa Barbara, CA 93106*

After attending this presentation, attendees will be acquainted with validation testing in forensic linguistic evidence, specifically author identification and current results of litigation-independent research. Attendees will be able to assess any author-identification methods they encounter in light of current research results and standards.

This presentation will impact the forensic science community by providing an example of validation testing that tests both statistical and machine-learning classification methods and by providing current research on the reliability of computational forensic linguistic author identification and verification. These problems are becoming more important due to the internet.

This presentation focuses on validation testing of methods for determining the author of a document in the forensic setting using a traditional forensic methodology and a novel biometric methodology. As Chaski explains, there are three current approaches to forensic author identification: forensic stylistics, computational stylometry, and forensic computational linguistics.[1] Stylistics has its roots in handwriting identification; stylometry in computer science and digital humanities; and forensic computational linguistics in linguistic theory and computational linguistics. Among the many ways that these three approaches differ, the most important is the attitude and productivity of a litigation-independent validation testing program to determine error rates and standard operating protocols with data requirements. Currently, only the forensic computational linguistics approach carries on a validation testing program on forensically feasible data (i.e., experimentally collected data or actual known data from cases and investigations, where all the data is "ground truth" for the tested method). Traditionally, computational forensic linguistics has focused on using hard-to-imitate, low-salience features that are psychologically real and theoretically valid: these features are syntactic structures categorized in a particular way.[2] Further, this feature set has yielded high accuracy (94%-95%) for author identification in two different datasets, one experimentally collected and the other from case investigations. Currently, this methodology has been used for both identification and verification. When used for an identification, the method requires a "line-up" of suspect authors who are each tested pair-wise against each other, with the resulting statistical model used to classify the questioned document to one or the other. If there is a large group of suspects, a binomial test can then be used on the multiple pairwise classifications, but often there are only two suspects (for reasons related to the case outside any linguistic analysis and unknown to the forensic linguist). When used for a verification, the questioned document must be long enough that it can be split into segments and tested against the known suspect documents; although the discriminant function procedure forces the documents into two groups, a single author's documents often cannot be distinguished into two groups.

Using this same feature set on different datasets, this study is testing traditional forensic methodology using a pairwise procedure to determine levels of accuracy and data quantity with both statistical and machine learning classifiers. Thus the current experiments are testing if previous results using discriminant function analysis can be replicated at the 94%-95% accuracy with other statistical classifiers such as logistic regression. Further, the current experiments also test if machine learning classifiers like Support Vector Machine can attain even higher accuracy.

Again using the same feature set on different datasets, this study is also testing a novel biometric methodology to determine if this feature set can be used for verification. Distance and similarity matrices are being used with threshold settings to see how well this feature set performs as a verification metric. Further, a Bayesian analysis is being tested to see if both identification and verification protocols in this framework can be conducted.

References:

1. Chaski. C.E. 2013. "Best Practices and Admissibility in Forensic Author Identification." Journal of Law & Policy, Brooklyn Law School, Brooklyn, New York.
2. Chaski, C.E. 2005. "Who's At the Keyboard? Recent Results in Authorship Attribution." International Journal of Digital Evidence. Volume 4:1. Spring 2005. Available at http://www.ijde.org.

Author Identification, Author Verification, Bayesian Likelihood

** Presenting Author*