

Review

Scalability Issues in Authorship Attribution.
Kim Luyckx. *Brussels: University Press
Antwerp, 2010. xvi + 180 pp. ISBN 978-90-
5487-823-0. € 29,95 (paperback).*

The field of automated (non-traditional) authorship attribution is (not to put a fine point on it) a mess. While many significant accomplishments have been achieved, the field is highly fragmented, with little to no general theory or deep understandings of the strengths and weaknesses of different methods. Indeed, there is very little agreement, if any, on standard evaluation methods, so it is nearly impossible to really measure progress in the field as a whole. A large part of the reason for this state of affairs is the highly multidisciplinary history of the field (as well as the dearth of large, well curated, publicly available corpora for testing and evaluation). There have been few attempts at large-scale systematic experimentation comparing different attribution techniques on different sorts of problems, with the goal of deriving more general understandings than particular studies can give.

Scalability Issues in Authorship Attribution, by Kim Luyckx, represents an important, though necessarily incomplete, step towards the goal of a general empirical theory of (automated non-traditional) authorship attribution. The book's aim is to systematically examine questions of *scalability* in authorship attribution, i.e. what effect variations in topic distribution, the number of features, the number of candidate authors, and the sizes of the texts have on attribution reliability. It is a focused monograph, and does not claim to provide a general survey of the field (nor does it); excellent background on the field can be found in Juola's (2006) and Stamatatos's (2009) recent surveys.

Part I (Chapters 1 through 3) of *Scalable Issues* describes the goals and methodology of the work,

which examines questions of attribution scalability squarely from a standard machine learning perspective, dividing the technical aspects of the problem into feature selection and classifier learning. Thus, while the work addresses one of the major approaches to authorship attribution, conclusions are less applicable for approaches that do not involve machine learning, such as Burrows's (2002) Delta method, or explicit classification, such as clustering or dimensionality reduction (e.g. Hoover (2003) and Binongo (2003), as well as many others). It would be helpful to see extension of the study's methodology to such approaches as well.

Within the text classification approach, the book covers the range of generally applied methods fairly well. Language features that are explored as inputs to authorship attribution include commonly used complexity features (type/token ratio, word and sentence length, etc.), lexical features (frequencies of content and function words as well as *n*-grams), character *n*-gram frequencies, and syntactic features (part-of-speech *n*-grams, shallow parsing chunk types, grammatical relations, etc.). The most common feature selection methods are compared (in Chapter 4, discussed below), including chi-squared, information gain, and various term frequency-based criteria. The primary machine learning method used is the Tilburg memory-based learner (TIMBL) system for memory-based learning, based on the *k*-nearest neighbor approach, which essentially classifies test examples according to the categories (authors) assigned to the closest training examples (in geometric terms). An oddity is that while the vast majority of the experiments described in the book are done using TIMBL, the few results (in section 6.5.2) in which TIMBL's performance is compared to that of the support vector machine (SVM) method, another dominant

technique in text classification, SVM tends to work noticeably better. One wonders, therefore, whether the conclusions of the earlier experiments might have been different had SVM been applied instead of TIMBL.

Part II describes several experiments that examine different scalability questions in authorship attribution. Chapter 4 explores variations in experimental design, specifically feature selection and its effect on topic-independence, and cross-validation as a way of measuring the ability of a method to generalize in a scalable fashion across topics. The chapter does an excellent job of examining a variety of feature selection methods, and how the novel notion of a ‘topic frequency threshold’ for selection can lead to greater topic-invariance. Somewhat less clear was the section on cross-validation, which left me somewhat confused as to whether cross-validation was viewed as a method of attribution or a method for *evaluating* attribution techniques. It should have been made clearer that it is the latter.

Chapter 5 explores variation in number of candidate authors and its effect on attribution accuracy. As might be expected, the larger the number of candidate authors, the lower the attribution accuracy. While this effect holds within each data set, the difficulty of attribution for different data sets is shown to vary as well. Luyckx also finds that performance on small numbers of authors systematically overestimates performance for larger numbers of authors, concluding rightly that experiments must explicitly test performance for large as well as for small numbers of authors.

Finally, Chapter 6 examines the effects of variation and imbalance in data size among the candidate authors. Here too, the primary result is negative, that the text classification approaches tested (including both TIMBL and SVM) tend to not work well in situations with smaller amounts of training data per author. One interesting, and to my knowledge novel, result is that there is a greater decrease in accuracy at smaller data sizes for feature sets that perform better at larger data sizes. This points towards the need to integrate multiple types of features for robust attribution.

As noted, while the studies described in *Scalability Issues* are clearly limited in scope, the work occupies an important place in moving authorship attribution research from being a scientific art to being an empirical science.

One distinctive aspect of the work is the parallel examination of authorship questions in multiple languages (English and Dutch), with an effort to make experiments in the different languages comparable. This is a step up from most studies which examine particular questions in one language, but more extensive studies need to be done to derive conclusions that can be plausibly generalized across languages and language families.

A key aspect of the methodological design, and rightfully so, is the attempt to look at multiple topics of texts. As is well recognized, the same features that tend to vary with authorship of a text may also vary with topic; indeed, different authors may tend to write on different topics, so that topic itself may be used as an indicator (however fallible) of authorship. Hence an attribution model that works well for texts of a certain topic may fail utterly if tested on texts of a different topic or genre. *Scalability Issues* is one of very few studies (and likely the most extensive to date) that addresses the question of attribution topic-independence head-on, utilizing the multiple text topics in each of its test corpora for this purpose. A related issue, not addressed here, is that of genre (or text type, but we can bracket the question of how to define ‘genre’ here). Since many of the same lexical and syntactic features that are used as authorship cues also vary across different text types, an attribution model that works for, say, novels will not likely work as well for, say, critical essays. While one of the corpora studied here (the Dutch Authorship Benchmark Corpus) does include texts in three different genres, the question of cross-genre attribution is, quite reasonably, left for future work.

The way forward in authorship attribution research is, primarily, clarification and standardization of experimental methodology and the concomitant development of a body of rigorous (and general) empirical results. *Scalable Issues* is a significant contribution to the field which helps to move it in this direction. The book is well worth

reading, both for the specialist in authorship studies as well as the technically savvy non-specialist with an interest in the field.

References

- Binongo, J. N. G.** (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, **16**(2): 9–17.
- Burrows, J.** (2002). ‘Delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.
- Hoover, D. L.** (2003). Frequent collocations and authorial style. *Literary and Linguistic Computing*, **18**(3): 261–86.
- Juola, P.** (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.
- Stamatatos, E.** (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–556.

Shlomo Argamon
 Illinois Institute of Technology, USA
 doi:10.1093/llc/fqr048