Deep Sentence-Level Authorship Attribution

Stephen Macke
Department of Computer Science
Stanford University
smacke@cs.stanford.edu

Jason Hirshman
Department of Mathematics
Stanford University
hirshman@stanford.edu

Abstract

We examine the problem of authorship attribution in collaborative documents. We seek to develop new deep learning models tailored to this task. We have curated a novel dataset by parsing Wikipedia's edit history, which we use to demonstrate the feasiblity of deep models to multi-author attribution at the sentence-level. Though we attempt to formulate models which learn stylometric features based on both grammatical structure and vocabulary, our error analysis suggests that our models mostly learn to recognize vocabulary-based cues, making them non-competitive with baselines tailored to vocabulary-based features. We explore why this may be, and suggest directions for future models to mitigate this shortcoming.

1 Introduction

We seek to apply deep learning methods to attribute portions of a collaborative document to individual authors. In particular, we are working with a portion of Wikipedia's edit history to determine authorship of multi-author articles.

Machine learning has been applied to the problem of author identification in the past, but to our knowledge it has not been applied in the case of collaborative documents. When multiple authors contribute to a single piece of writing, they may interact in complex ways, meaning that traditional approaches based on per-document statistics will not work. We hypothesize that a deep learning approach may allow a model to learn stylistic features of authors at a finer granularity, which may in turn be used to determine document authorship with some level of confidence.

We parse the edit history of selected Wikipedia pages in order to isolate sentence-level edits attributable to particular authors. We then attempt to create models that classify sentences as belonging to a given author.

Of the deep models we implemented, we found recursive neural networks to be most effective at distinguishing authorship. Unfortunately, as we increased the number of authors, we found that it was not competitive with a very simple multinomial Naive Bayes baseline. In the subsequent sections, we present these results and explore the data to determine why this might be the case. We also outline a possible extension where a sentence's context or place within the document is included in the model.

1.1 Formal Problem Statement

Given a document D consisting of sentences $S = \{s_i\}$ written by authors $A = \{a_j\}$, we wish to recover the function $f : S \to A$ that maps each sentence to the author that wrote it. We will learn f by observing a series of example sentences in other documents from those same authors. In other words we have a set of ordered pairs (s_k, a_k) where $a_k \in A$ but $s_k \notin S$.

The documents come from English Wikipedia where over 25 million people contribute to over 4 million articles. To make the problem manageable, we will restrict A to a small set of prolific authors, $|A| = 10$. Our estimate for f, $\hat{f}$, will map to a set of probabilities that represent the proportion of the sentence that the model believes to be attributable to each author. So, $\hat{f} : S \to [0, 1]^{|A|}$.

We will evaluate our success based on the proportion of sentences we correctly attribute. We simply consider the author with the highest predicted probability to be the one predicted by the classifier. A simple accuracy can then be computed. As random guessing would result in an accuracy of $\frac{1}{|A|}$, we hope to do significanly better.

2 Related Work

Authorship attribution is a well-studied task in natural language processing. The classic problem involves determining who wrote the unclaimed Federalist Papers. Jockers and Witten survey the various ways machine learning has been applied to this task [2] [3]. In each of the methods, the models rely on hand-coded vocabulary and stylometric features. Deep learning moves beyond these hand coded features and allows for a more flexible model. Neural networks were actually applied to this problem back in 1996 [7], but those networks were shallow and could not leverage new natural language processing techniques in deep learning [1].

Additionally, prior work on authorship attribution is mostly concerned with document-level models for single-author documents, as opposed to our sentence-level formulation for multi-author documents [4]. Typically, this work relies on aggregate statistics from the entire document pending classification. Our formulation distinguishes itself in that aggregate statistics from previously-unseen documents are not as useful, as the document may have been generated by several different authors.

3 Approach

3.1 Data Collection Algorithm

The actual process of obtaining sentence-level authorship information from the Wikipedia edit history is nontrivial. When a user makes an edit to an article, the resulting revision appears to be entirely the work of that user, from the perspective of Wikipedia. The raw data has no notion of a diff. To further complicate things, a user could, for example, reorder paragraphs without providing novel article content. A simple diff tool will fail to recognize that article A' which is a permutation of the paragraphs of article A is effectively the "same" article from our perspective.

To remedy this, we built a heuristic tool which, given a revision and current sentence-level author information, builds an in-memory inverted index of the revision, mapping trigrams to sentences. Sentences from subsequent revisions are scored against existing sentences by Jaccard similarity on trigram sets. If a new revision's score s falls below some threshold, the sentence is considered "new" and fully attributed to the new revision's author. Otherwise, the new author is given $1 - s$ "ownership" of the sentence (s would be 1 in the case where the author is solely permuting paragraphs). This processed is summarized graphically in figures 1 and 2.

As an example, in our data user "Sam Francis" wrote the following: `All anarchists have a fundamental critique of government, a vision of a society without government, and a proposed method of reaching such a society.`

Later, user "Miguil enwiki" revised this to: `Anarchist theories have a fundamental critique of government , a vision of a society without government , and a proposed method of reaching such a society.`

Our algorithm scored the updated sentence as follows:

```
{
    "Sam Francis": 0.8679245283018866,
    "Miguel˜enwiki": 0.13207547169811326
}
```

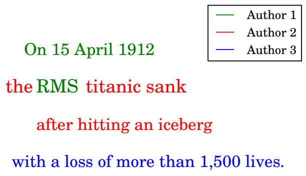

Figure 1: To determine ground-truth authorship attribution from revision history, we build an in-memory inverted index of each article (mapping ngrams to sentences) and score sentences in subsequent revisions against those seen so far.

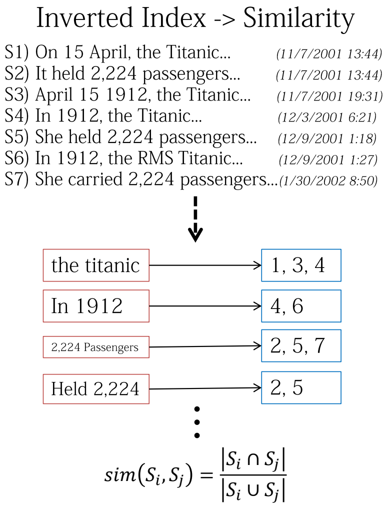
Inverted Index -> Similarity
S1) On 15 April, the Titanic... (11/7/2001 13:44)
S2) It held 2,224 passengers... (11/7/2001 13:44)
S3) April 15 1912, the Titanic... (11/7/2001 19:31)
S4) In 1912, the Titanic... (12/3/2001 6:21)
S5) She held 2,224 passengers... (12/9/2001 1:18)
S6) In 1912, the RMS Titanic... (12/9/2001 1:27)
S7) She carried 2,224 passengers... (1/30/2002 8:50)
the titanic
1, 3, 4
In 1912
4, 6
2,224 Passengers
2, 5, 7
Held 2,224
2, 5
$sim(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$

Figure 2: Illustration of ngram inverted index for sentence-level attribution.

Our attribution algorithm allows us to filter out sentences with ambiguous or unknown authors. In total, we use the above methods to isolate sentences from 10 prolific authors in order to build our authorship treebank. A few complications made it impossible to scale to larger numbers of authors:

1. Wikipedia authorship follows a heavy-tailed distribution – only a very few authors are prolific enough to generate data in the quantities that we require (hundreds to thousands of sentences per author), and any given subset of Wikipedia is likely to have only a very small fraction of these, which means we must tailor our scraping tools to target specific known prolific authors.
2. Of the prolific authors, a large number do not generate actual content, but instead add references, move content around, and police vandalism. As such, for each author we needed to perform a manual audit to ensure we only use original content from that author (content not copied from elsewhere), as our collection procedure was not capable of performing attribution accurately across separate articles.

Table 5 (in the appendix) shows the per-author sentence counts, as well as the train/dev/test proportions.

3.2 Models

To approximate $f : S \to A$, we found that the most effective deep learning method was an extremely simple recursive neural network with a single ReLU nonlinearity at each node, and a single softmax output only at the root node. Figure 3 illustrates this model. Our formulation of a recursive neural network is slightly different because only one prediction is made at the root node as opposed to predicting at each branch of the tree. Because the desired output did not vary across the sentence nor does predicting authorship from a single word make sense, predicting at each branch resulted in overfitting the training sentences.

We found that additional softmax outputs at non-root nodes, which have the interpretation of attributing authorship to grammatical substructures, ended up hurting performance as they tended to encourage overfitting. Additionally, we did not see significant improvements from tree-structured LSTM networks or tree-structured neural tensor networks, as described in [6] and [5], respectively.

We also attempted to train a recurrent-LSTM model both on fixed-length windows of sentences and by pooling together vectors computed across the sentences. However, we repeatedly ran into vanishing gradient issues, and when we could get a model to be trained, the accuracy was very poor.

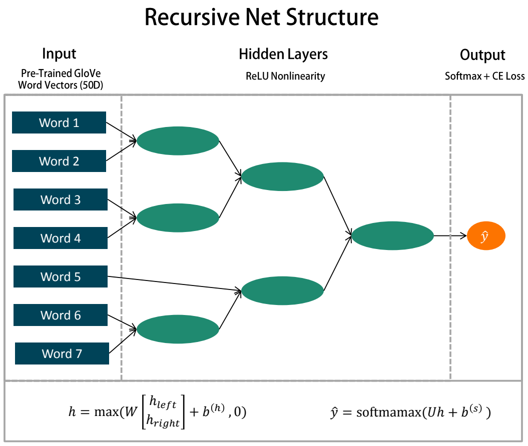
Recursive Net Structure
Input
Pre-Trained GloVe Word Vectors (50D)
Hidden Layers
ReLU Nonlinearity
Output
Softmax + CE Loss
Word 1
Word 2
Word 3
Word 4
Word 5
Word 6
Word 7
$\hat{y}$
$h = \max(W \begin{bmatrix} h_{left} \\ h_{right} \end{bmatrix} + b^{(h)}, 0)$
$\hat{y} = \text{softmamax}(Uh + b^{(s)})$

Figure 3: Illustration of the simple recursive model.

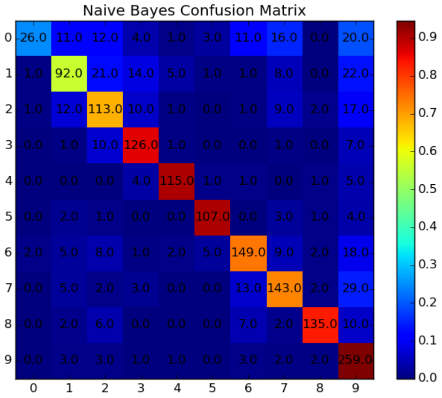

Figure 4: Confusion matrix for the multinomial Naive Bayes baseline.

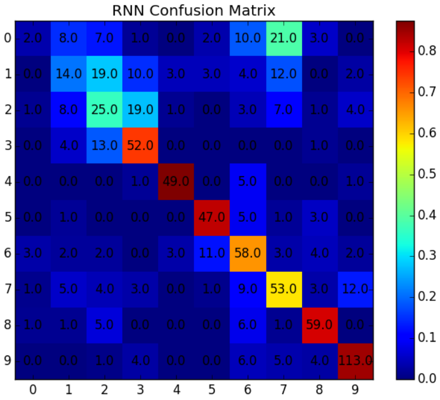

Figure 5: Confusion matrix for the recursive neural network.

4 Experiments

As mentioned in section 3, we collected and parsed sentences for 10 prolific authors. In total, we have 6,641 train sentences, along with 825 dev sentences and 834 test sentences for hyperparameter tuning and cross validation. Our results for three authors and a subset of our data are summarized in tables 1 and 2, and our results for all authors are summarized in tables 3 and 4.

Additionally, see figures 4 and 5 for a visualization of the confusion matrices for each model on the test set.

	Precision	**Recall**	**F1**
Train	0.97	0.97	0.97
Test	0.82	0.81	0.81

Table 1: Results for the Naive Bayes sentence-level authorship classifier, 3 authors.

	Precision	**Recall**	**F1**
Train	0.95467	0.95460	0.95463
Dev	0.81985	0.82178	0.82025
Test	0.84059	0.84135	0.84016

Table 2: Results for the RNN sentence-level authorship classifier, 3 authors.

	Accuracy
Train	0.9242697983
Test	0.7652694611

Table 3: Results for the Naive Bayes sentence-level authorship classifier, 10 authors.

	Accuracy
Train	0.674770
Dev	0.629530
Test	0.622691

Table 4: Results for the RNN sentence-level authorship classifier, 10 authors.

4.1 Discussion

Although the recursive neural network outperformed the Naive Bayes baseline marginally on the 3-author subset, its accuracies diminish much more rapidly as the number of authors increases. We believe that this due to one or both of the following reasons:

1. The recursive model is not expressive enough to fully capture stylometric nuances latent grammatical structures of sentences.
2. There is simply not enough variation in grammar and sentence parses between different authors to distinguish them, at least in a medium such as Wikipedia.

Indeed, most of the errors in the neural network appear to stem from overfitting on vocabulary-based features. For example, consider the folowing sentence:

```
(Parkwells:  ( ( ( ( ( The) ( ( ``) ( ( Oriental) ( ( '') (
alabaster))))) ( ( was) ( ( highly) ( ( esteemed) ( ( for) (
( making) ( ( ( small) ( ( perfume) ( ( bottles) ( ( or) ( (
ointment) ( vases)))))) ( ( called) ( alabastra))))))))) ( ;))
( ( ( the) ( ( vessel) ( name))) ( ( has) ( ( been) ( ( suggested)
( ( as) ( ( ( a) ( ( possible) ( source))) ( ( of) ( ( the) ( (
mineral) ( name))))))))))) ( .))
```

The model incorrectly attributes this sentence to "Materialscientist", likely because of the presence of the words "mineral" and "alabaster".

5 Conclusion

We found the deep architectures were generally outperformed by simple baselines. We believe this has two possible causes: either the models were unable to capture stylometric differences due to grammatical structures, or grammatical structures between authors in a medium such as Wikipedia are mostly indistinguishable. One possible future model inspired form this observation would train seperate units for each possible grammatical substructure (NP, PP, etc.). In the end, the models mainly learned to recognize differences in vocabulary.

It is interesting to see that the recursive neural network came close. The neural network has many fewer parameters than the naive bayes model since naive bayes has a parameter for each word in the vocabulary. With the fewer parameters, it is incapable of simply memorizing the words used by These observations also prompt a call for a future experiment where authors all must write about the same topic and then see whether the deep learning models can outperform classical stylometric techniques.

References

[1] F. A. Gers and J. Schmidhuber. Lstm recurrent networks learn simple context-free and context-sensitive languages. *Neural Networks, IEEE Transactions on*, 12(6):1333–1340, 2001.

[2] M. L. Jockers and D. M. Witten. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, page fqq001, 2010.

[3] V. Kešelj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264, 2003.

[4] M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[5] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[6] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[7] F. J. Tweedie, S. Singh, and D. I. Holmes. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30(1):1–10, 1996.

6 Appendix

Table 5: Number of train, dev, and test sentences for each author.

	Number of Samples		
Authors	**Train**	**Dev**	**Test**
Wikidea	1110	136	138
MaterialScientist	779	98	103
Parkwells	673	85	79
Rjensen	672	84	82
Jmabel	732	96	101
Iss246	695	83	79
JustinTime55	551	67	60
Wehwalt	533	70	76
Cwmhiraeth	513	56	62
Mannanan51	383	50	54

Table 6: Most frequent words by author.

	Wikidea	**MaterialScientist**	**Parkwells**	**Rjensen**	**Jmabel**	**Iss246**	**JustinTime55**	**Wehwalt**	**Cwmhiraeth**
1)	company	americium	state	british	andalusia	health	apollo	johnson	species
2)	act	retrieved	alabama	new	andalusian	psychology	lunar	speer	amphibians
3)	law	actinium	war	lincoln	spain	research	mission	president	frogs
4)	directors	isbn	jackson	american	new	work	module	hitler	salamanders
5)	health	nuclear	american	states	times	occupational	crew	tennessee	eggs
6)	rights	used	states	history	spanish	job	spacecraft	congress	water
7)	shareholders	actinides	people	army	region	journal	nasa	lincoln	body
8)	companies	alkaloids	county	union	york	psychologists	flight	senate	skin
9)	care	compounds	new	state	anadalucia	school	orbit	war	frog
10)	employees	uranium	british	confederate	percent	psychological	earth	party	caecilians