What constitutes “style” in authorship attribution?

Kalaivani Sundararajan
Florida Institute for Cybersecurity
University of Florida
kalaivani.s@ufl.edu

Damon L. Woodard
Florida Institute for Cybersecurity
University of Florida
dwoodard@ece.ufl.edu

Abstract

Authorship attribution typically uses all information representing both content and style whereas attribution based only on stylistic aspects may be robust in cross-domain settings. This paper analyzes different linguistic aspects that may help represent style. Specifically, we study the role of syntax and lexical words (nouns, verbs, adjectives and adverbs) in representing style. We use a purely syntactic language model to study the significance of sentence structures in both single-domain and cross-domain attribution, *i.e.* cross-topic and cross-genre attribution. We show that syntax may be helpful for cross-genre attribution while cross-topic attribution and single-domain may benefit from additional lexical information. Further, pure syntactic models may not be effective by themselves and need to be used in combination with other robust models. To study the role of word choice, we perform attribution by masking all words or specific topic words corresponding to nouns, verbs, adjectives and adverbs. Using a single-domain dataset, IMDB1M reviews, we demonstrate the heavy influence of common nouns and proper nouns in attribution, thereby highlighting topic interference. Using cross-domain Guardian10 dataset, we show that some common nouns, verbs, adjectives and adverbs may help with stylometric attribution as demonstrated by masking topic words corresponding to these parts-of-speech. As expected, it was observed that proper nouns are heavily influenced by content and cross-domain attribution will benefit from completely masking them.

1 Introduction

Authorship attribution attributes a text sample to a person based on either content or their writing style. On the other hand, stylometry involves authorship attribution only based on writing style and needs to be independent of topic or genre. Though authorship attribution has various applications, some important ones are security-related. eg. cybercrime investigation, digital forensics, countering identity deception in social networks, etc. In today’s connected world, most of our cyber-interactions are text-based on various platforms like messages, emails, blogs, tweets, forum posts etc. With rampant cybercrime attacks, it is important to perform robust authorship attribution or verification across these platforms. This calls for stylometric attribution approaches that model style and work well in cross-domain scenarios. Cross-domain scenarios include both cross-topic (same genre but different topics) and cross-genre (different genres) scenarios.

Various authorship attribution approaches have been surveyed in literature (Juola and others, 2008; Stamatatos, 2009; Neal et al., 2017). However, only some of these approaches focus specifically on cross-domain attribution. Function words have been touted as content-independent and hence reliable as a style marker (Juola and others, 2008). Syntactic information, both shallow (Luyckx and Daelemans, 2008) and deep (Baayen et al., 1996; Raghavan et al., 2010; Feng et al., 2012; Fuller et al., 2014; Björklund and Zechner, 2017), have also been proposed for ensuring topic or genre independence. Few other approaches, perform preprocessing to prevent topic/genre dependency when using lexical or character features (Stamatatos, 2012; Sapkota et al., 2014; Markov et al., 2017; Stamatatos, 2018).

Proceedings of the 27th International Conference on Computational Linguistics, pages 2814–2822
Santa Fe, New Mexico, USA, August 20-26, 2018.

In this paper, we study the role of different linguistic aspects on style, specifically sentence syntax and word choice. We use a language model to represent syntax and words as against using vector representations with machine learning approaches. We do so for two reasons: i) Vector representation of syntax has not been effective in past efforts (Stamatatos, 2009), ii) Vector representation of words or character n-grams mostly consist of function words and may not be helpful to study the role of lexical parts-of-speech (POS).

To study the role of syntax, we perform authorship attribution using a purely syntactic language model. The syntactic model is constructed using the Probabilistic Context-free grammars (PCFG) obtained from an author's training data. While the approach is similar to other syntactic approaches (Raghavan et al., 2010), we keep our model purely syntactic by removing any lexical information. We also add context to the rewrite rules by vertical and horizontal Markovization. To handle previously unseen rules, we also incorporate smoothing where the language model can back-off to lower order models.

To study the role of word choice, we perform attribution by masking out all words or specific topic words corresponding to different lexical POS. We do so because character n-gram based approaches have largely outperformed function word based approaches (Kestemont, 2014) indicating that lexical words may also help with authorship attribution. This analysis would help understand which of these lexical words may help represent style. Other approaches (Stamatatos, 2018) also mask out infrequent words (mostly lexical words) but we perform an in-depth analysis as to which of these lexical words are useful for representing style.

In this paper, we attempt to answer the following questions:

- Does sentence structure or syntax help with stylometric attribution in cross-domain settings?
- What are effects of words corresponding to different lexical POS on cross-domain attribution?
- Does masking topic words corresponding to various lexical POS help with stylometric attribution in cross-domain settings?

2 Methodology

In this section, we explain the approaches taken to answer the above questions pertaining to stylometric attribution. This could help us come up with better style representations that would work well in cross-domain scenarios.

2.1 Role of syntax

We use a purely syntactic language model to study the role of sentence structure or syntax in cross-domain stylometric attribution. A syntactic language model is obtained by constructing the probabilistic context-free grammar (PCFG) for each author using the constituency parse trees of sentences in their training posts. While some approaches (Raghavan et al., 2010) use the rewrite rules directly to construct PCFGs, we apply both vertical and horizontal Markovization (Klein and Manning, 2003) to the parse trees before constructing PCFGs. This helps to incorporate some context into the rewrite rules and improves parsing accuracy. To keep the approach purely syntactic, we remove the leaf nodes which contain the sentence words (Fuller et al., 2014; Björklund and Zechner, 2017). A test sample is attributed to the author whose PCFG yields the highest likelihood score. In order to account for unseen rules during test time, we also incorporate smoothing that allows the model to backoff to lower order syntactic language models. The rewrite rules of a sample sentence corresponding to different Markovization orders used in our model are shown in Figure 1.

We compare the syntactic language model with an analogous character language model (Teahan and Harper, 2003). This approach uses a lossless text compression method called Prediction by Partial Matching (PPM) (Cleary and Witten, 1984). With PPM, individual characters are encoded using the context provided by the preceding characters thus representing each author A using a separate language model p_A. To attribute an unknown sample $\mathbf{u}$ of length L, we compute its cross-entropy with respect to an

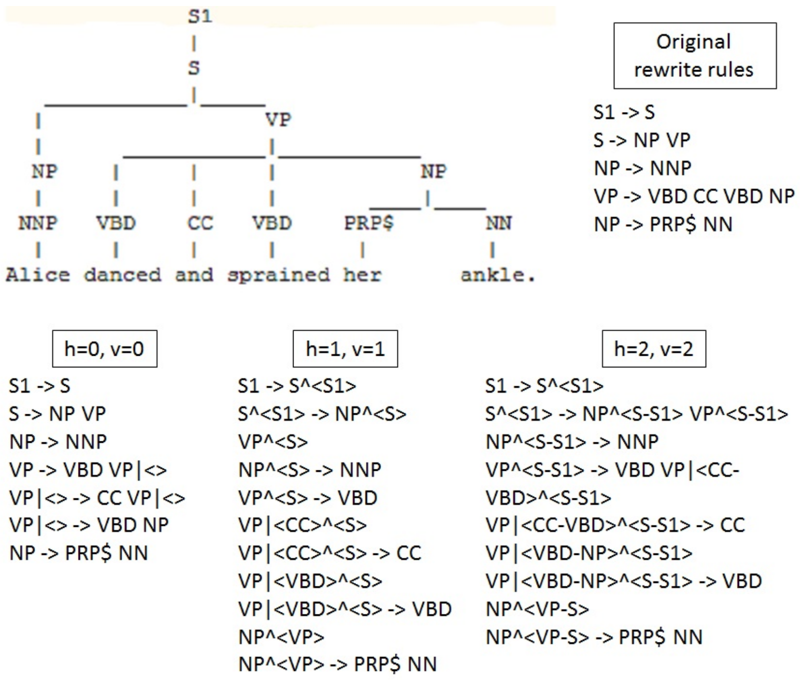


Figure 1: Parse tree of a sample sentence and rewrite rules under different orders of vertical (v) and horizontal (h) Markovization. The leaf nodes are excluded from the rewrite rules.

author model p_A as

$$\begin{aligned} H(p_A, u) &= -\frac{1}{L}\, log_2\, p_A(u) \\ &= -\frac{1}{L}\sum_{i=1}^{L} log_2\, p_A(x_i|context_i) \end{aligned}$$

where $context_i$ of character x_i is $x_1, x_2, ..., x_{i-1}$. The unknown sample is attributed to the author whose model yields least cross-entropy. For computational efficiency, the context is typically truncated to preceding n-1 characters i.e. $x_{i-n}, ..x_{i-1}$.

2.2 Role of word choice

Besides syntax, the choice of words used by a person also plays an important role in stylometric attribution. These words may have a primary purpose of being lexical or grammatical. Lexical words have meaning by themselves and are mostly content-specific. Such words typically include nouns, verbs, adjectives and adverbs. On the other hand, grammatical words do not have any meaning by themselves and specify the relationships between lexical words. Hence, they are independent of content and typically include determiners, prepositions, pronouns, etc. Conventionally, grammatical words, especially function words, have been proposed for stylometric authorship attribution since they are independent of content. However, character n-gram based approaches have largely outperformed function word based approaches (Kestemont, 2014) indicating that some lexical words may also help with authorship attribution. In order to understand which lexical words may help with stylometric attribution, we study the effects of masking all lexical words and certain topic words corresponding to different POS on authorship attribution.

2.2.1 Role of lexical POS

To analyze the effects of different lexical POS on stylometric attribution, we preprocess the posts to replace words corresponding to different POS with a predefined string <*T*>. We use the set of Penn TreeBank POS tags in our experiments. The following effects are analyzed using a character language model (Teahan and Harper, 2003) for both single-domain and cross-domain datasets:

- **orig:** This utilizes the original posts and are used for benchmarking.
- **no_NNP:** This utilizes posts in which all proper nouns (NNP, NNPS) are replaced by <*T*>.
- **no_NN:** This utilizes posts in which all common nouns (NN, NNS) are replaced by <*T*>.
- **no_VB:** This utilizes posts in which all verbs (VB, VBD, VBG, VBN, VBP, VBZ) except function words are replaced by <*T*>.
- **no_ADJ:** This utilizes posts in which all adjectives (JJ, JJR, JJS) are replaced by <*T*>.
- **no_ADV:** This utilizes posts in which all adverbs (RB, RBR, RBS) are replaced by <*T*>.

2.2.2 Role of topic words

The approach in Section 2.2.1 assumes that all words corresponding to a lexical POS may be content-specific. This may not be the case as some of these words may actually help with stylometric attribution. Hence, in this section, we analyze the effects of masking only the topic words corresponding to different lexical POS on attribution. We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to obtain the topic words for each lexical POS. Since each author may focus on different topics, we perform LDA for each author using their training posts. The topic words of all authors are then collated and used for masking. We analyze the same effects as Section 2.2.1 - *orig, no_topic_NNP, no_topic_NN, no_topic_VB, no_topic_ADJ* and *no_topic_ADV*.

3 Datasets

We use two datasets for our experiments as follows:

- **IMDB1M reviews** (Seroussi et al., 2011): This single-domain dataset consists of 204,809 posts and 66,816 reviews written by 22,116 users. For our experiments, we use only a subset of 674 users with more than 50 posts per person. The chosen data subset contains 160,482 posts across 674 users. On an average, the dataset consists of 238 posts/author, 347 characters/post, 73 words/post and 4 sentences/post.

- **Guardian10 corpus** (Stamatatos, 2012): This cross-domain dataset consists of opinion articles and book reviews written by 13 authors and up to 10 posts per topic/genre. On an average, this dataset consists of 34 posts/author, 6234 characters/post, 1202 words/post and 52 sentences/post. Cross-topic experiments are performed only using the opinion articles which are on four different topics - Politics, UK, World and Society. Cross-genre experiments involve training author models using opinion articles on all four topics and evaluating authorship attribution performance on book reviews.

4 Results

In this section, we report the results of our analysis detailed in Section 2.

4.1 Role of syntax

We use the BLLIP parser (Charniak and Johnson, 2005) trained on Wall Street Journal (WSJ) corpus to treebank an author's training data. We use vertical and horizontal Markovization order of two *(h=2, v=2)* since higher orders did not seem to improve performance. The leaf nodes are removed from the parse trees to keep the analysis purely syntactic. To test stylometric attribution, we evaluate the syntactic language model (Syntactic LM) on both single-domain IMDB1M and cross-domain Guardian10 datasets. We repeat these experiments using the syntactic language model with lexical information in its leaf nodes (Syntactic LM + Lexical) and also using the character language model (Character LM) (Teahan and Harper, 2003) for comparison. The single-domain and cross-domain attribution performances are reported in Tables 1 and 2 respectively.

Table 1: Single-domain performance on IMDB1M using syntactic language model

Scenarios	Precision(%)	Recall(%)	F-score(%)	Accuracy(%)
Syntactic LM	30.96	27.71	28.28	27.71
Syntactic LM + Lexical	44.31	43.52	42.26	43.53
Character LM	64.41	64.77	63.42	64.77

Table 2: Cross-domain performance on Guardian10 using syntactic language model

Domain	Scenarios	Precision(%)	Recall(%)	F-score(%)	Accuracy(%)
Cross-topic	Syntactic LM	21.85	22.22	18.40	24.14
	Syntactic LM + Lexical	27.30	25.15	21.27	26.56
	Character LM	76.89	67.41	67.02	70.41
Cross-genre	Syntactic LM	20.44	26.96	21.90	33.33
	Syntactic LM + Lexical	18.92	23.15	18.99	28.57
	Character LM	72.32	56.90	59.79	69.84

4.1.1 Discussion

We studied the role of sentence structure or syntax in stylometric attribution using a purely syntactic language model. For single-domain IMDB1M dataset, it can be inferred from Table 1 that using a purely syntactic model (*Syntactic LM*) seems to be ineffective as compared to using both syntax and lexical information (*Syntactic LM + Lexical*). This reaffirms the fact that single-domain attribution benefits largely from lexical information some of which may be topic-dependent. For cross-domain settings, it can be inferred from Table 2 that purely syntactic attribution (*Syntactic LM*) may be more helpful for cross-genre data than cross-topic data. Adding lexical information to purely syntactic model (*Syntactic LM + Lexical*) seems to improve cross-topic attribution while it deteriorates cross-genre attribution.

Nevertheless, purely syntactic models seem to perform poorly compared to character language models as is evident from in Tables 1 and 2. Character n-gram based approaches have been highly successful in authorship attribution because they represent character, lexical and syntactic information to varying extents (Luyckx, 2011; Kestemont, 2014) and are robust to morphological variations in language use (Sapkota et al., 2015). However, from a purely quantitative perspective, their success can be attributed to the fact that character n-grams have much more data points than function words or syntactic rules thereby yielding superior performance (Kestemont, 2014). Character n-grams span the same word or word combinations multiple times depending on the n-gram order and hence amplify the presence of possibly unique word/phrase choices by an author. PCFG-based syntactic language models do not have this advantage as the set of all possible syntactic rewrite rules is much lesser than the set of all possible character n-grams. Hence, the low performance of purely syntactic models does not necessarily imply that they are not useful in representing style. One can use these in combination with character language models to improve cross-domain performance.

4.2 Role of word choice

To analyze the role of word choice in stylometric attribution, we perform experiments on both single-domain and cross-domain datasets by masking out all lexical words and topic words corresponding to different POS.

4.2.1 Role of lexical POS

To study the effects of lexical words on authorship attribution, we mask all words corresponding to specific lexical POS as described in Section 2.2.1. We use *NLTK* toolkit (Bird and Loper, 2004) for POS tagging. We perform author identification on both single-domain and cross-domain datasets using the character language model (Teahan and Harper, 2003). We perform 4-fold cross-validation and report the average precision, recall, F-score and accuracy across all cross-validation folds.

For single domain IMDB1M dataset, the performance measures reflecting the effects of masking different lexical POS are reported in Table 3. Similarly, the performance measures for cross-topic and cross-genre experiments on Guardian10 dataset are reported in Table 4.

Table 3: Effect of lexical POS on single-domain authorship attribution using IMDB1M

Scenarios	**Precision(%)**	**Recall(%)**	**F-score(%)**	**Accuracy(%)**
orig	64.41	64.77	63.42	64.77
no_NN	53.78	50.47	50.73	50.47
no_topic_NN	62.19	62.33	61.12	62.33
no_NNP	59.66	58.38	57.83	58.38
no_topic_NNP	63.20	63.25	62.12	63.25
no_VB	62.78	62.99	61.71	62.99
no_topic_VB	63.23	63.48	62.17	63.48
no_ADJ	63.12	63.25	62.04	63.25
no_topic_ADJ	63.48	63.71	62.45	63.71
no_ADV	63.42	63.91	62.53	63.91
no_topic_ADV	64.02	64.41	63.07	64.41

Table 4: Effect of lexical POS on cross-domain authorship attribution using Guardian10

Domain	**Scenarios**	**Precision(%)**	**Recall(%)**	**F-score(%)**	**Accuracy(%)**
Cross-topic	orig	77.79	70.41	69.50	70.41
	no_NN	80.46	69.10	69.35	69.10
	no_topic_NN	84.13	73.13	72.17	73.13
	no_NNP	85.86	80.30	79.98	80.30
	no_topic_NNP	86.25	75.48	74.68	75.48
	no_VB	81.12	67.33	66.77	67.33
	no_topic_VB	80.13	71.41	70.21	71.41
	no_ADJ	78.18	69.49	68.26	69.49
	no_topic_ADJ	83.25	72.73	71.07	72.74
	no_ADV	76.31	67.92	67.13	67.92
	no_topic_ADV	84.17	73.22	72.24	73.22
Cross-genre	orig	78.99	69.84	69.94	69.84
	no_NN	79.18	71.43	71.52	71.43
	no_topic_NN	82.92	73.02	72.80	73.02
	no_NNP	84.52	82.54	80.97	82.56
	no_topic_NNP	81.68	71.43	71.44	71.43
	no_VB	81.82	74.60	74.02	74.60
	no_topic_VB	80.37	73.02	72.34	73.02
	no_ADJ	76.37	69.84	69.61	69.84
	no_topic_ADJ	79.45	71.43	71.14	71.43
	no_ADV	78.71	73.02	72.42	73.02
	no_topic_ADV	80.89	74.60	73.93	74.60

4.2.2 Role of topic words

In this section, we hypothesize that not all words corresponding to lexical POS are content-specific. Hence, we experiment with masking certain topic words corresponding to different lexical POS. The topic words are chosen using the LDA implementation in *gensim* Python toolkit. Only words corresponding to specific lexical POS are input to LDA. We experiment by varying the number of topics (t = 2,10,50) and number of words per topic (w=10,100). The topic words obtained from each author's training data are collated and used for masking. We perform 4-fold cross-validation experiments on both single-domain and cross-domain datasets. We report the performance measures for the best performing configuration using both IMDB1M (t=50, w=10) and Guardian datasets (t=10, w=10) in Tables 3 and 4.

4.2.3 Discussion

It is known that function words are independent of content and are useful for representing style. However, the success of character n-gram approaches in authorship attribution indicate that some lexical words may also be useful for authorship attribution. Besides, character n-gram approaches do not necessarily decouple style and content and using these approaches as such may deteriorate attribution in cross-domain settings. Hence, we studied the effect of masking all words or topic words corresponding to different lexical POS on both single-domain and cross-domain attribution as explained in Sections 2.2.1

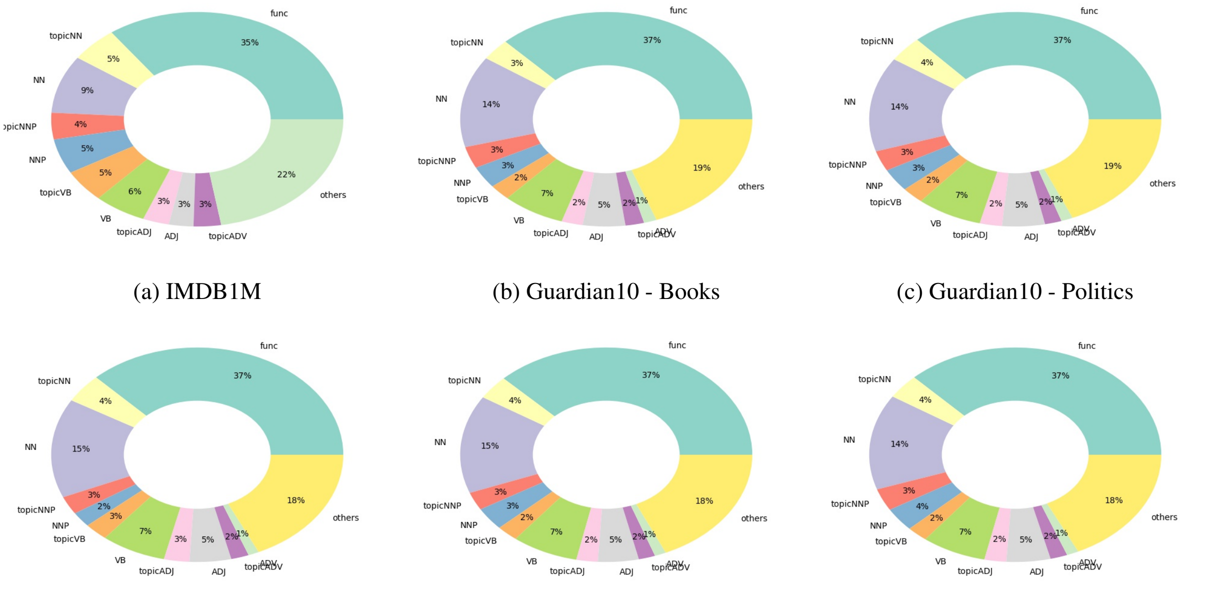
and 2.2.2. For all experiments, attribution with the character language model without masking any lexical POS ($orig$) is considered as the baseline.

For single-domain IMDB1M dataset, it can be observed from Table 3 that excluding all common nouns (no_NN) and proper nouns (no_NNP) affects the attribution performance drastically. In contrast, masking only topic words corresponding to common nouns (no_topic_NN) and proper nouns (no_topic_NNP) seems to improve attribution performance compared to masking them completely (no_NNP, no_NN). Nevertheless, this performance is lesser than that of $orig$. This shows the heavy influence of nouns and proper nouns in single-domain attribution. Masking all words or topic words corresponding to other lexical POS like verbs, adjectives or adverbs do not seem to impact the performance even though POS like verbs do contribute significantly to attribution as seen in Figure 2a. This could possibly be due to the heavy dependence on common nouns and proper nouns for attribution under those scenarios.

In contrast, for cross-domain Guardian10 dataset, it can be observed from Table 4 that excluding proper nouns (no_NNP) completely yields a marked improvement in performance for both cross-topic and cross-genre experiments. In fact, excluding only topic words corresponding to proper nouns (no_topic_NNP) performs poorly compared to no_NNP though it performs marginally better than $orig$. Hence, while performing cross-domain attribution, it would be wise to exclude proper nouns from style representations. With respect to common nouns, masking topic words (no_topic_NN) improves performance as compared to completely masking them (no_NN) for both cross-topic and cross-genre scenarios. This implies that some common nouns also help represent style as shown by their distribution in Figures 2b,2c,2d,2e and 2f. With respect to verbs, in cross-topic settings, masking topic words (no_topic_VB) marginally improves performance while completely masking them (no_VB) reduces performance. In cross-genre settings, masking all verbs completely (no_VB) improves performance significantly while masking only topic verbs (no_topic_VB) improves performance marginally compared to $orig$. This implies that choice of verbs have a significant influence in cross-genre settings as compared to cross-topic settings. For both adjectives and adverbs, masking certain topic words (no_topic_ADJ and no_topic_ADV) improves attribution performance as against completely masking them (no_ADJ and no_ADV) for both cross-topic and cross-genre settings. This suggests that some of these adverbs and adjectives may help represent style. The top 20 non-topic words corresponding to different lexical POS for both datasets are shown in Table 5.

Table 5: Top 20 non-topic lexical POS

Guardian10				**IMDB1M**			
Nouns	**Verbs**	**Adjectives**	**Adverbs**	**Nouns**	**Verbs**	**Adjectives**	**Adverbs**
nothing	given	real	probably	opinion	done	little	generally
something	found	better	instead	world	give	overall	somewhat
one	got	past	alone	kind	said	modern	also
end	asked	big	first	sense	put	nice	deeply
point	used	modern	enough	others	thought	new	together
work	came	large	especially	scene	do	fine	just
day	tell	true	soon	day	going	third	hopefully
moment	wants	recent	indeed	times	guess	impressive	perhaps
year	let	less	easily	man	watched	realistic	fairly
kind	turned	right	hardly	thing	getting	long	completely
fact	allowed	hard	perhaps	place	seeing	glad	definitely
question	told	obvious	only	someone	use	fair	first
things	seen	best	surely	story	call	enjoyable	seriously
example	making	wrong	though	reason	look	similar	long
place	seemed	easy	later	chance	saying	less	later
course	thought	full	less	plot	given	easy	therefore
story	mean	foreign	simply	mind	let	huge	best
evidence	run	private	certainly	death	gets	next	personally
business	trying	general	seriously	idea	enjoy	likely	obviously
nation	happen	long	instead	experience	released	happy	only



(a) IMDB1M

(b) Guardian10 - Books

(c) Guardian10 - Politics

(d) Guardian10 - Society

(e) Guardian10 - UK

(f) Guardian10 - World

Figure 2: POS distribution on single-domain and cross-domain datasets. POS with <1% contribution have been suppressed

One of the most pressing problems in this line of research is that there seems to be a dearth for publicly available large-scale cross-domain datasets. Authorship attribution approaches in literature largely demonstrate their results using either small datasets or large-scale single-domain datasets. This does not provide a clear picture of the factors that contribute to style which may be robust in cross-domain settings. The role of syntax or word choice or other aspects in representing style can be better ascertained with experiments on large-scale cross-domain datasets.

5 Conclusion

Authorship attribution approaches in literature focus mostly on single-domain attribution where content and style are highly entangled. This does not provide a clear picture of linguistic aspects that are robust to cross-domain settings. In this paper, we studied the role of sentence structure and word choice in representing style. We evaluated the role of syntax using a purely syntactic language model and show that syntax may be useful with cross-genre attribution while cross-topic attribution and single-domain attribution may be benefit from both syntax and lexical information. However, syntactic language models are not discriminative by themselves and need to be used in conjunction with more successful character n-gram models. We evaluated the role of word choice by masking off all words or certain topic words corresponding to different lexical POS. For common nouns, verbs, adjectives and adverbs, masking off certain topic words yield better performance suggesting that the remaining words corresponding to these lexical POS may help represent style. However, proper nouns seem to be heavily influenced by topic and cross-domain attribution may benefit from completely masking them.

References

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(3):121–132.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Johanna Björklund and Niklas Zechner. 2017. Syntactic methods for topic-independent authorship attribution. *Natural Language Engineering*, 23(5):789–806.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 173–180. Association for Computational Linguistics.

John Cleary and Ian Witten. 1984. Data compression using adaptive coding and partial string matching. *IEEE transactions on Communications*, 32(4):396–402.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Characterizing stylistic elements in syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1522–1533. Association for Computational Linguistics.

Simon Fuller, Phil Maguire, and Philippe Moser. 2014. A deep context grammatical model for authorship attribution.

Patrick Juola et al. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *CLfL@ EACL*, pages 59–66.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 513–520. Association for Computational Linguistics.

Kim Luyckx. 2011. *Scalability issues in authorship attribution.* ASP/VUBPRESS/UPA.

Ilia Markov, Efstathios Stamatatos, and Grigori Sidorov. 2017. Improving cross-topic authorship attribution: The role of pre-processing. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing*.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6):86:1–86:36, November.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42. Association for Computational Linguistics.

Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 93–102.

Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. 2011. Personalised rating prediction for new users using latent factor models. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 47–56. ACM.

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

Efstathios Stamatatos. 2012. On the robustness of authorship attribution based on character n-gram features. *JL & Pol'y*, 21:421.

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

William J Teahan and David J Harper. 2003. Using compression-based language models for text categorization. In *Language modeling for information retrieval*, pages 141–165. Springer.