Unsupervised Authorial Clustering Based on Syntactic Structure

Alon Daks and **Aidan Clark**
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
{alon.daks, aidanbclark}@berkeley.edu

Abstract

This paper proposes a new unsupervised technique for clustering a collection of documents written by distinct individuals into authorial components. We highlight the importance of utilizing syntactic structure to cluster documents by author, and demonstrate experimental results that show the method we outline performs on par with state-of-the-art techniques. Additionally, we argue that this feature set outperforms previous methods in cases where authors consciously emulate each other's style or are otherwise rhetorically similar.

1 Introduction

Unsupervised authorial clustering is the process of partitioning n documents written by k distinct authors into k groups of documents segmented by authorship. Nothing is assumed about each document except that it was written by a single author. Koppel et al. (2011) formulated this problem in a paper focused on clustering five books from the Hebrew Bible. They also consider a 'multi-author document' version of the problem: decomposing sentences from a single composite document generated by merging randomly sampled chunks of text from k authors. Akiva and Koppel (2013) followed that work with an expanded method, and Aldebei et al. (2015) have since presented an improved technique in the 'multi-author document' context by exploiting posterior probabilities of a Naive-Bayesian Model. We consider only the case of clustering n documents written by k authors because we believe that, in most cases of authorial decomposition, there is some minimum size of text (a 'document'), for which it can be reliably asserted that only a single author is present. Furthermore, this formulation precludes results dependent on a random document generation procedure.

In this paper, we argue that the biblical clustering done by Koppel et al. (2011) and by Aldebei et al. (2015) do not represent a grouping around true authorship within the Bible, but rather around common topics or shared style. We demonstrate a general technique that can accurately discern multiple authors contained within the Books of Ezekiel and Jeremiah. Prior work assumes that each prophetic book reflects a single source, and does not consider the consensus among modern biblical scholars that the books of Ezekiel and Jeremiah were written by multiple individuals.

To cluster documents by true authorship, we propose that considering part-of-speech (POS) n-grams as features most distinctly identifies an individual writer. The use of syntactic structure in authorial research has been studied before. Baayen et al. (1996) introduced syntactic information measures for authorship attribution and Stamatatos (2009) argued that POS information could reflect a more reliable authorial fingerprint than lexical information. Both Zheng et al. (2006) and Layton et al. (2013) propose that syntactic feature sets are reliable predictors for authorial attribution, and Tschuggnall and Specht (2014) demonstrates, with modest success, authorial decomposition using pq-grams extracted from sentences' syntax trees. We found that by combining the feature set of POS n-grams with a clustering approach similar to the one presented by Akiva (2013), our method of decomposition attains higher accuracy than Tschuggnall's method, which also considers grammatical style. Additionally, in cases where authors are rhetorically similar, our framework outperforms techniques outlined by Akiva (2013) and Aldebei (2015), which both rely on word occurrences as features.

This paper is organized as follows: section 2 outlines our proposed framework, section 3 clari-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 114–118,
Berlin, Germany, August 7-12, 2016.

fies our method in detail through an example, section 4 contains results, section 5 tests an explanation of our results, and section 6 concludes our findings and discusses future work.

2 Our Framework

Given n documents written by k distinct authors, where it is assumed that each document is written entirely by one of the k authors, our method proceeds in the following way:

First, represent each document as a frequency vector reflecting all n-grams occurring in the 'POS-translated' document.

Second, cluster documents into k groups using an unsupervised clustering algorithm.

Third, determine 'core elements', documents that most strongly represent authorship attributes of their respective clusters.

Fourth, use 'core elements' to train a supervised classifier in order to improve accuracies of documents that were not central to any cluster.

A key improvement our framework presents over prior techniques is in step one, where we represent documents in terms of POS n-grams. Specifically, each document, x_i, is transformed into a 'POS-translated' version, x_i', such that every word or punctuation symbol from the original document is replaced with its respective POS or punctuation token in the translated version. Consider the following sentences from a *New York Times* (NYT) column written by Paul Krugman: "Last week the Federal Reserve chose not to raise interest rates. It was the right decision." In the 'POS-translated' version these sentences appear as "JJ NN DT NNP NNP NN RB TO VB NN NNS PERIOD PRP VBD DT JJ NN PERIOD".[1] We use a POS tagger from the Natural Language Toolkit to translate English documents (Bird et al., 2009) and use hand annotations for the Hebrew Bible. Our framework will work with any text for which POS-annotations are obtainable. The requirement that k is a fixed parameter is a limitation of the set of unsupervised clustering algorithms available in step two.

3 Clarifying Details with NYT Columns

We shall describe a clustering of *New York Times* columns to clarify our framework. The NYT corpus is used both because the author of each document is known with certainty and because it is a canonical dataset that has served as a benchmark for both Akiva and Koppel (2013) and Aldebei et al. (2015). The corpus is comprised of texts from four columnists: Gail Collins (274 documents), Maureen Dowd (298 documents), Thomas Friedman (279 documents), and Paul Krugman (331 documents). Each document is approximately the same length and the columnists discuss a variety of topics. Here we consider the binary $(k = 2)$ case of clustering the set of 629 Dowd and Krugman documents into two groups.

Authors	1st	2nd	3rd
TF-PK	4 - 4	5 - 5	3 - 4
TF-GC	**3 - 5**	3 - 4	4 - 4
TF-MD	5 - 5	3 - 4	**3 - 5**
GC-PK	4 - 4	**3 - 5**	3 - 4
MD-PK	**3 - 5**	3 - 4	4 - 4
GC-MD	**3 - 5**	3 - 4	4 - 4

Table 1: The top three ranges for n-grams by F1 accuracy for each two-way split of NYT columnists. Here, TF = Thomas Friedman, GC = Gail Collins, MD = Maureen Dowd, PK = Paul Krugman.

In step one, the documents are converted into their 'POS-translated' form as previously outlined. Each document is represented as a frequency vector that reflects all 3, 4, and 5-grams that appear in the 'POS-translated' corpus. This range of n-grams was determined through validation of different values for n across several datasets. Results of this validation for the two way split over NYT columnists is displayed in Table 1. These results are consistent when validating against other datasets. Using 3, 4, and 5-grams, the resulting design matrix has dimension 629 by 302,395. We re-weight every element in the design matrix according to its term frequency–inverse document frequency.

In step two, we apply spectral clustering to the design matrix to partition the documents into two clusters. This is implemented with the Shi and Malik (2000) algorithm, which solves a convex relaxation of the normalized cuts problem on the affinity graph (Pedregosa et al., 2011). Edge-weights of the affinity graph are computed using a linear kernel. In the case of clustering several $(k > 2)$ authors, we apply the Yu and Shi (2003) algorithm to perform multiclass spectral clustering.

In step three, we calculate the centroid of each cluster produced by step two. For each document

[1] A list of POS tags and explanations:
http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Columnist	Cluster I	Cluster II
Dowd	**294**	4
Krugman	3	**328**

Table 2: Results when clustering 629 documents written by Maureen Dowd and Paul Krugman into two clusters.

x_i', we determine θ_i, the angle between that document and the centroid of its cluster, and call a document a 'core element' if θ_i is within 2 standard deviations of the average of θ_i in x_i''s cluster.

In step four, 'core elements' are used to train a 500 tree random forest where at each split the standard heuristic of $\sqrt{p}$ features are considered (here $p = 302,395$). Finally, we reclassify all 629 documents according to this random forest to produce our final class labels, summarized in Table 2. The final accuracy of the Dowd-Krugman clustering, measured as an F1-score, is 98.8%.

4 Results

All accuracy scores given in the rest of this paper are calculated using the F1-score. Because our technique contains stochastic elements, results reflect an average of 20 runs.

4.1 NYT Columns

When clustering over all six binary-pairs of NYT columnists, our framework achieves an average accuracy of 94.5%, ranging from 90.0% to 98.8%. Aldebei et al. (2015) addresses the slightly different problem of decomposing artificially merged NYT documents, and acknowledging the distinction between the two problems, our results are comparable to their accuracies which range from 93.3% to 96.1%.

4.2 *Sanditon*: An Uncompleted Novel

Another canonical authorship test is that of the novel *Sanditon*, a text left incomplete at the death of its author, Jane Austen, and finished some years later by an anonymous person known as "Another Lady." She closely emulated Austen's style and added 19 chapters to Austen's original 11. Researchers have long examined this text and most recently Moon et al. (2006) analyzed *Sanditon* using supervised techniques in the context of authorship attribution. Much progress has been made in the field since then, but examining *Sanditon* has fallen out of style. Our framework clusters Austen's chapters from Another Lady's with 93.8% accuracy, only mislabeling two documents.

4.3 Obama-McCain & Ezekiel-Jeremiah

In order to confirm our framework is accurate over a variety of documents, we considered campaign speeches from the 2008 presidential election. Collecting 27 speeches from President Obama and 20 from Senator McCain, we expected our technique to excel in this context. We found instead that our method performed exceptionally poorly, clustering these speeches with only 74.2% accuracy. Indeed, we were further surprised to discover that by adjusting our framework to be similar to that presented in Akiva and Koppel (2013) and Aldebei et al. (2015) – by replacing POS n-grams with ordinary word occurrences in step one – our framework performed very well, clustering at 95.3%.

Similarly, our framework performed poorly on the Books of Ezekiel and Jeremiah from the Hebrew Bible. Using the English-translated King James Version, and considering each chapter as an individual document, our framework clusters the 48 chapters of Ezekiel and the 52 chapters of Jeremiah at 54.7%. Aldebei et al. (2015) reports 98.0% on this dataset, and when considering the original English text instead of the POS-translated text, our framework achieves 99.0%. The simultaneous success of word features and failure of POS features on these two datasets seemed to completely contradict our previous results.

We propose two explanations. First, perhaps too much syntactic structure is lost during translation. This could certainly be a factor, but does not explain the Obama-McCain results. The second explanation comes from the wide consensus among biblical scholars that there was no single 'Ezekiel' or 'Jeremiah' entirely responsible for each book. Instead, the books are composites from a number of authors, sometimes written over the span of hundreds of years (McKane, 1986; Zimmerli, 1979; Mowinckel, 1914). Koppel et al. (2011) acknowledges this shortcoming in their original paper, and suggest that in this authorial interpretation their clustering is one of style, not authorship. We hypothesize that in both failed cases, accuracy is low because our assumption that only two authors were represented among the documents is incorrect. This theory holds for the Obama-McCain dataset, because Obama had up to three primary speechwriters during the '08 election and McCain likely had a similar number (Parker, 2008). Perhaps emulating syntactic patterns is more difficult than emulating word choice. If so, using word fea-

Author	Cluster I	Cluster II
Ezekiel 1	**37**	2
Ezekiel 2	1	**8**

Table 3: Results when clustering the Hebrew text of the Book of Ezekiel split over the two authors.

Author	Cluster I	Cluster II
Jeremiah 1	**21**	2
Jeremiah 2	0	**14**

Table 4: Results when clustering the Hebrew text of the Book of Jeremiah split over the two primary authors.

tures, a model can discern Obama's rhetoric from that of McCain. However, since the syntax of more than two individuals is present in the text, POS features cannot accurately cluster the documents into two groups. Our goal is for POS features to cluster more accurately than word features when the true authorship of the documents is correctly considered.

5 Testing Our Theory

We first attempt to cluster the Ezekiel and Jeremiah texts in the original Hebrew in order to test if too much syntactic structure is lost during translation. For the Hebrew text, we use hand-tagged POS information because a reliable automatic tagger was not available (van Peursen et al., 2015; Roorda, 2015). Clustering Ezekiel and Jeremiah using Hebrew POS features obtains 62.5% accuracy. This is an improvement over the English text, but still performs far worse than lexical feature sets.

We next attempt to cluster the Ezekiel and Jeremiah texts according to the authorial strata within each book that is widely agreed upon by biblical scholars, in order to test if incorrect authorial assumptions were causing the decrease in accuracy. Unfortunately, there is no public breakdown of Obama and McCain speeches by speechwriter, so testing our hypothesis is limited here to the biblical dataset.

We therefore cluster the Book of Ezekiel assuming there are two nested authors, which according to modern scholarship are Ezekiel 1 (chapters 1–39) and Ezekiel 2 (chapters 40–48) (Zimmerli, 1979). Summarized in Table 3, according to this division our framework clusters the Ezekiel chapters with 93.6% accuracy, mislabeling only three documents. We also consider the Book of Jeremiah, which is composed of two primary authors with four secondary authors. In clustering a corpus containing Jeremiah 1 (23 non-contiguous chapters) and Jeremiah 2 (14 non-contiguous chapters) (McKane, 1986), our framework divides the 37 chapters into two groups with 94.5% accuracy, mislabeling only two documents. These results are summarized in Table 4. When considering the 4-way split between Ezekiel 1, Ezekiel 2, Jeremiah 1 and Jeremiah 2, our method achieves 87.5% accuracy as summarized in Table 5.

Author	C I	C II	C III	C IV
Ezekiel 1	**32**	2	5	0
Ezekiel 2	1	**8**	0	0
Jeremiah 1	0	0	**21**	2
Jeremiah 2	0	0	0	**14**

Table 5: Results when clustering Ezekiel 1 and 2 and Jeremiah 1 and 2 simultaneously with $k = 4$.

When comparing these results with those obtained by looking at word frequencies in the original Hebrew texts partitioned into the four correct authors, we find that our approach performs significantly better. With word frequencies as features, our framework clusters Ezekiel 1 from Ezekiel 2 with only 76.3% accuracy, Jeremiah 1 from Jeremiah 2 with only 74.9% accuracy, and crucially, clusters the four-way between both Ezekiels and both Jeremiahs with only 47.9% accuracy. While lexical features outperform syntactic features when considering incorrect authorship, syntactic features substantially outperform lexical features when considering the true authorial divisions of Ezekiel and Jeremiah.

6 Conclusion and Future Work

We have demonstrated a new framework for authorial clustering that not only clusters canonical datasets with state-of-the-art accuracy, but also discerns nested authorship within the Hebrew Bible more accurately than prior work. While we believe it is possible for an author to emulate another author's word choice, it is much more difficult to emulate unconscious syntactic structure. These syntactic patterns, rather than lexical frequencies, may therefore be key to understanding authorial fingerprints. Finding testing data for this problem is difficult, since documents for which authorship is misconstrued or obfuscated but for which true authorship is known with certainty are rare. However, when clustering texts for which authorship is not known, one would wish to have a framework which most accurately discerns author-

ship, rather than rhetorical similarity. We believe that our framework, and syntactic feature sets in particular, clusters documents based on authorship more accurately than prior work. While we have shown that POS feature sets can succeed independently, future work should examine augmenting syntactic and lexical feature sets in order to utilize the benefits of each.

Finally, authorial clustering performs poorly when the number of true and expected authors within a corpus do not match. An important next step is to automatically identify the number of authors contained within a set of documents. We believe that a more reliable method of generating 'core elements' is essential, and should not be reliant on a predetermined number of authors.

Acknowledgments

We thank Laurent El Ghaoui, Professor of EECS and IEOR, UC Berkeley, and Ronald Hendel, Norma and Sam Dabby Professor of Hebrew Bible and Jewish Studies, UC Berkeley, for comments that greatly improved the paper.

References

Navot Akiva and Moshe Koppel. 2013. A generic unsupervised method for decomposing multi-author documents. *Journal of the American Society for Information Science and Technology*, pages 2256–2264.

Khaled Aldebei, Xiangjian He, and Jie Yang. 2015. Unsupervised decomposition of a multi-author document based on naive-bayesian model. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 501–505.

Harald Baayen, Hans Van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11:121–131.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1356–1364.

Robert Layton, Paul Watters, and Richard Dazeley. 2013. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*.

William McKane. 1986. *A Critical and Exegetical Commentary on Jeremiah*. Edinburgh, Edinburgh, UK.

Todd K. Moon, Peg Howland, and Jacob H Gunther. 2006. Document author classification using generalized discriminant analysis. *Proc. Workshop on Text Mining, SIAM Int'l Conf. on Data Mining*.

Sigmund Mowinckel. 1914. *Zur Komposition des Buches Jeremia*. Kristiania.

Ashley Parker. 2008. What would obama say? *New York Times*, January 20.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Dirk Roorda. 2015. Laf-fabric software. https://github.com/ETCBC/laf-fabric. GitHub repository.

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEE Transactions On Pattern Analysis And Machine Intelligence*, 22:888–905.

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

Michael Tschuggnall and Gunther Specht. 2014. Enhancing authorship attribution by utilizing syntax tree profiles. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 195–199.

W.T. van Peursen, M. Sc. C. Sikkel, and D. Roorda. 2015. Hebrew text database etcbc4b. http://dx.doi.org/10.17026/dans-z6y-skyh. DANS.

Stella X. Yu and Jianbo Shi. 2003. Multiclass spectral clustering. *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 1:313–319.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.*, 57(3):378–393, February.

Walther Zimmerli. 1979. *Ezekiel 1-2: A Commentary on the Book of the Prophet Ezekiel*. Fortress Press of Philadelphia, Philadelphia, US.