Lang Resources & Evaluation (2011) 45:1–4
DOI 10.1007/s10579-011-9136-1

Plagiarism and authorship analysis: introduction to the special issue

Efstathios Stamatatos · Moshe Koppel

Published online: 29 January 2011

1 Authorship attribution and plagiarism analysis

The Internet has facilitated both the dissemination of anonymous texts as well as easy "borrowing" of ideas and words of others. This has raised a number of important questions regarding authorship. Can we identify the anonymous author of a text by comparing the text with the writings of known authors? Can we determine if a text, or parts of it, has been plagiarized? Such questions are clearly of both academic and commercial importance.

The task of determining or verifying the authorship of an anonymous text based solely on internal evidence is a very old one, dating back at least to the medieval scholastics, for whom the reliable attribution of a given text to a known ancient authority was essential to determining the text's veracity. More recently, the problem of authorship attribution has gained greater prominence due to new applications in forensic analysis, humanities scholarship, and electronic commerce, and the development of computational methods for addressing the problem.

Over the last century and more, a great variety of methods have been applied to authorship attribution problems of various sorts. One can roughly trace the evolution of methods through three main stages. In the earliest stage researchers sought a single numeric function of a text to discriminate between authors. In a later stage, statistical multivariate discriminant analysis was applied to word frequencies and related numerical features. Most recently, machine learning methods and high-dimensional textual features have been applied to sets of training documents to

E. Stamatatos (✉)
Department of Information and Communication Systems Engineering, University of the Aegean,
83200 Karlovassi, Greece
e-mail: stamatatos@aegean.gr

M. Koppel
Department of Computer Science, Bar-Ilan University, 52900 Ramat-Gan, Israel
e-mail: koppel@cs.biu.ac.il

construct classifiers that can be applied to new anonymous documents. Several recent papers survey this literature and describe both text representation techniques and classification paradigms (Juola 2008; Koppel et al. 2009; Stamatatos 2009).

Roughly speaking, authorship identification divides into so-called *attribution* and *verification* problems. In the authorship attribution problem, one is given examples of the writing of a number of authors and is asked to determine which of them authored given anonymous texts. In the authorship verification problem, one is given examples of the writing of a single author and is asked to determine if given texts were or were not written by this author. As a categorization problem, verification is significantly more difficult than attribution.

One task of authorship analysis that has drawn special attention in recent years is the problem of plagiarism. The plagiarism problem itself can be divided into two types: *extrinsic* analysis and *intrinsic* analysis. In the extrinsic case, we wish to detect plagiarism by finding near-matches to a text in a database of texts. In intrinsic detection, we wish to show that different parts of a presumably single-author text could not have been written by the same author. Extrinsic plagiarism analysis is actually more closely related to algorithmic issues involving approximate pattern matching than to other authorship attribution problems. Intrinsic plagiarism analysis, however, is very tightly tied to the problem of authorship verification since stylistic inconsistencies within a text indicate parts written by different authors.

2 The PAN workshops

Despite considerable progress in research on these problems, authorship analysis and plagiarism detection have not yet matured as a discipline in the sense that the field has yet to develop standard, large-scale resources and consensus evaluation techniques for comparing different methods.

Consequently, the need has been felt in recent years to place the field of authorship analysis and plagiarism detection on a firm scientific footing. The PAN workshops in Amsterdam (2007); Patras (2008); San Sebastian (2009), and Padua (2010) brought together researchers for the purpose of mapping the relationships among the central challenges in the field, suggesting methods for meeting these challenges and establishing resources and standards for evaluation (Stein et al. 2007; Stein et al. 2008; Stein et al. 2009; Stein et al. 2010). Another important aim of the PAN workshops was to make a bridge between the scientific communities of authorship analysis and plagiarism analysis so that to handle problems with many similarities (e.g. authorship verification and intrinsic plagiarism analysis) more effectively. This volume was originally motivated by the progress made at the PAN meetings and includes more mature versions of some of the papers presented there (as well as some others).

The plagiarism detection competition introduced in conjunction with PAN 2009 and continued in 2010 as an evaluation campaign in the framework of CLEF 2010 provided a unique opportunity for researchers to test their approaches using common resources and evaluation criteria. Both external and intrinsic plagiarism detection tasks were included in the competition. Large scale corpora were built incorporating simulated and artificial plagiarism cases. Moreover, appropriate evaluation measures were defined focusing on the recall and precision of plagiarized passages as well as the ability of a tool to detect a plagiarized section as a whole or in several pieces (Potthast et al. 2010). The robustness of the participants' methods can be examined in detail according to several factors, e.g. dealing with texts of varying text-length and varying degree of obfuscation in plagiarized passages. We believe that the plagiarism detection competition should serve as a model of testbed development for the authorship attribution field generally.

3 LRE special issue

The call for papers for this special issue was published in January 2009. The five papers included in this volume (from among 13 submissions) cover the range of problems in authorship attribution and plagiarism detection.

Clough et al. "Developing a Corpus of Plagiarized Short Answers" provides a corpus of plagiarized short answers that can be used as a testbed for the evaluation of plagiarism detection methods. This new corpus significantly broadens existing testbeds by incorporating a previously under-developed, though important, genre of texts.

Lavergne et al. "Filtering Artificial Texts with Statistical Machine Learning Techniques" (originally presented at PAN-08) considers the problem of automatically distinguishing authentic texts from computer-generated texts typically used as filter for purposes of spamming. They demonstrate the strengths and weaknesses of simple methods based on lexico-graphic features and on language models and show that a new entropy-based method covers certain cases not handled by the simpler methods.

Potthast et al. "Cross-language Plagiarism Detection" (originally presented at PAN-08) considers the particularly difficult case of extrinsic plagiarism analysis in which near matches have to be found across languages. They systematically compare a number of algorithms for achieving this based on texts written in six major European languages.

Stein et al. "Intrinsic Plagiarism Analysis" (originally presented at PAN-07) notes that authorship verification and intrinsic plagiarism analysis are actually isomorphic problems. They exploit this connection to offer a systematic approach to intrinsic plagiarism analysis based on fundamental algorithmic building blocks and provide empirical comparisons of several approaches.

Finally, Koppel et al. "Authorship Attribution in the Wild" considers the problem of authorship attribution in cases where the candidate set might contain many thousands of candidate authors, possibly none of which is the actual author. They show that information retrieval methods combined with confidence measures based on random selection of feature subsets can yield high precision even for such difficult attribution problems.

Acknowledgments The editors gratefully thank Benno Stein, co-organizer of the PAN workshops, for his invaluable assistance throughout the editing process. Many thanks to Paolo Rosso (co-organizer of

PAN-09, and PAN-10), Martin Potthast and Alberto Barron-Cedeno (co-organizers of the competition on plagiarism detection held in conjunction with PAN-09 and PAN-10).

References

Juola, P. (2008). Author attribution. *Foundations and Trends in Information Retrieval, 1*(3), 233–334.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology, 60*(1), 9–26.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An evaluation framework for plagiarism detection. In: *Proceedings of the 23rd international conference on computational linguistics (COLING 2010)*, Association for computational linguistics, pp. 997–1005.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538–556.

Stein, B., Koppel, M., & Stamatatos, E. (Eds.). (2007). Plagiarism analysis, authorship identification, and near-duplicate detection (PAN-07). *SIGIR Forum*, *41*(2), 68–71, ACM.

Stein, B., Stamatatos, E., & Koppel, M. (Eds.). (2008). *ECAI 2008 Workshop on uncovering plagiarism, authorship, and social software misuse* (PAN-08).

Stein, B., Rosso, P., Stamatatos, E., Koppel, M., & Agirre, E. (Eds.). (2009). *SEPLN 2009 Workshop on uncovering plagiarism, authorship, and social software misuse* (PAN-09).

Stein, B., Rosso, P., Stamatatos, E., & Koppel, M. (Eds.). (2010). *CLEF 2010 Workshop on uncovering plagiarism, authorship, and social software misuse* (PAN-10).