Using Semantic Features for Authorship Attribution

Harry Smith
Columbia University
hs3061@columbia.edu

1 Introduction

Authorship Attribution (AA) is a problem in Natural Language Processing concerned with determining the author of a sample of text. AA is, in general, a supervised learning task wherein a corpus of documents of known authorship are available for training a classification model. An investigator might use AA models for a variety of tasks: identifying the author of a threatening post on an extremist web forum (Stamatatos, 2018), discerning information about a document's author without guessing their actual identity, or confirming if one particular person is truthful in claiming authorship of some text (Stamatatos, 2009). In a document, several levels of information are available for completing the classification task, ranging from syntactical and lexical to semantic and functional (Argamon et al., 2007). Prior research in AA has demonstrated the effectiveness of the "lower level" lexical features for classification, but research continues into developing efficient and widely applicable analysis of the semantic information contained in text. Furthermore, as media of writing have evolved with the ubiquity of computers, recent work has demonstrated that the success of AA strategies depends strongly on the genre of writing (Stamatatos, 2018). While traditional techniques found success on classic works of fiction, the emergence of new genres of writing like tweets or texts has necessitated the discovery of new methods in AA.

Our research attempts to assess how an author's use of features at a higher level than the syntactical or lexical contribute to that author's unique voice. In particular, we attempt to answer the following question: *how does the genre of writing influence the extent to which an author's semantic structure contributes to personal style?* While previous work has shown that syntax alone is a strong indicator of personal style, it is possible that certain genres of writing allow semantic qualities to be contribute more directly to an author's voice. That is, we hypothesize that the addition of semantic features will *slightly* improve AA model performance when examining documents in long, literary genres (e.g. novel, essay, etc.). Further, we anticipate that the addition of semantic features will *greatly* improve AA performance for short documents like tweets, emails, and chats where syntactic information is relatively sparse. Finally, we hypothesize that the inclusion of semantic features with standard lexical features provides a weaker improvement to distinguishing among authors of a given genre when the range of time from which the texts are drawn is larger.

2 Literature Review

We begin by reviewing the current landscape of methods available in AA. Stamatatos provides a useful summary in his review, categorizing the available methods by the types of features that they explore (2009). Stamatatos notes that the problem of AA is essentially two-fold: first, investigators must determine the *stylometric features* of interest from a document or corpus; second, they must decide the means by which they apply these features to actually attribute an author. Although the focus of our research is mainly on the former component of the AA problem, it remains important to consider exactly how the features an AA model incorporates will be used. Indeed, to properly address the hypotheses that we have posed, it will be crucial to control for the effects of latter component, likely by comparing the performance of feature choices in only one model of attribution. In his review, Stamatatos enumerates the different types of stylometric features and provides a working definition for each that we will adopt in

this study. First, he defines the *lexical features* , which are those that can be extracted by use of a tokenizer tool. Examples of these features include word counts, n-gram counts, and measures of vocabulary richness. Stamatatos notes, through summary of previous studies, that a "simple and successful method to define a lexical feature set" (Stamatatos, 2009) is to include only the counts of the most frequently used words, where the threshold for frequency rank is chosen between 100 and 1000. Second, he defines *character features,* which are gathered by observing the sequences of the characters in the text independent of the words that these make up. These features include character n-grams of fixed or variable length and character classes. Stamatatos notes that many studies through the past several decades have had good success with character features. Third, he defines *syntactic features* as those which examine the roles of words and sentence structures in writing. These features include part-of-speech (POS) tags and sentence or phrase structures. While Stamatatos highlights several studies which have successfully used these features, he notes that they are more difficult to derive from the original text than lower-level features. Most important for this study, Stamatatos defines *semantic features* as those which depend on the meaning of the text. Examples of semantic features include the use of synonyms, the topic of the writing, and the functions of sentences. He notes that these features are often noisy, that they are the most difficult to extract from the text, and that they have the most sparse record of successful use in the field of AA.

Thus, we proceed to a review of research that attempts to create or explore semantic features at this frontier. We turn to Argamon et al.'s work on the use of functional lexical features for text classification and authorship attribution (2007). Here, the authors leverage the notion of a *Systematic Functional Grammar* (SFG) to procedurally define the function of words in the text. Words are categorized into broad *systems,* which provide the general purpose for their inclusion: cohesion, assessment, and appraisal. Words tagged under the cohesion system provide elaboration, extension, or enhancement of meaning. Assessment words provide either modality, for the possibility or typicality of some event or object (e.g. its type or value), or comment, for the opinion of the author on the event or object (e.g. admission, validation, evaluation). While these systems admit a more complex classification than we have described here, it is sufficient to understand that the use of an SFG allows an investigator to use a machine learning model to assess the function of different words in the text and count their relative frequencies, thus producing features that have semantic meaning and that respect the context of the writing.

Argamon et al. explore the effectiveness of these features in several AA experiments. On a corpus of twenty 19th century novels, the authors perform ten-fold cross-validation of their attribution models using different feature sets admitted by including and excluding the relative frequencies of different systems of functional features. They find that in the task of pure AA, including any subset of the three aforementioned systems to a basic lexical feature set provides a slight yet statistically significant improvement. In the related tasks of book attribution (which book is a given excerpt from?) and nationality attribution (is the author British or American?), the authors find similar positive results. Thus, we find evidence that the addition of semantic features to an AA model may improve its performance.

Feng and Hirst (2013) provide a different potential semantic feature to include in AA models: *Local Discourse Coherence.* The authors build from earlier computational linguistic research showing how to model the references that authors naturally make to the entities they introduce in their text: for example, in this sentence, we reference "the authors", wherein we invoke "Feng and Hirst", mentioned in the sentence prior, as subjects. In order to model these patterns of coherence, the authors adopt the model of a document as an entity grid, where the rows represent the sentences of the text and the columns represent the entities that an author has referenced. Entries into the grid take one of four forms: *subject, object, other, none,* where the entries represent the type of reference made to the column's entity in the sentence represented by the row. The authors then extract estimated probabilities of given transitions (e.g. *[subject, none, subject]*) occuring as semantic features.

Feng and Hirst then evaluate the performance of a one-layered neural-network when provided with lexical features only, coherence features only, and both feature types together. In the task of pairwise AA for several 19th century novels written in English, the model is unable to match the per-

formance of lexical features when provided only the coherence features; however, for many of the pairwise tasks, the model's performance with *both* lexical and coherence features improves with statistical significance over cases where only lexical features are available. Similarly, in the task of one-class classification (i.e. Did Nathaniel Hawthorne author this document or not?), the combined feature set corresponds with a boost in performance over either feature set alone. While this represents yet another promising semantic feature set for AA, we note that both this feature set and the functional feature set both require the use of a separate machine learning model in order to automatically derive these features from text (Argamon et al., 2007; Hirst and Feng, 2013). This results in feature sets that are both computationally expensive to obtain and relatively noisy compared to lower level lexical- or character-based feature sets.

We move to conclude our review with brief looks at two other illuminating works. The first is another investigation by Stamatatos, wherein he investigates the use of text distortion as a preprocessing step to improve AA performance (Stamatatos, 2018). Stamatatos hypothesizes that by masking information related to the topic of the writing, AA models based on character features will improve when attempting to attribute authorship to documents from either many topics, many genres, or both. Although this work is based in low-level feature sets only, Stamatatos provides a relatively novel look at the performance of models when topic and genre are not held constant through the set of documents to be processed. In particular, this research supports our hypotheses that models will behave with varying performance depending on the genre of writing to which we apply them.

Finally, we turn to Hedegaard's research on applying semantic features to perform AA on translated works (2011). He performs a similar experiment to those that we have seen from Argamon and Hirst, but he presents a semantic feature set based on the frequencies of "semantic frames", which are sequences of words identified by their structure and meaning. Although the Hedegaard is vague with respect to the specific methods of this research, he notes that he uses semantic frames from FrameNet, a publicly available database of frames. This presents the possibility for a semantic feature set which does not require the prior tuning and exectution of a separate, novel machine learning model before beginning the AA task, which is highly desirable. The potential for its relative ease of use, coupled with the boosts in performance that Hedegaard observes from his models when including the semantic frame frequency features, suggests that using predefined features may represent a promising path towards understanding the role of semantics in authorship attribution.

References

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Steffen Hedegaard and Jakob Grue Simonsen. 2011. Lost in Translation: Authorship Attribution using Frame Semantics BT - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pages 65–70. Association for Computational Linguistics.

Graeme Hirst and Vanessa Wei Feng. 2013. Patterns of local discourse coherence as a feature for authorship attribution. *Literary and Linguistic Computing*, 29(2):191–198.

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

Efstathios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473.