Best Practices for Data Set Development in Authorship Attribution and Beyond

Erin Smith Crabb
University of Maryland College Park
College Park, MD 20740, USA
ecrabb@umd.edu

Abstract
With the rise in popularity of communication forms like social media, online forums, and text messaging, text documents have begun to play a more important role within civil and criminal investigations. In order to effectively and scientifically analyze these documents, analytical methods, whether they rely on machine learning techniques or qualitative analysis, must be tested.

Forensic linguists have proposed a variety of criteria designed to address the discrepancies and challenges faced by researchers and professionals in need of data for testing their approaches. However, a review of recent research shows that many data sets do not adequately meet many of these standards, especially that of collecting and utilizing ground-truth data, which can be especially difficult to gather from online settings.

Author Keywords
Data collection techniques; large-scale data; ethics in data use and collection; authorship attribution.

ACM Classification Keywords
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing --- linguistic processing; K.4.1
Introduction
Since the advent of textual communication, questions have been asked about the authorship of documents. For example, did Shakespeare really write 38 plays, 156 poems, and numerous other works (Montagne, 2008)? Was the Book of Isaiah written by one person or several (Adams, 1984)? The number of documents with questioned authorship is always on the rise; with the growing popularity of social media and short-form messages, anyone with a working internet connection has the ability to create and disperse content under any assumed identity. This poses a difficult problem for the consumer of the data: who actually wrote the text in question? Was it the author given on the byline or someone else? This problem is especially relevant to law enforcement, where lives can depend on the ability for an investigator to uncover the truth, even when faced with uncooperative or misleading evidence or witnesses. In order to address this growing need, a variety of computational and qualitative text analysis methods have been investigated and scientifically tested.

Proposed guidelines for authorship attribution
Not all methods are created equal, and the diversity of approaches and data sets used for testing them lead to difficulties comparing the results obtained by different studies. For example, Chaski (2005) tests her algorithm for authorship identification on genre-diverse texts elicited from participants in an experimental environment, while Grant (2013) focuses exclusively on determining authorship of SMS text messages sent and received from cellular telephones using a qualitative analytical process. Despite their differences, however, both approaches have played roles in criminal investigations, and as such, should share a minimum number of characteristics which legitimate expert witness testimony.

The ruling in Daubert v. Merrell Dow Pharmaceuticals laid out the criteria that determine admissibility of all forms of expert witness testimony in United States courtrooms. Based on the ruling, Chaski (2013) outlines a list of ten criteria that she believes should be met in order to observe best practices in forensic linguistics. She states that methods for authorship attribution should be:

1. Developed independent of any litigation;
2. Tested for accuracy outside of any litigation;
3. Tested for accuracy on “ground truth” data;
4. Able to work reliably on “forensically feasible” data;
5. Tested for known limits correlated to specific accuracy levels;
6. Tested for any errors of individual testing techniques that could cause accumulated error when combined with other techniques;
7. Replicable;
8. Related to a specific expertise and academic training;
9. Related to standard (“generally accepted”) techniques and within the specific expertise and academic training; and
10. Related to uses outside of any litigation in industries or fieldwork in the specific expertise (Chaski, 2013, p. 334).
While these criteria are meant to address methods, several of them can apply or apply exclusively to the data upon which those methods are tested. I propose that data sets used for testing methods should (1) be developed independently from litigation, (3) contain data with a known ground truth, (4) be "forensically feasible", (6) be examined for any collection-related errors or imbalances (especially within large sets) which could affect algorithmic outcomes, and when appropriate, (9) related to standard collection techniques documented by scholarly literature. If researchers ensure that their data sets follow these recommended practices to the best of their ability, they will take great strides toward being able to certify that the resulting programs are thoroughly tested and based on the most current, valid research available. Let us examine some of these (1, 3, and 4) in turn.

**Best practices for data set development**

In order to incorporate Chaski’s (2013) list of best practices into data set development, let us identify some concrete steps that can be followed by researchers when collecting their data. The first criterion (1), “Developed independent of any litigation,” is one that is commonly met by data sets used by many quantitative researchers. Many researchers collect their own data from online sources such as Twitter, online forums, blogs and newspapers, and train their algorithms on it; they then test their established programs on evidence collected during investigations (Argamon et al., 2007; Chaski, 2005; Chaski, 2013; Stamatos, 2009). However, separating training and test data can be more problematic for researchers and professionals relying on qualitative methods; for example, Grant (2013) does not describe data used to test or develop his approach, aside from the evidence provided for the investigation he had been undertaking. Despite this added challenge, however, those using qualitative methods ought be able to satisfy this guideline by providing examples of their approach as applied to data gathered online or from other known sources.

While meeting the letter of (3) (incorporating data with a known ground truth) seems simple, it is very challenging to establish ground truth outside of experimental space. Due to the ease of adopting alternative identities online, verifying the authors of online content can be difficult, especially when it is collected as part of a large data set and may not be directly viewed by a human. In order to circumvent this problem, some researchers have adopted other methods for trying ground truth (Afroz et al., 2014; Johansson, Kaati, & Shrestha, 2013, Stolerman et al., 2014). However, all of these include one assumption: that the account’s content is being created and distributed by one person, unless it is otherwise credited. This is frequently not a reasonable assumption; plagiarism, whether intentional or not, is not rare online.

According to Chaski (2013, p. 337), “For the methods [to be accurate], ground-truth data must be forensically feasible, i.e., the same kind of data that is obtained in actual cases.” However, forensic feasibility possesses a flexible definition; the features studied for a case involving 10 years of corporate records could be very different from what are needed to compare two short, essay-length writing samples (Basim, 1967). In addition, there are challenges to the definition posed by researchers. For example, Almishari, Kaafar, and Oguz (2014, p. 4) acknowledge that the number of tweets
selected for each author in their study was subjective, and that they chose authors with a large number of posts. Samples will not be as abundant for all cases (see Basim (1967), in which only one known sample was found). Thus, it is important to acknowledge that forensic feasibility will differ depending on the investigation at hand, and using data sets with samples of similar size and genre will be the best way to ensure similar conditions for training and test data.

In conclusion, despite the proliferation of text-based data, standards for creating data sets have yet to be developed. This is in part due to the diversity of data sources and forms, which is both a strength and a challenge. The criteria advocated here support the creation of data sets able to evaluate and test both qualitative and quantitative approaches, offering further opportunities for methodological comparison and theoretical advancement.

Acknowledgements
I would like to thank my advisor, Dr. Paul Jaeger, for his help in editing and refining this paper.

References