Models for Authorship Attribution based on Corpus Characteristics

Furaha Damien
McGill University

Irene Grose
McGill University

Albert Faucher
McGill University

{furaha.damien, irene.grose, albert.faucher}@mail.mcgill.ca

Abstract

This paper provides an analysis intended to aid model selection when developing a solution for an application-specific instance of authorship attribution. In this work, we replicate similarity-based methods with Integrated Syntactic Graphs (ISGs) presented by Gómez-Adorno et al. (2016). We also examine supervised learning with Support Vector Machines (SVMs) using a "locally-weighted bag of histograms" feature vector, following Escalante et al. (2011). The aim of this investigation is to evaluate the performance of both models on a range of corpora with varying characteristics, including the number of candidate authors, the number of documents per author, and the content type of each source. While overall performance was low, a negative trend with the number of authors is observed, which is not mitigated by increase in documents per author. Interesting results are seen when tests are run on mixed corpora.

1 Introduction

1.1 Motivation

In the past, methods for authorship attribution in natural language processing research have covered machine learning approaches (e.g. Escalante et al., 2011), and similarity-based approaches (e.g. Gómez-Adorno et al., 2016). These works have demonstrated that machine learning methods for authorship attribution can yield higher performance on the same input corpus, when measured in accuracy; however, machine-learning methods are considered unsuitable when the number of classes (i.e. authors) in the corpus is excessively high (Koppel et al., 2014).

Given that authorship attribution can have great practical benefit in fields such as law enforcement or digital forensics (Stamatatos, 2009), where the pool of potential authors could be very large, it would be useful to have a standard for which method - machine learning or similarity comparison - might serve better based on the characteristics of the available corpus.

1.2 Hypothesis and Experimental Approach

A review of the literature on authorship attribution identified two implementations for each approach which will be replicated in this experiment. For machine-learning with SVMs, the proposed "locally-weighted bag of words" (LOBOW) feature vector was demonstrated to give high performance on the C10 subset of the RCV1 corpus (Escalante et al., 2011). For a similarity-based method, the modelling of an author profile as an integrated syntactic graph (ISG) was described by Gómez-Adorno et al. (2016) for the same corpus and will be used as a starting model instead.

Our investigation into the impact of corpora characteristics covers three main settings: the number of authors in the corpus, the number of documents available per author, and the breadth and variation of writing styles present in the corpus. We rely on the availability of pre-existing corpora, but also examine model performance on corpora collected by scraping online forums.

For each corpus type, there is an intuitive hypothesis that we seek to confirm experimentally. For example, one might predict that labelling documents for two authors would be easier than for fifty authors. Similarly, more documents per author would provide a better representation of an author's writing style, and if the corpus contains authors with highly distinctive writing styles it would naturally be easier to distinguish them. For each experimental conditions, we seek to confirm or deny the presence of these trends, while comparing the results of an SVM model against an ISG model.

The findings of this experiment are not highly conclusive, but there is a suggestion that characteristics of a corpus can impact model performance in authorship attribution. The testing procedure brings light to another dimension of the task that should be considered: the time cost for training and testing a model. Even if one model is theoretically more powerful, the cost of implementation is a necessary consideration in deciding which model to use for a given task, particularly if subject to constraints in computational power.

The remainder of the report will provide an overview of related works in Section 2; a description of the corpora, experimental set-up, and testing procedure in Section 3; results and evaluation measures in Section 4; and finally, discussion and conclusion in Section 5.

2 Related Works

2.1 Authorship Attribution

The case of authorship attribution can be represented as a single-label multi-class classification task, in which the goal is to automatically identify the author of a given text from a set of predefined candidates (Statamatos, 2009). Usual tasks in text classification strive to find thematic similarities between texts based on content, but in authorship attribution the aim is to model a style of writing independently of context (Escalante et al., 2011). Because of this, methods for representing documents in a way that provides information about writing style are necessary to perform well in authorship attribution tasks. Standard machine learning techniques including Bayesian classifiers and neural networks have been applied in the past (Escalante et al., 2011), but Support Vector Machines (SVMs) in particular are considered one of the best solutions, in main part due to an ability to handle thousands of features (Gómez-Adorno et al., 2016). Several previous works have explored the use of graphical structures to represent documents for similarity-based approaches to text classification, but the use of an ISG is most relevant to this experiment due to the general and practical method for its construction (Gómez-Adorno et al., 2016). Many relevant papers seem to focus on the idea of finding the ideal document representation method for authorship attribution tasks in general. This is useful to our analysis as our goal is to aid model selection for specific authorship attribution tasks, document representation is an important aspect to consider when devising a solution to an NLP problem. The most relevant papers will be briefly summarized below.

2.2 Integrated Syntactic Graphs

Gómez-Adorno et al. (2016) propose that graph structures are intuitive ways to represent textual data, and that shortest path traversal on ISGs could be a good method for generalizing feature extraction of textual patterns. An ISG is built using linguistic features over several domains of language; the goal is to provide as much relevant and important information as possible. Patterns found by such a method can be applied in many different document analysis tasks, including authorship attribution, and have demonstrated comparable performance to state-of-the-art techniques in the past (Gómez-Adorno et al., 2016).

2.3 Local Histograms of Character N-grams

In contrast to the emphasis on textual content in the ISG representation, Escalante et al. (2011) focus solely on stylistic information conveyed through local histograms over n-grams at the character level. The locally-weighted bag of words framework (LOBOW) preserves sequential information, which can be highly indicative of writing style for certain authors. This method of text representation outperformed many state of the art techniques for authorship attribution tasks, and was reportedly well-performing even in unfavourable conditions.

3 Experimental Methods

In order to investigate the performance of cosine similarity models against machine learning methods, several tests were performed repeatedly with both an ISG parser and a multiclass SVM.

3.1 Model Design

3.1.1 Libraries Used

As per Gómez-Adorno et al. (2016), the ISG parser was built using the dependency parser provided in the Stanford CoreNLP library (v.3.9.2) in Java, with feature extraction relying on the CoreNLP document preprocessor combined with supporting functions in nltk (v.3.4.5). For the SVM, a multiclass OneVsRestClassifier provided in sklearn (v0.22) was used to train the model; feature extraction was done primarily without library support

3.1.2 Building the ISG Parser

Based on the best-performing model reported in Gómez-Adorno et al. (2016), the ISG parser implemented for this experiment follows a profile-based approach, wherein each author is represented by a feature matrix summarizing linguistic features in various domains across a set of training documents. Following the procedure described by Gómez-Adorno et al. (2016), the ISG is structured as a tree. Nodes represent unique word type and part-of-speech, and edges are tagged with the dependency relation.

For a given author, the ISG is built up iteratively on a sentence-by-sentence basis, with common nodes being collapsed into the existing tree. Given a sentence, the dependency tree is calculated first. For each edge in the dependency tree, the endpoints are added or combined with existing nodes to the ISG, and the edge is tagged with the dependency relation. After all training sentences have been combined, the shortest path from the root node to every other node in the tree is found using Djikstra's algorithm. Each shortest path is used to build up the feature matrix for a given author: rows in the matrix indicate the endpoint of the path, and columns represent unique linguistic features including word type, dependency relation, and part-of-speech tag.

The final component of the ISG model is the comparison mechanism through a modified cosine similarity score. The following function provided in Gómez-Adorno et al. (2016) was used to score similarity between two feature matrices:

$$Sim(A,B) = \sum_{(i=1)}^{m} \left(\frac{\sum_{(j=1)}^{|V|} A_{i,j} \cdot B_{i,j}}{\sum_{j=1}^{|V|} \sqrt{(A_{i,j})^2} \sqrt{(B_{i,j})^2}} \right)$$

where A and B are the input matrices corresponding to documents D_1, D_2; m is the number of rows in the matrix shared by A and B; and $|V|$ is the size of the linguistic feature set.

3.1.3 Building the SVM

Although training and testing the model itself is trivial using sklearn, feature extraction of LOBOW is more complex to implement. Unlike a bag-of-words model, LOBOW begins by representing a document as a function x(i,j), where input $i \in \{1, \ldots, N\}$ is a position in the character sequence and input $j \in V$ is a character-level trigram. The function outputs an 'add-δ' smoothed value of 1 if j is present at i and 0 otherwise (Lebanon et al., 2007); in this experiment the maximum number of features was limited to 2500 trigrams and no preprocessing (e.g. stemming, lemmatization, stop word removal) was done. The LOBOW feature vector preserves sequential information, lost in the usual bag-of-words, by computing the j^{th} component of the vector as $\int_0^1 x(ceil(tN), j) K_{\mu,\sigma}(t)dt$ (Lebanon et al., 2007). This can be viewed as a sum across all positions in the document of the product of the document function $K_{\mu,\sigma}(t)$ as the smoothing function. From Escalante et al. (2011), the smoothing function

$$K_{\mu,\sigma}(t) = \begin{cases} \frac{N(t;\mu,\sigma)}{\phi((1-\mu)/\sigma)-\phi((0-\mu)/\sigma)} & t \in [0,1] \\ 0 & t \notin [0,1] \end{cases}$$

is parameterized by μ to indicate a fixed location in the document and σ to control amount of sequential information preserved – as $\sigma \to \infty$ this reduces to a bag-of-words.

The secondary component to the SVM model is the kernel function. The kernel function used for this experiment is the Eucidean function described in Escalante et al. (2011), which computes a Euclidean distance between the input vectors.

3.2 Experimental Settings and Model Parameters

Several simplifications were made for the sake of testing in this experiment. Unfortunately, the large performance cost made it prohibitive to perform an extensive model selection process. A best effort to experiment across parameter values on very small data sets was made, and the full experiments were only run once for the chosen parameter values.

For the SVM, rather than integrating over all possible positions during feature extraction, a coarse Riemann sum is computed instead – a step size of 0.1 and 0.05 were both tested, and the kernel positions vector μ was fixed at two positions located one-thirds and two-thirds in the document. For comparison, the best settings found by Escalante et al. (2011) fixed μ at 20 locations over the document, but for performance reasons this was infeasible to replicate. The value of σ is set to 0.2 for all tests following the best setting found by Escalante et al. (2011). A value of δ = 0.01 was chosen for smoothing.

For the ISG, small experiments were run using lemmatized input vs non-lemmatized input. Otherwise, the same feature set of POS tags, word types, and dependency relations was used for all tests.

3.3 Corpus Processing

There were four primary corpora used for testing in this experiment. They are referred to as "GB" for the Project Gutenberg corpus available through nltk, "WP" for a web-scraped corpus of stories submitted to reddit.com/r/writingprompts, "HP" for works of fanfiction web-scraped from fanfiction.net's Harry Potter community, and "Blog" for a subset of the open-source blogger corpus (Schler et al., 2006). A secondary corpus referred to as "PP" was web-scraped from fanfiction.net's Pride and Prejudice community. There are 11 authors in GB, 247 authors in HP, 2459 authors for WP, 840 authors for Blog, and 183 authors in PP. For each test setting described below, a different random sampling of authors was drawn from each corpus with restrictions on the number of documents (sentences) taken per author, and the minimum size of a document allowed. Basic text-cleaning (e.g. whitespace stripping, fixing the encoding, sentence segmentation) was performed on the web-scraped corpora, and the blogger corpus was parsed into an acceptable format from the original XML files. The corpora were primarily left unmodified however, due to the possibility for unique attributes of writing style to manifest in choices such as punctuation, capitalization, etc.

3.4 Test Conditions

Over the course of testing, the SVM and ISG models were evaluated for two different conditions on the input corpus: the number of authors (3, 5, 10, and 50), and the number of documents used per author (20, 50, 100). Preliminary experiments indicated little effect from modifying document length itself, so this condition was excluded. The two models were further tested on specially constructed corpora that mixed documents and authors across different sources. In the first condition, an equal number of authors (2 or 3 each) were drawn from the GB corpus and compared against authors from the WP, HP, and Blog corpora. In the second condition, documents written by Jane Austen in the GB corpus were extracted to test specifically against documents written by 10 fanfiction authors in the PP corpus. This test condition used two modes: in the first, all fanfiction writers were grouped into the same 'author' class and compared against Jane Austen. In the second, each author was kept separate for a total of 11 authors in the corpus.

The testing procedure for the SVM follows the standard practice: the input corpus was split into 80% training data and 20% testing data using sklearn, and after computing the feature matrix, a classifier model was trained and scored for accuracy on the reserved test set.

For the ISG, a similar approach was used: after splitting the corpus line-by-line into 80% training data and 20% test data, all lines written by the same author were grouped and passed as input to the ISG parser. Each test document for the ISG composed of only one document, and the output author label was chosen as the author profile which had the highest similarity score

4 Results

For each of the tests described above, the primary evaluation measure used was accuracy. The score for each of the test conditions are as follows:

4.1 Number of Authors in Corpus

# of Authors	GB	HP	wp	Blog
3	0.25(0.25)	0.42(0.33)	0.33(0.50)	0.33(0.33)
5	0.20(0.40)	0.20(0.15)	0.15(0.10)	0.30(0.25)
10	0.05(0.05)	0.12(0.12)	0.10(0.23)	0.12(0.12)

Table 1: SVM % Accuracy for N Authors [20 Documents per Author]. Cells report accuracy when estimating the integral using a step size of dt=0.10 vs dt=0.05 as ”0.10(0.05)”.

# of Authors	GB	HP	wp	Blog
3	0.25	0.42	0.58	0.42
5	0.15	-	-	-
10	0.10	0.08	0.23	0.05
50	-	0.03	0.04	0.02

Table 2a: ISG % Accuracy for N Authors [20 Documents per Author]

# of Authors	GB	HP	wp	Blog
3	0.33	0.23	0.20	0.23
5	0.28	-	-	-
10	0.09	0.11	0.06	0.15
50	-	0.04	0.04	0.05

Table 2b: ISG % Accuracy for N Authors [50 Documents per Author]

# of Authors	GB	HP	wp	Blog
3	0.325	0.41	0.00	0.17
5	0.25	0.30	0.20	0.10
10	0.08	0.00	0.10	0.05

Table 2c: ISG % Accuracy for N Authors [20 Documents per Author + Lemmatization]

Tables 1 and 2a report the accuracy scores of tests run with the SVM model and ISG model (trained and tested as described in Section 3) when the number of authors is varied in the corpus; tables 2b and 2c reflect some extra test conditions used on the ISG model. In the data, there appears to be a fairly strong downward trend in accuracy for both models as authors increase, implying a negative impact on both methods. In fact, performance barely surpasses a random guessing classifier in these tests. Initial goals included testing the machine learning method for authorship attribution on a corpus of over 100 authors, but all tests for more than 10 authors were terminated early due to excessively long training periods. Some preliminary tests were done with the ISG model on 100 authors, but the results - while slightly faster to come - also followed the sharp decline in accuracy with near-zero scores.

For the extra conditions on the ISG, in the 50 document case there is an improvement over the 20 document case for the Gutenberg corpus but a general decline in the other three corpora. In the lemmatized input case, results seem fairly consistent in the Gutenberg corpus but not in the others. There does not appear to be enough consistency in the data to make a conclusion about the impact of these variations on the ISG model, and in both cases the relative performance against the SVM is roughly equivalent.

4.2 Number of Documents in Corpus

# of Docs	GB	HP	wp	Blog
50	0.20	0.16	0.15	0.10
100	0.28	0.24	0.25	0.17

Table 3: SVM % Accuracy for N Documents per Author. [N = 50, 10 authors; N = 100, 5 authors]

To accompany the ISG tests with 50 documents, tests were ran on an SVM model trained with 50 documents and 100 documents respectively. Here, there is a consistent increase in performance across corpora when the number of documents increases. Compared with the same number of authors on 20 documents reported in Table 1, there is also a general improvement to SVM performance with more data. Preliminary testing on document length showed no similar change in score, but this could potentially be related to the length-normalization step when extracting LOBOW features.

The observation that the SVM model appears to improve with more data while the ISG does not seems to be reflective of the nature of each method: with machine-learning, more training data provides more opportunity to learn patterns and make connections. Similarity-based methods, on the other hand, might experience a smaller beneficial impact in comparison.

The training and testing time for these tests spanned roughly 8 hours from start to completion for the SVM. The extent to which this is caused by insufficient computational power is unclear, but it seems evident that there is a sizeable drawback of increasing corpus size when extra data exists in the time it takes to train the model.

4.3 11 Authors in a One Vs. Rest Corpus

# of Authors	Accuracy
2	0.91
11	0.07

Table 4a: SVM % Accuracy on One vs Rest Corpus for 11 Authors [20 Documents per Author]

In the 2 author case, documents of 10 authors with 20 documents each are combined into a single ”Wrong Author” class and compared. In the 11 author case, the documents are separated back into original author labels.

# of Authors	Accuracy
2	0.82
11	0.07

Table 4b: ISG % Accuracy on One vs Rest Corpus for 11 Authors [20 Documents per Author]

In this set of tests, documents written by Jane Austen were compared with 10 authors who are in theory aiming to emulate Austen’s writing style. Due to the fact that works of fanfiction use the same names and settings, there is also likely to be overlap in lexical features that get recorded by the ISG. The most interesting observation with these tests is how wildly the accuracy changes when testing ”one vs the rest” compared to the standard testing set up. This could in part be due to the fact that the classifier has a good chance of guessing correctly at random for only two authors, but it seems likely that there are other explanations. Further tests that would have been interesting, given the time, would include tests varying the identity of the ”one” author, and tests where the ”rest” weren’t necessarily related in writing style to the ”one”.

4.4 N Authors on a Mixed Corpus

# of Authors	GB + HP	GB + WP	GB + Blog
2	0.44	0.38	0.31
3	0.46	0.54	0.38

Table 5: SVM % Accuracy on Mixed Corpus for N Authors [20 Documents per Author]

# of Authors	GB + HP	GB + WP	GB + Blog
3	0.25	0.30	0.17

Table 6: ISG % Accuracy for N Authors on Mixed Corpus [20 Documents per Author]

Legend
GB : Project Gutenberg Corpus(Source: nltk)
HP : Harry Potter FanFiction Corpus (Source: fanfiction.net, webscraping)
wp : Writing Prompts Corpus (Source: reddit.com, webscraping)
Blog : Blog Authorship Corpus (Source: Koppel et al., 2006)

This set of tests aimed to examine performance when the corpus content was more distributed across writing styles. A qualitative examination of the four corpora used shows that documents are fairly easy to classify by their source (as a human), based on factors like the punctuation, grammar, and content, with the most violent discrepancy in writing style existing between the Gutenberg novels and the Blogger corpus. As a result, equal numbers of documents were drawn from the corpora pairs above and tests were run. Performance of the SVM is better than ISG model, although in general both models’ scores are not significantly different from the non-mixed corpus case.

4.5 Evaluation Measures

The use of accuracy as an evaluation measure is a simplistic way to capture the performance of both the SVM and ISG. The greater the proportion of the test set they accurately classified, the better they learned how to distinguish authors. Our focus on accuracy also reflects the choices made in Escalante et al. (2011) and Gómez-Adorno et al. (2016).

Data beyond accuracy scores were collected during the testing phase, but are excluded due to space constraints from this report. Some of this data may help explain the observed performance measured in accuracy, however. For the SVM, the precision, recall, F score, and confusion matrix were saved; for the ISG, the mean and standard deviation of similarity scores with candidate authors were saved. In the case of the ISG, one interesting qualitative observation is that the standard deviation of similarity scores for each test document were all generally quite small, as was the difference in scores between the best scoring and next best scoring candidates. This implies that there could be a degree of ”goodness” that is unrepresented by a binary right or wrong measure.

Other evaluation measures could be very applicable to the authorship attribution case. For example, tests on the one-vs-rest schema corpus showed by far the best performance across the corpora and test conditions. As an automatic evaluation measure, the ability to correctly accept or reject a single author - authorship verification from a known list of candidates - might display different trends than seen here. It is also interesting to consider how human evaluation measures could be used to judge the relative goodness of the SVM vs ISG, whether through comparing accuracy of a machine model to a human judge, or by asking humans to score how close an incorrect machine is to the actual.

5 Discussion and Conclusion

The original intent of this experiment was to replicate the best-performing ISG and SVM models described in Gómez-Adorno et al. (2016) and Escalante et al. (2011) respectively, in order to test their relative performance on a variety of input data sets and confirm or reject the existence of trends in accuracy. To address the first point, the general performance of the SVM model was not very different from the ISG model under our test conditions. Unfortunately, our limited ability to properly test and extend the SVM model means that it is difficult to draw a strong conclusion about the relative performance of the two. It may yet be possible that at a more drastic variation on the parameters of the corpora, greater differences in performance would be seen, but this cannot be inferred from the test results in this experiment alone. Other test conditions, such as variation on the length of test documents used for the ISG, or the maximum size of the vocabulary for the SVM, would have also helped develop stronger conclusions about the relative performance of each model.

Regarding the trends in accuracy, the first most salient point is how low the overall performance was of both models across all test conditions. There are several possible explanations for this result. In both of these papers, a large range of parameters were tested and evaluated for the C10 subset of the RCV1 corpus: for example, the number of kernel positions and custom kernel function for the SVM, or the use of different linguistic feature sets for feature extraction for the ISG. We aimed to use the ”best-reported” settings identified in these papers, but a preferable approach would have been to re-tune and verify for each corpus condition how to best parameterize the models. In our experiments, the lack of extensive model selection due to constraints of time and computational power may be one main cause for the comparatively lower performance across all tests when judged against the original papers. Despite the overall low scores, there are still some trends that can be identified, as highlighted in Section 4.

6 Statement of Contributions

All members of Group 1 worked on various aspects of the project, and multiple teamwork meetings were organized. Albert predominantly worked on data clean up, citations, and report writing. Irene collected and processed the corpora, implemented the SVM and ISG model testing, and collected the raw test results. Furaha’s main contributions were to the SVM model implementations, along with LaTeX formatting and writing of the report.

References

[1] Argamon, Shlomo, Juola and Patrick (2011). Overview of the International Authorship Identification Competition at PAN-2011. [online] (1177). Available at: http://www.clef-initiative.eu/documents/71612/86377/CLEF2011wn-PAN-ArgamonEt2011.pdf [Accessed 10 Dec. 2019].

[2] Bird, S., Loper, E. and Klein, E. (2009). Natural Language Processing with Python. [online] Natural Language Toolkit. Available at: https://www.nltk.org/ [Accessed 10 Dec. 2019].

[3] Escalante, H., Solorio, T. and Montes-y-Gomez, M. (2011). Local Histograms of Character N-grams for Authorship Attribution. Proceedings of the 49th Annual

Meeting of the Association for Computational Linguistics: Human Language Technologies, 19-24, pp.288-298.

[4] Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D. and Gelbukh, A. (2016). Automatic Authorship Detection Using Textual Patterns Extracted from Integrated Syntactic Graphs. Sensors, 16(9), p.1374.

[5] Koppel, M. and Winter, Y. (2013). Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology, 65(1), pp.178-187.

[6] Lebanon, Mao, Y. and Dillon, J. (2007). The locally weighted bag of words framework for document representation. Journal of Machine Learning Research, 8, pp.2405–2441.

[7] Manning, Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. (2014). Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, [online] pp.55-60. Available at: https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf [Accessed 9 Dec. 2019].
bibitemPedregosa, F., Varoquaux, G. and Gramfort, A. (2011). Scikit-learn: Machine Learning in Python. JMLR 12, [online] pp.2825-2830. Available at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf [Accessed 9 Dec. 2019].

[8] Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of Age and Gender on Blogging. Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. [online] Available at: https://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-039.pdf [Accessed 10 Dec. 2019].

[9] Socher, R., Bauer, J. and Manning, C. (2019). Parsing with compositional vector grammars. Proceedings of ACL Conference 2013, 4–9, pp.455–465.

[10] Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), pp.538-556.