Authorship Attribution Using Text Distortion

Efstathios Stamatatos
University of the Aegean
83200, Karlovassi, Samos, Greece
stamatatos@aegean.gr

Abstract

Authorship attribution is associated with important applications in forensics and humanities research. A crucial point in this field is to quantify the personal style of writing, ideally in a way that is not affected by changes in topic or genre. In this paper, we present a novel method that enhances authorship attribution effectiveness by introducing a text distortion step before extracting stylometric measures. The proposed method attempts to mask topic-specific information that is not related to the personal style of authors. Based on experiments on two main tasks in authorship attribution, closed-set attribution and authorship verification, we demonstrate that the proposed approach can enhance existing methods especially under cross-topic conditions, where the training and test corpora do not match in topic.

1 Introduction

Authorship attribution is the task of determining the author of a disputed text given a set of candidate authors and samples of their writing (Juola, 2008; Stamatatos, 2009). This task has gained increasing popularity since it is associated with important forensic applications, e.g., identifying the authors of anonymous messages in extremist forums, verifying the author of threatening email messages, etc. (Abbasi and Chen, 2005; Lambers and Veenman, 2009; Coulthard, 2013), as well as humanities and historical research, e.g., unmasking the authors of novels published anonymously or under aliases, verifying the authenticity of literary works by specific authors, etc. (Koppel and Seidman, 2013; Juola, 2013; Stover et al., 2016).

The majority of published works in authorship attribution focus on *closed-set attribution* where it is assumed that the author of the text under investigation is necessarily a member of a given well-defined set of candidate authors (Stamatatos et al., 2000; Gamon, 2004; Escalante et al., 2011; Schwartz et al., 2013; Savoy, 2013; Seroussi et al., 2014). This setting fits many forensic applications where usually specific individuals have access to certain resources, have knowledge of certain issues, etc. (Coulthard, 2013) A more general framework is *open-set attribution* (Koppel et al., 2011). A special case of the latter is *authorship verification* where the set of candidate authors is singleton (Stamatatos et al., 2000; van Halteren, 2004; Koppel et al., 2007; Jankowska et al., 2014; Koppel and Winter, 2014). This is essentially a one-class classification problem since the negative class (i.e., all texts by all other authors) is huge and extremely heterogeneous. Recently, the verification setting became popular in research community mainly due to the corresponding PAN shared tasks (Stamatatos et al., 2014; Stamatatos et al., 2015).

In authorship attribution it is not always realistic to assume that the texts of known authorship and the texts under investigation belong in the same genre and are in the same thematic area. In most applications, there are certain restrictions that do not allow the construction of a representative training corpus. Unlike other text categorization tasks, a recent trend in authorship attribution research is to build cross-genre and cross-topic models, meaning that the training and test corpora do not share the same properties (Kestemont et al., 2012; Stamatatos, 2013; Sapkota et al., 2014; Stamatatos et al., 2015).

One crucial issue in any authorship attribution approach is to quantify the personal style of authors, a line of research also called stylometry (Stamatatos, 2009). Ideally, stylometric fea-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1138–1149,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

tures should not be affected by shifts in topic or genre variations and they should only depend on personal style of the authors. However, it is not yet clear how the topic/genre factor can be separated from the personal writing style. Function words (i.e., prepositions, articles, etc.) and lexical richness features are not immune to topic shifts (Mikros and Argiri, 2007). In addition, character *n*-grams, the most effective type of features in authorship attribution as demonstrated in multiple studies (Grieve, 2007; Stamatatos, 2007; Luyckx and Daelemans, 2008; Escalante et al., 2011) including cross-topic conditions (Stamatatos, 2013; Sapkota et al., 2015), unavoidably capture information related to theme and genre of texts. Features of higher level of analysis, including measures related to syntactic or semantic analysis of texts, are too noisy and less effective, and can be used as complement to other more powerful low-level features (van Halteren, 2004; Argamon et al., 2007; Hedegaard and Simonsen, 2011).

In this paper, we propose a novel method that is based on text distortion to compress topic-related information. The main idea of our approach is to transform input texts to an appropriate form where the textual structure, related to personal style of authors, is maintained while the occurrences of the least frequent words, corresponding to thematic information, are masked. We show that this distorted view of text when combined with existing authorship attribution methods can significantly improve their effectiveness under cross-topic conditions in both closed-set attribution and authorship verification.

2 Related Work

Previous work in authorship attribution focuses mainly on stylometric features that capture aspects of personal writing style (Gamon, 2004; Luyckx and Daelemans, 2008; Escalante et al., 2011; Schwartz et al., 2013; Tschuggnall and Specht, 2014; Sidorov et al., 2014). In addition, beyond the use of typical classification algorithms, several attribution models that are specifically designed for authorship attribution tasks have been proposed (Koppel et al., 2011; Seroussi et al., 2014; Qian et al., 2014). Basic approaches and models are reviewed by Juola (2008) and Stamatatos (2009). In addition, recent studies in authorship verification are surveyed in (Stamatatos et al., 2014; Stamatatos et al., 2015).

An early cross-topic study in authorship attribution using a very small corpus (3 authors and 3 topics) showed that the identification of authors of email messages is not affected too much when the training and test messages are on different topics (de Vel et al., 2001). Based on another small corpus (2 authors and 3 topics) Madigan, *et al.* (2005) demonstrated that POS features are more effective than word unigrams in cross-topic conditions. The *unmasking* method for author verification of long documents based on very frequent word frequencies was successfully tested in cross-topic conditions (Koppel et al., 2007) but Kestemont, *et al.* (2012) found that its reliability was significantly lower in cross-genre conditions. Function words have been found to be effective when topics of the test corpus are excluded from the training corpus (Baayen et al., 2002; Goldstein-Stewart et al., 2009; Menon and Choi, 2011). However, Mikros and Argiri (2007) demonstrated that function word features actually correlate with topic. Other types of features found effective in cross-topic and cross-genre authorship attribution are punctuation mark frequencies (Baayen et al., 2002), LIWC features (Goldstein-Stewart et al., 2009), and character *n*-grams (Stamatatos, 2013). To enhance the performance of attribution models based on character *n*-gram features, Sapkota et al. (2015) define several *n*-gram categories and then they combine *n*-grams that correspond to word affixes and punctuation marks. Combining several topics in the training set seems also to enhance the ability to identify the authors of texts on another topic (Sapkota et al., 2014). More recently, Sapkota et al. (2016) proposed a domain adaptation model based on structural correspondence learning and punctuation-based character *n*-grams as pivot features.

Text distortion has successfully been used to enhance thematic text clustering by masking the occurrences of frequent words while maintaining the textual structure (Granados et al., 2011; Granados et al., 2012). That way, the clustering model was no longer confused by non relevant information hidden in the produced distorted text (Granados et al., 2014). An important conclusion drawn by these studies was that, in cases the textual structure was not maintained, the performance of clustering decreased despite the fact that the same thematic information was available.

3 Text Distortion Views

In this paper we propose a method to transform texts by applying a text distortion process before extracting the stylometric features. The main idea is to provide a new version of texts that is more topic-neutral in comparison to the original texts while maintaining most of the information related with the personal style of the author. Our method is inspired by the text distortion approach introduced by Granados, *et al.* (2011; 2012) but it significantly differs from that since it is more suitable for authorship attribution rather than thematic clustering. In more detail, given the k most frequent words of the language W_k, the proposed method transforms the input $Text$ as described in Algorithm 1.

Algorithm 1 DV-MA

```
Input: $Text$, $W_k$
Output: $Text$
Tokenize $Text$
for each token $t$ in $Text$ do
    if lowercase($t$) $\notin W_k$ then
        replace each digit in $t$ with #
        replace each letter in $t$ with *
    end if
end for
```

We call this method *Distorted View with Multiple Asterisks* (DV-MA). Alternatively, a single symbol can be used to replace sequences of digits/letters meaning that the token length information is lost. This version of the proposed method, called *Distorted View with Single Asterisks* (DV-SA) is illustrated in Algorithm 2.

Algorithm 2 DV-SA

```
Input: $Text$, $W_k$
Output: $Text$
Tokenize $Text$
for each token $t$ in $Text$ do
    if lowercase($t$) $\notin W_k$ then
        replace any sequence of digits in $t$ with a single #
        replace any sequence of letters in $t$ with a single *
    end if
end for
```

Note that DV-MA does not affect the length of input text while DV-SA reduces text-length. The proposed text transformation is demonstrated in the example of Table 1 where an input text is transformed according to either DV-MA or DV-SA algorithms. In each example, W_k consists of the k most frequent words of the BNC corpus[1]. As can be seen, each value of k provides a distorted view on the text where the textual structure is maintained but some, mainly thematically related, information is masked. In the extreme case where k=0 all words are replaced by asterisks and the only information left concerns word-length, punctuation marks and numbers usage. When k=100, function words remain visible and it is possible to extract patterns of their usage. Note that capitalization of letters remain unaffected. When k increases to include thousands of frequent words of BNC more topic-related information is visible. In general, the lower the k, the more thematic information is masked. By appropriately tuning parameter k, it is possible to decide how much thematic information is going to be compressed.

The text distortion method described in Granados, *et al.* (2011; 2012) has also been applied to the input text of Table 1 for k=1,000. In comparison to that method, the proposed approach is different in the following points:

- We replace the occurrences of the least frequent words rather than the most frequent words since it is well known that function word usage provide important stylometric information.
- Punctuation marks and other symbols are maintained since they are important style markers.
- Capitalization of original text is maintained.
- We treat numbers in a special way in order to keep them in the resulting text but in a more neutral way that reflects the stylistic choices of authors. For example, note that in each example of Table 1 both $15,000 and $17,000 are transformed to the same pattern. Thus, the proposed methods are able to capture the format used by the author and discard the non-relevant information about the exact numbers. In the case of DV-SA, any similar number (e.g., $1,000, $100,000) would have exactly the same transformation.

[1] https://www.kilgarriff.co.uk/bnc-readme.html

Original text	The cars, slightly smaller than the Ford Taurus and expected to be priced in the $15,000-$17,000 range, could help GM regain a sizeable piece of the mid-size car market, a segment it once dominated.
DV-MA, k=0	*** **** , ******** ******* **** *** **** ****** *** ******** ** ** ****** ** *** $## , ### - $## , ### ***** , ***** **** ** ****** * ******** ***** ** *** *** - **** *** ****** , * ******* ** **** ********* .
DV-MA, k=100	The **** , ******** ******* than the **** ****** and ******** to be ****** in the $## , ### - $## , ### ***** , could **** ** ****** a ******** ***** of the *** - **** *** ****** , a ******* it **** ********* .
DV-MA, k=1,000	The **** , ******** ******* than the **** ****** and expected to be ****** in the $## , ### - $## , ### range , could help ** ****** a ******** ***** of the *** - size car market , a ******* it once ********* .
DV-SA, k=1,000	The * , * * than the * * and expected to be * in the $# , # - $# , # range , could help * * a * * of the * - size car market , a * it once * .
Granados et al. (2012), k=1,000	*** cars slightly smaller **** *** ford taurus *** ******** ** ** priced ** *** ******* ******* ***** ***** **** gm regain * sizeable piece ** *** mid **** *** ****** * segment ** **** dominated

Table 1: An example of transforming an input text according to DV-MA and DV-SA algorithms using different values of k.

The new version of texts after the application of the above distortion processes can then be used to extract regular stylometric features like character n-grams and word n-grams. The resulting features are expected to be more topic-neutral and therefore more useful to an authorship attribution model that is applied to cross-topic problems. One crucial issue now is the appropriate selection of parameter k that reflects how much thematic information is going to remain in the representation of text. As will be explained in the next section the most suitable value of k can be estimated using either the training or a validation corpus and it reflects the thematic differences in texts by the same authors.

4 Experimental Settings

In this section we are going to examine the effectiveness of the proposed text distortion method when combined with regular stylometric features and existing authorship attribution approaches. In more detail, the following features, popular in previous authorship attribution studies, are extracted from text: character n-grams and token n-grams[2]

In both cases, following the suggestions of previous work, the most frequent n-grams of the training corpus are included in the feature set (Stamatatos, 2009). Towards this direction, there are two alternatives: either selecting the top d most frequent n-grams or selecting all words with at least f_t occurrences in the training corpus. In this study, we adopt the latter approach. Thus, for each of the above type of features there are two parameters: the order (length) of n-grams (n) and the frequency threshold (f_t). In addition, when the proposed text distortion method is used, an additional parameter is introduced, the k most frequent words of the language.

In most previous studies, predefined values of n and f_t (or d) are used (Hedegaard and Simonsen, 2011; Schwartz et al., 2013; Qian et al., 2014; Sapkota et al., 2014; Sapkota et al., 2015). However, the appropriate tuning of these parameters is crucial especially in attribution methods that are tested in cross-topic or cross-genre conditions (Stamatatos, 2013). In this paper, we estimate the most appropriate values of the three above parameters by performing grid search on the training corpus or a validation corpus (separate from the test corpus) (Jankowska et al., 2014). In more detail, the following initial set of values are examined: $n \in \{3, 4, 5, 6, 7, 8\}$ for character n-grams and $n \in \{1, 2, 3\}$ for token n-grams, $f_t \in \{5, 10, 15, ..., 50\}$, and $k \in \{0, 100, 200, ..., 500, 1000, 2000, ..., 5000\}$. In case of ties, the parameter settings that correspond to the lowest feature set size are selected.

In each of the experiments presented in the following sections, the effect of the proposed text distortion approach is examined when combined with an existing well-known attribution model. We are going to examine the following three cases:

- Baseline: original input texts are used (no text distortion).
- DV-MA: the input texts are distorted using the Algorithm 1.

[2] We avoid to use the term word n-grams to put emphasis on the fact that all tokens are taken into account including punctuation marks and numbers.

- DV-SA: the input texts are distorted using the Algorithm 2.

5 Closed-set Attribution

5.1 Corpora

First, we examine the case where a given closed-set of candidate authors is given. Multiple corpora are nowadays available for this task. We selected to use the following two corpora:

1. CCAT-10: this is a subset of the *Reuters Corpus v.1* comprising 10 authors and 100 texts per author belonging to the CCAT thematic category (corporate and industrial news). This corpus has been used in several previous studies (Plakias and Stamatatos, 2008; Escalante et al., 2011; Sapkota et al., 2015). Separate training and test corpora of equal size are provided.

2. Guardian: this is a corpus of articles from *The Guardian* UK newspaper. It includes opinion articles by 13 authors on 4 thematic areas (politics, society, UK, and world) as well as book reviews by the same authors. It has been used in previous work that focused on authorship attribution in cross-topic and cross-genre conditions (Stamatatos, 2013; Sapkota et al., 2014). Following the practice of previous studies, we use at most 10 texts per author in each category of this corpus.

It is important to highlight the main difference among the above corpora. CCAT-10 authors tend to write newswire stories on specific subjects and this is consistent in both training and test corpora. On the other hand, in the Guardian corpus the texts by one author cover several thematic areas and two genres (opinion articles and book reviews). Therefore, it is expected that an authorship attribution method that is not robust to topic shifts will be less effective in the Guardian corpus. In CCAT-10, it is the combination of personal style and preferred thematic nuances that define each class (author).

To make this difference among the above corpora more clear, Table 2 shows the top fifteen words of each corpus with respect to their χ^2 value and a total frequency of at least five. As expected, most of these words correspond to named-entities and other topic-related information. For each word, the total term frequency *tf*, document frequency *df* (number of different documents where the word appears), and author frequency *af* (number of different authors in whose documents the word appears) are also provided. As can be seen, in CCAT-10 there are multiple words that are author-specific and tend to appear in multiple documents by that author (both *tf* and *df* are high). Thus, these words are useful indicators of authorship for that specific corpus. On the other hand, in the Guardian corpus, it is not easy to find words that appear in multiple documents of the same author and do not appear in documents by other authors (when *af* is low, *df* is also very low).

CCAT-10				**Guardian**			
Word	***tf***	***df***	***af***	**Word**	***tf***	***df***	***af***
Prague	133	74	1	dam	14	3	1
crowns	168	43	1	technologies	6	2	1
ING	41	34	2	Congo	12	2	1
PX50	39	29	1	DRC	17	2	1
Wood	27	27	1	Rwandan	25	2	1
Patria	27	24	1	speakers	7	3	2
fixing	37	27	1	theft	8	3	2
Futures	23	21	1	columnist	6	2	2
Barings	58	22	2	enriched	6	2	2
pence	70	20	1	whatsoever	6	2	2
Banka	52	18	1	combatants	7	2	2
Petr	17	17	1	Gadafy	9	2	2
Czechs	49	17	1	wellbeing	9	2	2
Grenfell	48	17	1	Libya	21	2	2
derivatives	31	16	1	allusions	6	4	1

Table 2: Top fifteen words with respect to χ^2 in CCAT-10 and Guardian corpora together with occurrence statistics.

5.2 Attribution Model

From each text of the corpus (either in its original form or in the proposed distorted view) the stylometric features are extracted (either character n-grams or token n-grams) and then a SVM classifier with a linear kernel is built. Such a simple attribution model has been extensively used in previous work and proved to be an effective approach to closed-set authorship attribution (Plakias and Stamatatos, 2008; Stamatatos, 2013; Sapkota et al., 2014; Sapkota et al., 2015).

The BNC corpus is used to estimate the most frequent words of the English language. For each model, the appropriate parameter settings for n, f_t, and k are estimated based on grid search as described in Section 4.

5.3 Results

First, we apply the examined models to the CCAT-10 corpus. Based on 10-fold cross-validation on the training corpus we estimate the best parame-

		Acc.	n	f_t	k	N
Character n-grams	Baseline	**80.6**	6	15		18,859
	DV-MA	78.2	7	35	2,000	5,426
	DV-SA	77.4	7	35	4,000	5,708
Token n-grams	Baseline	80.0	1	20		1,805
	DV-MA	79.2	2	5	4,000	10,199
	DV-SA	79.4	2	10	4,000	4,023

Table 3: Accuracy results of closed-set attribution on the CCAT-10 corpus. For each model, parameter settings (n, f_t, k) and number of features (N) are also shown.

ter settings for each model. Then, the best models are applied to the test corpus. Table 3 presents the results of this experiment. As can be seen, the baseline models are the most effective ones using both character and token *n*-gram features. However, only the DV-SA model using character *n*-grams is statistically significantly worse according to a McNemar's test with continuity correction (p<0.05) (Dietterich, 1998). For character *n*-gram features, both the baseline models and the proposed models are based on long *n*-grams (n=6 or 7), longer than usually examined in authorship attribution studies. This reflects the topic-specificity of classes in that corpus since longer character *n*-grams are better able to capture thematic information. Moreover, the proposed distortion-based models were based on high values of k confirming that thematic information is important in that corpus. As concerns the token *n*-gram features, the baseline model was based on unigrams while bigrams were selected for the proposed methods.

Next, we applied the examined models on the opinion articles of the cross-topic Guardian corpus as follows: one thematic category was used as training corpus, another was used as validation corpus (to estimate the best parameter settings) and the remaining two categories were used as test corpus. Since there are four thematic categories in total, all 12 combinations were examined and the results are shown in Table 4. Here the results favour the proposed methods. Note that the average value of k is low indicating that most thematic information is masked. For character *n*-gram features, in almost all cases the distorted view models (both DV-MA and DV-SA) outperform the baseline. In many cases the difference with respect to the baseline is very high, especially for character *n*-gram models. According to a McNemar's test with continuity correction (p<0.05) on the overall performances, DV-MA based on character *n*-grams is significantly better than all other models except the corresponding DV-SA model. The average k value of this model is very low (150) meaning that essentially only function words remain unmasked. It should be mentioned that the baseline character *n*-gram models in most cases are more effective than baseline token *n*-gram models. However, in average, they are worse than token *n*-grams due to their poor performance when the Society texts are used for training. This thematic category contains the least number of texts (Stamatatos, 2013). All examined models are based on significantly reduced feature sets in comparison to the CCAT-10 corpus indicating that in cross-topic conditions the least frequent features are not so useful.

In another experiment using the Guardian corpus, a cross-genre scenario was followed where the training and evaluation corpora come from different genres. In more detail, the book reviews category of that corpus was used as training corpus, one thematic category of opinion articles was used as validation corpus (to estimate the best parameter settings) and the remaining three thematic categories of opinion articles were used as test corpus. Again, all 4 combinations were examined (each time using a different validation corpus). Note that since the training and validation corpus belong to different genres we expect the attribution models to capture the cross-genre variation. What makes this experiment challenging is again the cross-topic variation since validation and test corpora do not share thematic properties. Table 5 presents the results for all tested models. Again, the proposed distortion-based models perform much better in comparison to the baseline models. In terms of overall performance, a McNemar's test shows that DV-MA using character *n*-grams is significantly better (p<0.05) than the rest of the models. Note also that the feature set size for most models in this experiment is further reduced in comparison to the previous experiment.

Tables 4 and 5 also show the average values of best parameter settings for the experiments related with the Guardian corpus. In comparison to the CCAT-10 corpus (Table 3), we see that shorter character *n*-grams are used for the Guardian corpus while parameter k is much smaller. All these reflect the cross-topic nature of this corpus. Note that the proposed method is able to take advantage of this fact by masking topic-specific information.

			Character *n*-grams			Token *n*-grams		
Train	**Valid.**	**Test**	**Baseline**	**DV-MA**	**DV-SA**	**Baseline**	**DV-MA**	**DV-SA**
P	S	U&W	83.1	86.0	**87.4**	78.3	77.8	79.2
P	U	S&W	84.4	**89.9**	89.4	73.7	82.7	81.0
P	W	S&U	**88.8**	**88.8**	85.5	83.6	85.5	84.9
S	P	U&W	34.0	44.4	46.4	**48.5**	**48.5**	**48.5**
S	U	P&W	32.1	47.7	47.2	**49.1**	45.8	**49.1**
S	W	P&U	35.4	48.8	49.0	50.5	**53.6**	51.6
U	P	S&W	69.8	**80.7**	69.3	68.7	72.1	65.4
U	S	P&W	71.6	69.1	**72.5**	67.7	67.2	65.1
U	W	P&S	76.4	**82.2**	79.3	75.9	70.7	75.9
W	P	S&U	71.7	**84.9**	82.9	71.1	80.9	78.9
W	S	P&U	70.8	87.8	**88.6**	68.3	77.2	79.2
W	U	P&S	76.4	**91.0**	90.8	73.6	85.1	82.8
Average Accuracy			66.2	**75.1**	74.0	67.4	70.6	70.1
		n	3.8	4.1	4.1	1.2	1.1	1.4
		f_t	33.3	27.5	30.4	32.9	30.4	30.4
		k		541.7	283.3		425	425
		N	3,665.5	1,649.8	1,722.8	528.7	382.6	403.7

Table 4: Accuracy results of closed-set attribution on the Guardian corpus in cross-topic conditions. P, S, U, and W correspond to Politics, Society, UK, and World thematic categories. In each row, best performance is in boldface. Average parameter settings and average number of features (N) are also given.

			Character *n*-grams			Token *n*-grams		
Train	**Valid.**	**Test**	**Baseline**	**DV-MA**	**DV-SA**	**Baseline**	**DV-MA**	**DV-SA**
B	P	S&U&W	38.3	**60.1**	58.1	43.9	49.8	51.4
B	S	P&U&W	41.0	**57.7**	52.0	47.7	49.7	50.0
B	U	P&S&W	39.3	**57.4**	49.6	40.1	48.5	51.8
B	W	P&S&U	40.5	59.1	**59.5**	48.0	50.8	52.4
Average Accuracy			39.4	**58.9**	55.8	44.0	49.7	51.9
		n	4	3.8	4	1.8	1.5	2
		f_t	45	41.3	35	32.5	35	31.3
		k		150	800		1,650	775
		N	1,563.8	694	1,082.8	184.3	451.3	330.3

Table 5: Accuracy results of closed-set attribution on the Guardian corpus in cross-genre conditions. B corresponds to book reviews while P, S, U, and W correspond to Politics, Society, UK, and World thematic categories of opinion articles. In each row, best performance is in boldface. Average parameter settings and average number of features (N) are also given.

5.4 Effect of Parameter k

So far, the parameter settings of the proposed models, as well as the baseline methods, were obtained using a validation corpus. To study the effect of the newly introduced parameter k, we performed an additional experiment this time using character n-gram features with fixed $n = 4$ and $f_t = 5$ and varying k values ($k \in \{100, 200, ..., 1000, 1500, ..., 5000\}$). Similar, for baseline models we used the same fixed n and f_t values. Figure 1 shows the performance of DV-MA, DV-SA, and baseline models on the CCAT-10 and Guardian corpora. Note that the results for CCAT-10 are directly comparable to Table 3. On the other hand, for the Guardian corpus we present the average performance of all possible 12 combinations (using one thematic category as training corpus and another thematic category as test corpus). This is not directly comparable to Table 4 (where the test corpus of each case consists of two thematic categories).

As can be clearly seen in Figure 1 the effect of parameter k in DV-MA and DV-SA models is crucial. In the case of CCAT-10, performance in general increases with k reflecting the topic-specific information per author in this corpus. Despite the fact that the baseline approach is better than the proposed models in this corpus, a carefully selected k value (around 3,500) makes DV-MA equally effective to the baseline. On the other hand, in the Guardian cross-topic corpus, the proposed DV-MA and DV-SA models are better than the baseline for all examined k values. The best performance is obtained for low k and the accuracy decreases as k increases. This clearly shows that the use of topic-specific information in this corpus negatively affects the effectiveness of authorship attribution models. It is also notable that, in both corpora, the differences between DV-MA and DV-SA are not significant. However, DV-MA is slightly better than DV-SA in most of the cases.

6 Authorship Verification

6.1 Corpora

Recently, the PAN evaluation campaigns focused on the authorship verification task and several benchmark corpora were developed for this task (Stamatatos et al., 2014; Stamatatos et al., 2015). Each PAN corpus consists of a number of verification problems and each problem includes a set of known documents by the same author and exactly one document of unknown authorship. The task is to decide whether the known and the unknown documents are by the same author. In this paper, we use the following corpora:

PAN-2014: It includes 6 corpora covering 4 languages and several genres: Dutch essays (PAN14-DE), Dutch reviews (PAN14-DR), English essays (PAN14-EE), English novels (PAN14-EN), Greek newspaper articles (PAN14-GR) and Spanish newspaper articles (PAN14-SP). Each corpus is divided into training and test sets. Within each verification problem all documents belong to the same genre and fall into the same thematic area. Details of these corpora are presented in (Stamatatos et al., 2014).

PAN-2015: In this collection, within each verification problem the documents can belong to different genres and thematic areas. It includes a cross-genre corpus in Dutch (PAN15-DU), cross-topic corpora in English (PAN15-EN) and Greek (PAN15-GR) and a mixed (partially cross-topic and cross-genre) corpus in Spanish (PAN15-SP). More details of these corpora are provided in (Stamatatos et al., 2015).

6.2 Verification model

In this study, we use the authorship verification approach proposed by Potha and Stamatatos (2014). In more detail, this is a *profile-based* approach meaning that first it concatenates all available known documents and then it extracts a single representation from the resulting document (Stamatatos, 2009). The top L_k *n*-grams of the known text are then compared to the top L_u *n*-grams of the unknown text and if the similarity is above a predefined threshold the unknown text is attributed to the author of the known texts. Note that parameters L_k and L_u essentially replace f_t that was used in previous experiments. All necessary parameters for this method, including n and k for the proposed method are estimated using the training part of each PAN corpus. Moreover, the most frequent words for each language are extracted from the corresponding training corpus.

6.3 Results

Table 6 shows the performance of the authorship verification models on the 10 PAN benchmark corpora based on the area under the Receiver-Operating-Characteristic curve (AUC-ROC). This evaluation measure was also used in PAN evaluation campaigns and the presented results can be directly compared to the ones reported by PAN organizers (Stamatatos et al., 2014; Stamatatos et al., 2015). In average, the proposed distortion-based models surpass the performance of the baseline models with both character *n*-gram and token *n*-gram features. The baseline model is better only in the case of the most challenging cross-genre PAN15-DU corpus. However, its performance essentially resembles random guessing (0.5). DV-SA models seem more competitive than DV-MA in the author verification task.

Table 6 also shows the performance of DV-Opt that corresponds to the best model (either DV-MA or DV-SA using either character or token *n*-grams) that can be selected by optimizing the performance (AUC-ROC) on the training corpus, separately for each one of the 10 corpora. The practice of using different models and settings per verification corpus is common in previous work (Seidman, 2013; Jankowska et al., 2014). As can be seen in Table 6, DV-Opt is better than any other single model in average performance and a one-tailed *t*-test shows that it is significantly better (at the 5% level) than both baseline models and DV-MA using character *n*-grams. The performance of DV-Opt is directly comparable to the results of PAN participants reported in (Stamatatos et al., 2014; Stamatatos et al., 2015) since the best models are selected based on information obtained from the training corpus. The last column of Table 6 compares DV-Opt with the overall winners of PAN-2014 (Khonji and Iraqi, 2014) and PAN-2015 (Bagnall, 2015). DV-Opt achieves better results in comparison to PAN

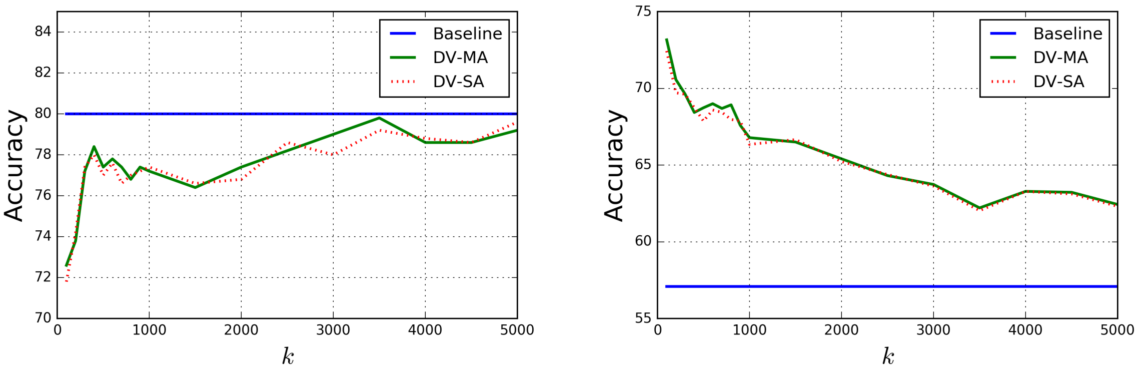

Figure 1: Performance of DV-MA, DV-SA, and baseline models based on character n-grams for fixed $n = 4$ and $f_t = 5$ and varying k values on CCAT-10 corpus (left) and the Guardian cross-topic corpus (right).

Corpus	Character n-grams			Token n-grams				Diff. PAN
	Baseline	DV-MA	DV-SA	Baseline	DV-MA	DV-SA	DV-Opt	Winner
PAN14-DE	0.975	0.961	**0.979**	0.948	0.919	0.900	0.961	+0.048
PAN14-DR	0.643	0.686	0.700	0.691	0.690	**0.704**	**0.704**	-0.032
PAN14-EE	0.528	0.591	0.582	0.626	**0.690**	0.606	**0.690**	+0.091
PAN14-EN	0.696	0.708	**0.733**	0.714	0.695	0.732	0.695	-0.055
PAN14-GR	0.625	0.783	0.779	0.794	0.838	**0.853**	0.783	-0.106
PAN14-SP	0.770	0.784	0.802	0.718	**0.838**	0.832	0.832	-0.066
PAN15-DU	**0.519**	0.470	0.509	0.247	0.433	0.505	0.509	-0.191
PAN15-EN	0.766	0.721	**0.770**	0.706	0.752	0.743	**0.770**	-0.041
PAN15-GR	0.710	**0.720**	0.706	0.672	0.674	0.646	**0.720**	-0.162
PAN15-SP	0.690	0.818	0.813	**0.852**	0.841	**0.852**	**0.852**	-0.034
Average	0.692	0.724	0.737	0.697	0.737	0.737	**0.752**	-0.055

Table 6: AUC-ROC scores of the examined authorship verification models. Last column shows the difference of DV-Opt with respect to the overall PAN-2014 and PAN-2015 winners.

winners in two corpora (PAN14-DE and PAN14-EE) while its performance is notably worse than PAN winners in PAN14-GR, PAN15-DU, and PAN15-GR. It should be underlined that the verification method used in this paper is an *intrinsic* approach while both PAN-2014 and PAN-2015 winners followed an *extrinsic* approach (where additional documents by other authors are considered in order to transform the verification problem to a binary classification task). Extrinsic models tend to perform better (Stamatatos et al., 2015).

7 Conclusions

In this paper, we presented techniques of text distortion that can significantly enhance the robustness of authorship attribution methods in challenging cases where the topic of documents by the same author varies. The proposed algorithms transform texts into a form where topic information is compressed while textual structure related to personal style is maintained. These algorithms are language-independent, do not require complicated resources, and can easily be combined with existing authorship attribution methods. Experimental results demonstrated a considerable gain in effectiveness when using the proposed models under the realistic cross-topic conditions in both closed-set attribution and author verification tasks. On the other hand, when the corpora are too topic-specific where the texts by a given author are consistently on certain subjects different than the ones of the other candidate authors, the distortion methods seem not to be helpful. Parameter k can be carefully adjusted to reflect topic properties of a given corpus.

More experiments are needed in the case of cross-genre conditions to estimate if the proposed method is also able to compress genre information and at the same time maintain properties related to personal style of authors. It would also be interesting to examine whether the distorted views of texts can be useful to other style-based text categorization tasks, including author profiling and genre detection.

References

A. Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE*, 20(5):67–75.

Shlomo Argamon, Casey Whitelaw, Paul J. Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. 2002. An experiment in authorship attribution. In *6th JADT*, pages 29–37.

Douglas Bagnall. 2015. Author Identification using multi-headed Recurrent Neural Networks. In L. Cappellato, N. Ferro, J. Gareth, and E. San Juan, editors, *Working Notes Papers of the CLEF 2015 Evaluation Labs*. CLEF and CEUR-WS.org.

Malcolm Coulthard. 2013. On admissible linguistic evidence. *Journal of Law and Policy*, XXI(2):441–466.

Olivier Y. de Vel, Alison Anderson, Malcolm Corney, and George M. Mohay. 2001. Mining email content for author identification forensics. *SIGMOD Record*, 30(4):55–64.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Hugo Jair Escalante, Thamar Solorio, and Manuel Montes-y Gomez. 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*, pages 288–298, Portland, Oregon, USA, June. Association for Computational Linguistics.

Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of Coling 2004*, pages 611–617, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Jade Goldstein-Stewart, Ransom Winder, and Roberta Sabin. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 336–344, Athens, Greece, March. Association for Computational Linguistics.

Ana Granados, Manuel Cebrián, David Camacho, and Francisco de Borja Rodríguez. 2011. Reducing the loss of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1090–1102.

Ana Granados, David Camacho, and Francisco de Borja Rodríguez. 2012. Is the contextual information relevant in text clustering by compression? *Expert Systems with Applications*, 39(10):8537–8546.

Ana Granados, Rafael Martínez, David Camacho, and Francisco de Borja Rodríguez. 2014. Improving NCD accuracy by combining document segmentation and document distortion. *Knowledge and Information Systems*, 41(1):223–245.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Steffen Hedegaard and Jakob Grue Simonsen. 2011. Lost in translation: Authorship attribution using frame semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 65–70, Portland, Oregon, USA, June. Association for Computational Linguistics.

Magdalena Jankowska, Evangelos Milios, and Vlado Keselj. 2014. Author verification using common n-gram profiles of text documents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 387–397, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Patrick Juola. 2008. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1:234–334.

Patrick Juola. 2013. How a computer program helped reveal J. K. Rowling as author of A Cuckoo's Calling. *Scientific American*.

Mike Kestemont, Kim Luyckx, Walter Daelemans, and Thomas Crombez. 2012. Cross-genre authorship verification using unmasking. *English Studies*, 93(3):340–356.

Mahmoud Khonji and Youssef Iraqi. 2014. A slightly-modified GI-based author-verifier with lots of features (ASGALF). In *CLEF 2014 Labs and Workshops, Notebook Papers*. CLEF and CEUR-WS.org.

Moshe Koppel and Shachar Seidman. 2013. Automatically identifying pseudepigraphic texts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1449–1454, Seattle, Washington, USA, October. Association for Computational Linguistics.

Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the American Society for Information Science and Technology*, 65(1):178–187.

Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8:1261–1276.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

Maarten Lambers and Cor J. Veenman. 2009. Forensic authorship attribution using compression distances to prototypes. In Zeno J.M.H. Geradts, Katrin Y. Franke, and Cor J. Veenman, editors, *Computational Forensics: Third International Workshop*, volume 5718 of *Lecture Notes in Computer Science*, pages 13–24. Springer Berlin Heidelberg.

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 (Coling 2008)*, pages 513–520, Manchester, UK, August. Coling 2008 Organizing Committee.

David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. 2005. Author identification on the large scale. In *Proceedings of the Meeting of the Classification Society of North America*.

Rohith Menon and Yejin Choi. 2011. Domain independent authorship attribution without domain adaptation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 309–315, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

George K. Mikros and Eleni K. Argiri. 2007. Investigating topic influence in authorship attribution. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN*.

Spyridon Plakias and Efstathios Stamatatos. 2008. Author identification using a tensor space representation. In *Proceedings of the 18th European Conference on Artificial Intelligence ECAI*, pages 833–834.

Nektaria Potha and Efstathios Stamatatos. 2014. A profile-based method for authorship verification. In *Artificial Intelligence: Methods and Applications - Proceedings of the 8th Hellenic Conference on AI, SETN*, pages 313–326.

Tieyun Qian, Bing Liu, Li Chen, and Zhiyong Peng. 2014. Tri-training for authorship attribution with limited training data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 345–351, Baltimore, Maryland, June. 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, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

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, Denver, Colorado, May–June. Association for Computational Linguistics.

Upendra Sapkota, Thamar Solorio, Manuel Montes, and Steven Bethard. 2016. Domain adaptation for authorship attribution: Improved structural correspondence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2226–2235, Berlin, Germany, August. Association for Computational Linguistics.

Jacques Savoy. 2013. Authorship attribution based on a probabilistic topic model. *Information Processing and Management*, 49(1):341–354.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micromessages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891, Seattle, Washington, USA, October. Association for Computational Linguistics.

Shachar Seidman. 2013. Authorship verification using the impostors method notebook for PAN at CLEF 2013. In *Working Notes for CLEF 2013 Conference*.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2014. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Benno Stein, Martin Potthast, Patrick Juola, Miguel A. Sánchez-Pérez, and Alberto Barrón-Cedeño. 2014. Overview of the author identification task at PAN 2014. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.*, pages 877–897.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. Overview of the author identification task at PAN 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

Efstathios Stamatatos. 2007. Author identification using imbalanced and limited training texts. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications*, DEXA '07, pages 237–241, Washington, DC, USA. IEEE Computer Society.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60:538–556.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21:421–439.

Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century african author. *Journal of the American Society for Information Science and Technology*, 67(1):239–242.

Michael Tschuggnall and Günther Specht. 2014. Enhancing authorship attribution by utilizing syntax tree profiles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 195–199, Gothenburg, Sweden, April. Association for Computational Linguistics.

Hans van Halteren. 2004. Linguistic profiling for authorship recognition and verification. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 199–206, Barcelona, Spain, July.