Authorship Attribution in Multi-author Documents

Tim Althoff, Denny Britz, Zifei Shan
Department of Computer Science, Stanford University
{althoff, dbritz, zifei}@cs.stanford.edu

Abstract

Authorship attribution, the task of identifying the author of a document, has been applied to works of historical value such as Shakespeare's plays or the political Federalist papers but is still highly significant today. We introduce the novel problem of authorship attribution in multi-authored documents and focus on scientific publications. Multi-authored documents present hard challenges for authorship attribution. We propose several ideas how these can be addressed and evaluate when such models perform well. To this end we present a sentence-based prediction model that also allows to estimate which sentences were contributed by which author. We demonstrate using a stylometric approach that paper authors can be predicted with significant accuracy by exploiting authors' stylistic idiosyncrasies. This challenges the assumption that simply removing names from a paper submission ensures anonymity in a double-blind process.

1 Introduction

Authorship attribution, the science of identifying the rightful author of a document, is a problem of long-standing history. The main idea behind statistically or computationally supported authorship attribution is that by measuring textual features, we can distinguish between texts written by different authors.

Identifying the author of a document also has modern applications such as identifying and linking users across online communities or detecting fraudulent transactions and impersonation attacks.

Thus far, work has focused on predicting authors of single-authored documents such as plays (Mendenhall, 1887; Matthews and Merriam, 1993; Merriam and Matthews, 1994), political essays (Mosteller and Wallace, 1964), or blog posts (Narayanan et al., 2012). In contrast, this paper introduces the problem of authorship attribution in multi-authored documents such as academic research papers. To the best of our knowledge, this problem has never been tackled before.

Many computer science conferences employ a double-blind submission process, relying on the assumption that identifying the authors of submitted papers is impossible at at least impractical. We challenge this notion by presenting an approach that is able to identify the authors of anonymous papers with significant accuracy by exploiting authors' stylistic idiosyncrasies.

Identifying the authors of multi-authored documents presents new challenges. Since it is unclear which author contributed which part

of a document, we lack ground truth that could be used to build a model for each author. Employing a document-level perspective as done in previous work may confuse idiosyncrasies of several authors (see Section 3.1). We propose employing a sentence-level perspective in which we predict an author for each individual sentence and then aggregate those to paper-level author predictions. We empirically evaluate all approaches on both synthetic and real world data and discuss where such techniques might fail.

2 Related Work

A large body of research exists on attributing authorship on the document level. The pioneering study of Mendenhall (Mendenhall, 1887) in the 19th century marks the first attempt to quantify writing style on the plays of Shakespeare, which later was followed by statistical studies by Yule (Yule, 1939; Yule, 1944) and Zipf (Zipf, 1932). Recent approaches can broadly be categorized into varying among the dimensions of feature selection, model selection, and candidate selection (Stamatatos, 2009). Features can be divided into lexical (token- and word features), character (character n-grams), syntactic (POS tags, phrase structure), semantic (synonymous and dependencies) and application-specific features (Stamatatos, 2009). More sophisticated features such as local histograms (Escalante et al., 2011) and grammatical errors (Koppel and Schler, 2003) have also been explored. Authorship attribution approaches taking into account only self-citations often perform well (Hill and Provost, 2003) compared to their supervised counterparts.

A variety of supervised and unsupervised Machine Learning methods have been applied to the problem of authorship attribution (Stamatatos, 2009). Simple similarity-based models (nearest-neighbors) also perform surprisingly well (Koppel et al., 2012) and often outperform more "sophisticated" supervised classifiers such as SVMs (Narayanan et al., 2012). The size of the candidate author set is another important dimension and only few researchers have applied attribution models to web-scale data (Narayanan et al., 2012).

Predicting authors in multi-authored scientific publications has been out of focus of the scientific community thus far (to be best of our knowledge). However, some work has looked at using scientific publications to predict gender (Sarawgi et al., 2011; Bergsma et al., 2012) and whether or not the publication was written by a native speaker or submitted to a workshop instead of a conference (Bergsma et al., 2012).

3 Approach

3.1 Author mixing

One key drawback of employing a paper-level perspective on authorship attribution is that it might fail to disentangle different stylistic influences and falsely attribute the paper to an incorrect author. Let A and B be the correct authors of a paper who have quite different styles of writing. Thus, A has a low probability of actually writing B's sentences and vice versa. Then, an incorrect author C who stylistically resembles both A and B in some way might actually have a higher probability of being a paper author than A or B. We coin this issue "author mixing".

3.2 Notation

Our corpus (P, A) consists of a set of papers $P = \{p_1, \ldots, p_N\}$ and a set of authors $A = \{a_1, \ldots, a_K\}$. Each paper p_i is divided into sentences $s_j^i \in p_i$ that we assume were written by a single author $\texttt{author}_\texttt{s}(s_j^i) = a_k$. While we assume that a sentence is written by a single author we do not actually observe author labels on sentence level. We only observe paper level

authors and assume that the sentence author is one of the paper authors. Mathematically, given paper level annotation $\texttt{author}_\texttt{p}(p_i) = \{a_1, a_7, a_9\}$ we assume

$$\forall s_j^i \in p_i : \texttt{author}_\texttt{s}(s_j^i) \in \texttt{author}_\texttt{p}(p_i)$$

Our goal is to learn to differentiate author styles from the ambiguous ground truth labels $\texttt{author}_\texttt{p}$ such that we can assign each sentence in a paper to a single author.

We represent this mapping from sentence to author as a matrix M that, for all sentences, contains the single author that is believed to have written that sentence, e.g. $M(s_j^i) = a_k$. The authors' individual styles are captured through parameters $\theta = (\theta_1, \ldots, \theta_K)$ that includes an independent parameter vector for each author $a_k \in A$.

3.3 Sentence authorship model

The core of our model describes how likely a given author a is to have generated a given sentence s, $p(a|s)$. We represent a sentence through features $F(s) \in \mathcal{R}^n$ (for more details on our features please refer to section 4.3). We use the following logit model:

$$\begin{aligned} p(a|s) = & \quad \text{logit}^{-1}(\langle \theta_a, F(s) \rangle) \\ = & \quad \frac{1}{1 + \exp(-\langle \theta_a, F(s) \rangle}) \end{aligned}$$

Note that this model formulation is equivalent to a log-linear model:

$$\log p(\texttt{author}_\texttt{s}(s) = a|s) = \langle \theta_a, F(s) \rangle - \log Z$$

where Z is a normalizing constant

$$Z = 1 + \exp(\langle \theta_a, F(s) \rangle).$$

Further note that we have K of these models, one for each author.

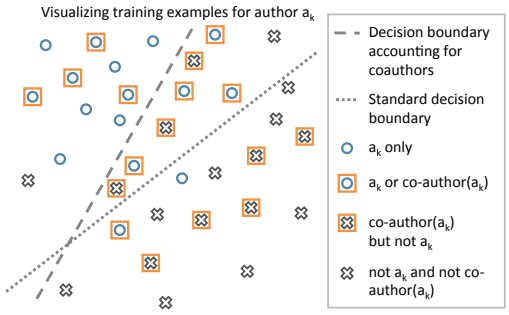

Figure 1: Training procedure considering co-author negative examples in addition to random negative examples

3.4 Learning to differentiate author styles from ambiguous labels

A core challenge in learning such logit models is that we lack sentence-level author labels. As described above, we only observe paper level author labels, i.e. $\texttt{author}_\texttt{s}(s_j^i) \in \texttt{author}_\texttt{p}(p_i)$ but we have no basis for assuming any concrete author yet.

Let the authors of a paper p_i be $\texttt{author}_\texttt{p}(p_i) = \{a_1, a_2\}$. To train a model for author a_1 we need both positive examples (sentences that a_1 did write) and negative example (sentences that a_1 did not write). For any sentence $s_j^i \in p_i$ we do not know whether it was written by a_1 or a_2, i.e. we have ambigious sentence labels. Negative examples are much easier to come by, sentences from any paper that a_1 did not co-author would provide true negative examples.

Since we have no basis for assuming, a priori, that sentence s_j^i was definitly written by either a_1 or a_2 we include all sentences that a_1 might have written as positive examples when training an author model for a_1. Obviously, this will include false positive examples that a_2 actually wrote and we need to make sure that the model does not only capture the combined style

of a_1 and a_2. To this end, we propose to carefully select negative examples that differentiate a_1's style from that of a_2. We can achieve this by including positive examples from a_2 in as negative examples for a_1. In fact, we do not only include but positive examples from all coauthors of a_1. This idea is visualized in Figure 1.

In the following, we describe this idea more formally. We denote the set of all coauthors of a given author a_k by

$$\texttt{coauthor}(a_k) = \{a_l \in A \mid \forall p \in P : a_l \in \texttt{author}_\texttt{p}(p) \Rightarrow a_k \in author_p(p)\}$$

Formally, we use the following set of positive examples for a_k:

$$\texttt{POS}(a_k) = \{s_j^i \in p_i \mid a_k \in \texttt{author}_\texttt{p}(p_i)\}$$

And the following set of negative examples for a_k:

$$\texttt{NEG}(a_k) = \{s_j^i \in p_i \mid \exists a_l \in \texttt{coauthor}(a_k) : a_l \in \texttt{author}_\texttt{p}(p_i) \wedge a_k \notin \texttt{author}_\texttt{p}(p_i)\}$$

We further add random sentences to $\texttt{NEG}(a_k)$ since authors have a varying number of coauthors and those with few would end up with very few negative examples otherwise.

3.5 Refining author models through Expectation Maximization

Modeling authorship on sentence level gives us the opportunity to further refine our author models (this is not possible on paper level). The approach described above uses all sentences that could possibly have been written by author a_k as positive examples (recall $\texttt{POS}(a_k)$). However, this will include many sentences that were actually written by one of a_k's coauthors. Based on our confidence about which sentences were likely written by a_k we can filter the positive examples to yield a cleaner set of positive examples for a_k. We now formalize this intuition.

For the likelihood of the full corpus (P, A) we treat each sentence independently:

$$P(M, \theta | P, A) = \prod_{p_i \in P} \prod_{s_j^i \in p_i} p(M(s_j^i) | s_j^i).$$

The parameter inference problem then becomes

$$\hat{M}, \hat{\theta} = \text{argmax}_{M,\theta}\, P(M, \theta | P, A) - \Omega(M, \theta),$$

where $\Omega(M, \theta)$ is a regularizer on the parameters to avoid overfitting.

Since this optimization depends on both M and θ we proceed by coordinate ascent on (M, θ), i.e. by alternately optimizing

$$M^i = \text{argmax}_M\, P(M, \theta^i | A, P)$$

and

$$\theta^{i+1} = \text{argmax}_\theta\, P(M^i, \theta | A, P)$$

until convergence, i.e. until M^i differs from M^{i-1} on less than a prespecified number of sentences (alternatively, one can alternate for a constant number of iterations).

Being a local optimization procedure, coordinate ascent is sensitive to initialization. Therefore, we initialize the style parameters for each author θ^0 by training independent logistic regression models using the procedure described in Section 3.4.

Optimizing for M then is a simple inference step where we dependently assign each sentence to the most likely possible author:

$$M(s_j^i) = \text{argmax}_a\, p(a | s_j^i, \theta)$$

Based on this assignment, optimizing for θ then becomes estimating K independent logis-

	Synthetic dataset	Real-world dataset
Authors	108	100
Publications	360	234
Sentences	78,577	206,300

Table 1: Descriptive statistics of the datasets

tic regression models for all authors based on the assignment M.

3.6 Predicting paper authors from sentence authors

Our proposed model gives us predictions on sentence-level $M(s_j^i)$. While those sentence-level predictions are interesting in its own right (e.g. to estimate contribution of different authors) we ultimately want to predict authors on paper-level. To this end, we aggregate our sentence-level predictions to paper level predictions by having each sentence s_j^i vote for its most likely author (namely, $M(s_j^i)$).

Compared to this hard voting scheme we also experimented with a soft voting scheme where each author gets a fractional vote depending on their confidence to have written this particular sentence. However, empirically we found that soft-voting across all sentences of a paper suffers from the same problems that paper-level predictions do (see Section 3.1). Hard voting performed best in most cases.

4 Experimental Setup

We evaluate our hypothesis on both synthetic and real-world data. We obtained scientific publications available through arXiv. The complete set of PDF documents includes about 700,000 publications. We converted all PDF documents into raw text files using Apache Tika[1]. For documents in either one of the two datasets we produced sentence tokenizations and annotations using Stanford CoreNLP[2] on the converted text files. The datasets are summarized in Table 1.

4.1 Synthetic data

In order to evaluate our sentence-level predictions we generated an synthetic dataset with sentence-level labels as follows. Let A be the set of authors such that every author $a \in A$ has more than 10 single-authored papers and more than 120 paragraphs with at least 500 words. Let P_a be the set of paragraphs of more than 500 words written by $a \in A$. We generate a document by randomly picking 3 authors $a_1, a_2, a_3 \in A$ and sampling 40 paragraphs from $P_{a_1}, P_{a_2}, P_{a_3}$ without replacement. We repeat this procedure until not enough paragraphs are left to generate a full document. This procedure yields a corpus of 108 authors, 360 publications, 14,027 paragraphs and 78,577 sentences. Each author gets approximately the same number of publications and paragraphs.

4.2 Real-world data

To test our model in the real world, we subsample the arXiv dataset. We are interested in distinguishing authors in similar fields, therefore we start from a list of authors ("bootstrap list") working in the field of social and information networks.[3] We sample a set of papers written by any of these authors and their coauthors, controlling each author to have at most 10 papers. We get a corpus of 234 papers, 100 authors (the 8 authors in the "bootstrap list" and their 92 coauthors) and 206,300 sentences.

[1] http://tika.apache.org/

[2] http://nlp.stanford.edu/software/corenlp.shtml

[3] These authors include: Alex Pentland, Bernardo Huberman, Brian Karrer, Johan Ugander, Jon Kleinberg, Jure Leskovec, Lada Adamic, and Lars Backstrom (sorted by first name).

4.3 Features

The parameters of our models correspond to features categorized into content and style. Table 2 contains a detailed description of the feature space we are considering. We tried to make a clear distinction between content and style features, but we acknowledge that some features can be regarded as capturing both content and style (e.g. character n-grams).

5 Results

In addition to the two models presented in section 3 we evaluate a paper-level logistic regression classifier using the same feature set as described above. We also present results for majority-class baselines.

5.1 Evaluation metrics

We evaluate paper-level predictions using precision@1 and average precision (Manning et al., 2008) . Precision@1 is defined as the fraction of papers where the most likely author (ranked highest according to our classifier) is one of the correct co-authors of the paper. Average Precision (Manning et al., 2008) complements precision@1 in that it captures correct predictions beyond the highest-ranked author. We also provide numbers for sentence-level precision evaluated on the synthetic dataset (sentence-level ground truth for real-world data is not available).

5.2 Results

Tables 3 and 4 show our results. We can see that aggregated predictions outperform paper-level predictions in some cases. This is mainly due to the mixing issue discussed in section 3.1. All our models use L2 regularization with the regularization parameter optimized using 5-fold cross validation. It is interesting to note that even though sentence-level precision is relatively low, aggregating the sentence predictions leads to good paper-level predictions.

5.3 Discussion

This section discusses our finding and specifically investigates why sentence-level models do not outperform paper-level models in all cases.

First, we recognize basic assumptions of our models: We assume everyone writes something (). Obviously this might not hold in the real-world since all authors might do something but not all might contribute writing. Given that we assign all sentences in a paper as positive examples to each paper author, in a way, we assume uniform distribution of author contributions. When lacking good negative examples we might fail to reject many sentences for an author thereby overestimating their contribution. Future work should make use of scientific publications with annotation on sentence- or paragraph-level to investigate this.

Second, we assume that each sentence is written by a single author but recognize that multiple authors could have contributed to its writing. Our toy dataset was generated from single-author papers that presumably were written by its single author but this assumption might not hold on real-world papers.

Third, while the sentence-level perspective has clear advantages (see Section 3.1) we must acknowledge some limitations of the proposed approach. Since we aggregate sentence-level predictions by voting to paper-level predictions, even when a sentence gives away authorship (e.g. by an obvious self-citation) this would only contribute a single vote among many. In contrast, such features could very strongly influence paper-level predictions. Future work could investigate hybrid approaches that aim at the best of both worlds.

Fourth, assigning single authors to sentences and aggregating these hard votes to paper-level predictions could be improved. For instance,

Type	Feature Name	Description	Examples
Content	Word N-grams	Unigrams, bigrams, and trigrams of all words.	"a distance matrix", "quartet topologies embedded"
	Character N-grams	Bigrams and trigrams of characters.	"NWD", "Ref"
Stylistic	Character unigrams	Unigrams of characters.	'%', '{', '?', 'α'
	POS tag N-grams	Unigrams and bigrams of POS tags (tagged by NLTK default Penn Treebank POS tagger).	DT NN, PRP VBP, -LRB- NN
	Function word N-grams	unigrams, bigrams, and trigrams of function words.	"under our", "but all the"
	Lengths	Sentence length and average word length of each sentence.	81 chars, 17 words, average word length=4.7
	Sentence start	First word in sentence if it is a function word.	"We", "It", "However"
	Transition words	First word after a punctuation if it is a function word.	"However, we observe" would give the feature "we".
	Punctuation sequence	Concatenation of all the punctuations in a sentence.	".", ",.", ",;,.", ",()."
	Sentence shape	Punctuation sequence and number of words ("M" if more than 3) between each two punctuations.	"3,M;2,M."

Table 2: Features

	Majority class	Paper-level	LR sentence	EM sentence
Precision@1	0.04	0.93	0.97	**0.99**
Average Precision	0.04	0.69	0.85	**0.86**
Sentence-level precision	0.04	-	0.17	**0.20**

Table 3: Experimental results on the synthetic dataset

one might one want to use high-confidence votes. In our empirical evaluations we find that this does not lead to significant performance improvements. For most sentences many authors will have a high likelihood since most sentences do not contain very unique and discriminative features. Based on more common features, e.g. character n-grams, many authors will be assigned a high-likelihood. One could address this by casting a vote if and only if the first ranked author is significantly more likely than the second ranked author. We also ran experiments in which we assigned a sentence to multiple authors. While this leads to more stability across EM iterations we do not always find significant performance improvements over the paper-level model (on the real-world dataset).

	Precision	Average Precision
Style	0.55	0.36
Content	0.89	0.68
All	**0.93**	**0.69**

Table 5: Precision of paper-based models with varying set of features

	Majority class	Paper-level	LR sentence	EM sentence
Precision@1	0.02	**0.35**	0.30	0.27
Average Precision	0.04	**0.41**	0.34	0.31
Sentence-level precision	-	-	-	-

Table 4: Experimental results on the real-world dataset

5.4 Feature variations

We also explored how varying the set of features affects prediction accuracy. For the paper-level model we varied the feature set to use content-only, style-only and a combination of both features (all). The results are shown in Table 5. We see that content-features account for most of the performance of the model. As expected, style features alone have less predictive performance. However, we note that the performance of pure style features is still quite remarkable and that adding them to content features can give another boost (all features).

6 Future Work

Future work should investigate the impact of individual features on the predictive performane of our author models. We also hypothesize that higher-level knowledge such as citations, co-authorship correlation, author order, sentence sequence order, and domain knowledge in scientific writing is likely to improve performance significantly. We performed initial experiments using a Conditional Random Field (CRF) model with promising results.

To better understand the performance gap between paper-level and sentence-level models on the real-world dataset, error analysis using a real-world dataset with ground truth annotation on sentence level should be performed.

Another line for future work is to investigate the impact of author pool size to the task: how many papers for each author do we need to train a good model for an author? How does increasing the number of authors affect the difficulty for prediction?

7 Conclusion

In this paper, we introduce the novel problem of authorship attribution in multi-authored documents, and focus on scientific publications.

A core challenge in this setting is the lack of ground truth of which authors write which parts of the publication. To address this challenge, we propose and evaluate different models including (1) paper-level logistic regression, (2) sentence-level logistic regression models along with different aggregation frameworks, and (3) an Expectation-Maximization technique to further refine author models. We also propose a novel supervision method to obtain training examples in the presence of ambiguous authorship labels.

We devise a set of stylistic features as well as content-based features, and find that paper authors can be predicted with significant accuracy by exploiting authors' stylistic idiosyncrasies.

Based on our evaluation on a synthetic dataset as well as a real-world dataset, we conclude several results: (1) in some cases sentence-level models can lead to significant performance improvements. (2) We are able to identify authors with high accuracy demonstrating a tenfold improvement over the majority class baseline.

Our results challenge the notion that simply removing names from a paper submission ensures anonymity in a double-blind submission process and hope that this work can serve as a basis for future research on authorship attribution in multi-authored documents.

References

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337. Association for Computational Linguistics.

Hugo Jair Escalante, Thamar Solorio, and Manuel Montes-y Gómez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 288–298. Association for Computational Linguistics.

Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review?: author identification using only citations. *ACM SIGKDD Explorations Newsletter*, 5(2):179–184.

Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, pages 72–80. Citeseer.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Yaron Winter. 2012. The fundamental problem of authorship attribution. *English Studies*, 93(3):284–291.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Robert AJ Matthews and Thomas VN Merriam. 1993. Neural computation in stylometry i: An application to the works of shakespeare and fletcher. *Literary and Linguistic Computing*, 8(4):203–209.

Thomas Corwin Mendenhall. 1887. The characteristic curves of composition. *Science*, (214S):237–246.

Thomas VN Merriam and Robert AJ Matthews. 1994. Neural computation in stylometry ii: An application to the works of shakespeare and marlowe. *Literary and Linguistic Computing*, 9(1):1–6.

Frederick Mosteller and David Wallace. 1964. Inference and disputed authorship: The federalist.

Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the feasibility of internet-scale author identification. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 300–314. IEEE.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86.

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

G Udny Yule. 1939. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3-4):363–390.

George Udny Yule. 1944. *The statistical study of literary vocabulary*. CUP Archive.

George Kingsley Zipf. 1932. Selected studies of the principle of relative frequency in language.