Masking Topic-related Information to Enhance Authorship Attribution

Efstathios Stamatatos

University of the Aegean

83200 - Karlovassi, Samos, Greece

stamatatos@aegean.gr

Abstract

Authorship attribution attempts to reveal the authors of documents. In recent years, research in this field has grown rapidly. However, the performance of state-of-the-art methods is heavily affected when text of known authorship and texts under investigation differ in topic and/or genre. So far, it is not clear how to quantify the personal style of authors in a way that is not affected by topic shifts or genre variations. In this paper, a set of text distortion methods are used attempting to mask topic-related information. These methods transform the input texts into a more topic-neutral form while maintaining the structure of documents associated with the personal style of author. Using a controlled corpus that includes a fine-grained range of topics and genres it is demonstrated how the proposed approach can be combined with existing authorship attribution methods to enhance their performance in very challenging tasks, especially in cross-topic attribution. We also examine cross-genre attribution and the most challenging, yet realistic, cross-topic-and-genre attribution scenarios and show how the proposed techniques should be tuned to enhance performance in these tasks. Finally, we demonstrate that there are important differences in attribution effectiveness when either conversational genres, non-conversational genres or a mix of them are considered.

Masking Topic-related Information to Enhance Authorship Attribution

Introduction

Authorship attribution (AA) is the task of revealing the author of a disputed text given a set of candidate authors and text samples by each one of them (Juola, 2008; Koppel, Schler, & Argamon, 2009; Stamatatos, 2009). This task has gained increasing popularity in recent years since it is associated with important applications mainly in (a) digital text forensics, e.g., identifying the authors of anonymous messages in extremist forums, verifying the author of threatening email messages, etc. (Abbasi & Chen, 2005; Lambers & Veenman, 2009; Coulthard, 2013), (b) humanities and historical research, e.g., unmasking the authors of novels published anonymously or under aliases, verifying the authenticity of literary works by known authors, etc. (Koppel & Seidman, 2013; Juola, 2013; Stover, Winter, Koppel, & Kestemont, 2016), and (c) social media analytics, e.g., verifying whether two reviews published by different user accounts are by the same person, explore the authenticity of tweets by a specific user, etc. (Almishari & Tsudik, 2012; Rocha et al., 2017).

The majority of published works in AA 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, Fakotakis, & Kokkinakis, 2000; Gamon, 2004; Escalante, Solorio, & Montes-y Gomez, 2011; Schwartz, Tsur, Rappoport, & Koppel, 2013; Savoy, 2013; Seroussi, Zukerman, & Bohnert, 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, Schler, & Argamon, 2011) that fits internet scale applications where anyone could be the author of the disputed text. A special case of open-set attribution is *authorship verification* where the set of candidate authors is singleton (Stamatatos et al., 2000; van Halteren, 2004; Koppel, Schler, & Bonchek-Dokow, 2007; Jankowska, Milios, & Keselj, 2014; Koppel & 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, authorship verification gained popularity in research community mainly due to the corresponding PAN shared tasks (Stamatatos et al., 2014, 2015).

In AA it is not always realistic to assume that the texts of known authorship (training texts) and the texts under investigation (test texts) 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 that share these properties with the test corpus. For instance, the text under investigation could be an email about politics while the only available texts by the candidate authors could be blogs, essays, and tweets about sports and business. A recent trend in AA research is to examine cross-domain models, meaning that the training and test corpora do not share the same properties in terms of topic and genre (Kestemont, Luyckx, Daelemans, & Crombez, 2012; Stamatatos, 2013; Sapkota, Solorio, Montes, Bethard, & Rosso, 2014; Stamatatos et al., 2015). The following cross-domain cases can be defined:

- *Cross-topic attribution*: The topic of test documents is different than the topic of training documents. All documents in both training and test corpora belong in the same genre. The training documents could be on a single topic/thematic area (Stamatatos, 2013) or a collection of topics/thematic areas (Schein, Caver, Honaker, & Martell, 2010; Sapkota et al., 2014).

- *Cross-genre attribution*: Training and test documents belong to different genres. All documents belong to the same thematic area. This is more challenging than cross-topic attribution since genre intermingles with personal style of author to affect the surface style of documents (Kestemont et al., 2012; Stamatatos, 2013).

- *Cross-topic-and-genre attribution*: The most challenging case where both topic and genre of test documents differ with respect to topic and genre of training documents.

So far, research in AA attempted to find stylometric measures and attribution methods that remain relatively stable over topic shifts and genre variations. Intuitively, function words (i.e., prepositions, articles, etc.) seem robust features in cross-topic attribution, and Goldstein-Stewart, Winder, and Sabin (2009) as well as Menon and Choi (2011) reported encouraging results. However, Mikros and Argiri (2007) demonstrated that these features are not immune to topic shifts. In addition, character *n*-grams, the most effective type of features

in AA as demonstrated in multiple previous studies (Grieve, 2007; Stamatatos, 2007; Luyckx & Daelemans, 2008; Escalante et al., 2011), found to be robust and more effective than function words in cross-topic conditions (Stamatatos, 2013; Sapkota et al., 2014). These low-level features 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 only be used as complement to other, more powerful, low-level features (van Halteren, 2004; Argamon et al., 2007; Hedegaard & Simonsen, 2011). As a result, it is not yet clear how the topic/genre factor can be reduced in the quantification of personal writing style. Moreover, it remains unclear whether cross-topic and cross-genre attribution should be handled with the same unified approach or they need separate models specifically designed to deal with their individual characteristics.

Text distortion has been used successfully in Granados, Cebrián, Camacho, and de Borja Rodríguez (2011) and Granados, Camacho, and de Borja Rodríguez (2012) to enhance thematic text clustering by masking the occurrences of frequent words while maintaining the textual structure. That way, the clustering model is no longer confused by non relevant information hidden in the produced distorted text (Granados, Martínez, Camacho, & de Borja Rodríguez, 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. Inspired by these findings, Stamatatos (2017) proposed the use of text distortion techniques in AA so that topic-related information is compressed. The main idea 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. In this paper, we extend this study and show that this distorted view of text when combined with existing AA methods can significantly improve their effectiveness in challenging cross-domain conditions.

The main contributions of this paper are listed below:

- A set of text distortion techniques (two of them newly introduced) are applied to AA tasks. The proposed techniques achieve to significantly enhance the effectiveness of existing AA methods in challenging conditions.

- To the best of our knowledge, we systematically examine for the first time the most challenging case in AA, that is cross-topic-and-genre attribution.

- We demonstrate how the proposed text distortion techniques can be appropriately tuned to handle cross-topic, cross-genre, or cross-topic-and-genre attribution.

- We show how the proposed text distortion techniques can be combined with two existing AA methods, an instance-based approach and a profile-based approach (Stamatatos, 2009).

- The differences between conversational and non-conversational genres are studied when performing cross-topic or cross-topic-and-genre attribution.

In comparison to the conference paper that introduced the main idea of applying text distortion in AA (Stamatatos, 2017), the current study has the following important differences:

- A richer set of text distortion techniques is examined.

- We combine the proposed text distortion techniques with two closed-set attribution methods (rather than a single one).

- A controlled corpus is used where topic, genre, and demographics of authors are very specifically controlled (rather than using corpora of edited journalistic texts in coarse-grained thematic categories).

- We focus on cross-topic attribution where the training corpus comprises texts on several topics (rather than a single general thematic area).

- Cross-genre and the most challenging case of cross-topic-and-genre attribution are studied.

- A richer set of genres is examined and differences between conversational and non-conversational genres are discussed.

Previous Work

In this section, we focus on previous work in cross-domain AA. Comprehensive general surveys on authorship attribution are provided by Juola (2008), Koppel et al. (2009), Stamatatos (2009), and Rocha et al. (2017). In addition, recent studies in authorship verification are summarized in Stamatatos et al. (2014, 2015).

Cross-domain AA studies are limited due to lack of appropriate corpora that should consist of texts in multiple topics and genres by the same authors. To obtain such corpora, one basic direction is the use of literature works (Koppel et al., 2007; Kestemont et al., 2012) or scientific books (Menon & Choi, 2011). Another option is to use journalistic texts, like newswire stories (Mikros & Argiri, 2007) or newspaper articles and book reviews (Stamatatos, 2013). Yet another option is to use emails (de Vel, Anderson, Corney, & Mohay, 2001) or social media texts (Madigan et al., 2005). In all these cases, topic is mainly characterized by coarse-grained thematic areas. Finally, some corpora that control topic and genre have been built (by hiring people for writing under controlled conditions) to explore cross-domain attribution (Baayen, van Halteren, Neijt, & Tweedie, 2002; Goldstein-Stewart et al., 2009). The latter approach is certainly the most reliable providing a fine-grained range of topics and genres. However, controlled corpora are limited in volume, e.g., providing only one document per topic and genre by a certain author.

A few cross-domain AA studies aim to explore whether a certain AA method performs well in challenging conditions. An early cross-topic work by de Vel et al. (2001) showed that an AA method specifically designed for email messages was not affected too much when the training and test messages were on different topics using a very small corpus (3 authors and 3 topics). The *unmasking* method for author verification of long documents was successfully tested in cross-topic conditions by Koppel et al. (2007). Later, Kestemont et al. (2012) demonstrated that the reliability of this method was significantly lower in cross-genre conditions when both prose and theatrical works were considered. Sapkota et al. (2014) explored the performance of several baseline AA methods in cross-topic conditions and found that combining several topics in the training set seems to enhance the ability to identify the authors of texts on another topic.

Most cross-domain AA studies focus on the examination of reliable stylometric features that are not affected by topic shifts and genre differences. Madigan et al. (2005) demonstrated that part-of-speech features are more effective than word unigrams in cross-topic 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 & 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 AA are punctuation mark frequencies (Baayen et al., 2002), LIWC features (Goldstein-Stewart et al., 2009), and character *n*-grams (Stamatatos, 2013; Sapkota et al., 2014). To enhance the performance of attribution models based on character *n*-gram features, Sapkota, Bethard, Montes, and Solorio (2015) define several *n*-gram categories and then they combine *n*-grams that correspond to word affixes and punctuation marks. More recently, Sapkota, Solorio, Montes, and Bethard (2016) proposed a domain adaptation model based on structural correspondence learning and punctuation-based character *n*-grams as pivot features.

One main finding of previous studies in cross-domain AA is that low-level features (like character *n*-grams) are more effective than more sophisticated features based on syntactic analysis of texts (Sapkota et al., 2014, 2015). It has also been demonstrated that a simple AA method when appropriately tuned can be effective in cross-domain conditions (Stamatatos, 2013). However, it is not yet possible to extract a text representation scheme that is not affected significantly by topic shifts and genre differences. Moreover, it is not clear whether topic and genre differences should be handled by the same way or by separate approaches (that should be determined).

Text Distortion Methods for Authorship Attribution

The main idea of the proposed approach is to transform the input texts (in both training and test corpora) before being processed by an AA method so that topic-related information is compressed. To achieve this, occurrences of not-so-frequent words are replaced by symbols. Moreover, numbers are transformed to a format that captures their structure while hiding their specific value. All these transformations maintain the structure of text (including

capitalization and punctuation mark usage) that is more likely to be associated with the personal style of the author. Actually, we apply text distortion methods that mask information mainly associated with topic preferences. Hence, we call the transformed text, a *distorted view* (DV) of the original text.

In more detail, a list of the most frequent words of the language of input texts is needed. Such a list can be obtained by a general large corpus or it can be estimated by the training corpus. The texts are tokenized and the occurrences of words in this list will remain intact. The occurrences of all other words will be masked according to a specific text distortion technique. Moreover, numbers will be transformed into a general form hiding their specific value. Let W_k be the list of k most frequent words of the language. We study the following text distortion methods:

- DV-MA: Every word not included in W_k is masked by replacing each of its characters with an asterisk (*). Every digit in the text is replaced by the symbol #. We call this technique *Distorted View - Multiple Asterisks* (DV-MA) (Stamatatos, 2017).

- DV-SA: Every word not included in W_k is masked by replacing each word occurrence with a single asterisk (*). Every sequence of digits in the text is replaced by a single symbol #. We call this technique *Distorted View - Single Asterisk* (DV-SA). Note that this technique reduces the length of texts (Stamatatos, 2017).

- DV-EX: Every word not included in W_k is masked by replacing each of its interior characters with an asterisk (*). As a result the pair of exterior characters remain intact. Every digit in the text is replaced by the symbol #. We call this technique *Distorted View - Exterior Characters* (DV-EX).

- DV-L2: Every word not included in W_k is masked by replacing each of its characters with an asterisk (*) except the last two characters. Every digit in the text is replaced by the symbol #. We call this technique *Distorted View - Last 2 Characters* (DV-L2).

An example of transforming a sentence according to the above text distortion techniques is provided in Table 1 using as W_k the 300 most frequent words of the British National Corpus

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=300	The **** , ******** ******* than the **** ****** and ******** to be ****** in the $## , ### - $## , ### ***** , could help ** ****** a ******** ***** of the *** - **** *** market , a ******* it **** ********* .
DV-SA, k=300	The * , * * than the * * and * to be * in the $# , # - $# , # * , could help * * a * * of the * - * * market , a * it * * .
DV-EX, k=300	The c**s , s******y s*****r than the F**d T****s and e******d to be p****d in the $## , ### - $## , ### r***e , could help GM r****n a s******e p***e of the m*d - s**e c*r market , a s*****t it o**e d*******d .
DV-L2, k=300	The **rs , ******ly *****er than the **rd ****us and ******ed to be ****ed in the $## , ### - $## , ### ***ge , could help GM ****in a ******le ***ce of the *id - **ze *ar market , a *****nt it **ce *******ed .

Table 1
An example of transforming a sentence according to the text distortion techniques proposed in this study.

(BNC)[1] that largely correspond to function words. As can be seen, each technique provides a distorted view of the input text where the textual structure is maintained but some, mainly thematically related, information is masked.

DV-MA is inspired by the text distortion method described in Granados et al. (2011). In comparison to that method, the proposed approach is more suitable for AA tasks. More specifically, the main differences with the approach of Granados et al. (2011) are following:

- We replace the occurrences of the least frequent words rather than the most frequent words since it is well known that function words provide important stylometric information.

- Punctuation marks and other symbols are maintained since they are important style markers.

- Capitalization of original text is maintained because it is a stylistic choice.

- 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 Table1 both $15,000 and $17,000 are transformed to the same pattern.

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

Thus, the proposed methods are able to capture the format used by the author and discard the non-relevant information about the exact numbers.

DV-SA is a modification of DV-MA where token (word or number) length information is lost. Thus, a comparison of DV-MA and DV-SA performance will show whether length of tokens contribute to enhance attribution effectiveness. DV-EX is inspired by psychological studies indicating that exterior letters are more important than interior letters in sentence reading (Jordan, Thomas, Patching, & Scott-Brown, 2003; Johnson & Eisler, 2012). Finally, DV-L2 is an attempt to maintain word suffixes that usually indicate morpho-syntactic information (e.g., tense, number, part-of-speech, etc.). Affix information has been found to be important in previous AA studies (Sapkota et al., 2015).

The value of parameter k is crucial to mask/reveal information of input texts. Table 2 shows an example sentence and its DV-MA distorted views corresponding to several values of k, using again the top k most frequent words of BNC as W_k. In the extreme case where k=0 all words are masked and the only information left concerns word-length, punctuation marks and numbers usage. When k=100, some 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.

Corpus

In this study, we use the corpus introduced in Goldstein-Stewart et al. (2009). This is a controlled corpus in terms of genre, topic and demographics of subjects. It includes samples by 21 undergraduate students covering six genres (blog, email, essay, chat, discussion, and interview) and six topics (church, gay marriage, privacy rights, legalization of marijuana, war in Iraq, gender discrimination) in English. To ensure that the produced samples are on the same specific aspect of the topic, a short question was given to subjects (e.g., Do you think the Catholic Church needs to change its ways to adapt to life in the 21th Century?). In two genres

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-MA, k=5,000	The cars , slightly smaller than the Ford ****** and expected to be ****** in the $## , ### - $## , ### range , could help ** ****** a ******** piece of the *** - size car market , a ******* it once dominated .

Table 2

An example of transforming a sentence according DV-MA and varying values of parameter k.

(discussion and interview) the samples were audio recordings and they have been transcribed into text as accurately as possible maintaining information about pauses, laughs etc. For each person, there is exactly one sample for each combination of genre and topic, that is 36 samples in total. More details about the construction of this corpus are provided in Goldstein-Stewart et al. (2009).

In Table 3 some properties of the genres included in the corpus are presented. Note that email is considered a non conversational genre since each email sample in this corpus contains a single message. On the other hand, a chat, discussion, or interview sample contains a sequence of utterances by a person produced in a dialogue with another person. It is important to notice that samples of conversational genres in this corpus tend to be longer than non conversational samples.

This corpus provides certain advantages to evaluate AA methods. First, since it supports a fine-grained control over genre and topic, it allows the evaluation of AA methods in cross-topic and/or cross-genre settings ensuring that other factors that can affect performance (e.g., demographics of authors, distribution of samples over the authors) are diminished. Moreover, it includes a wide variety of genres, so it is possible to study how AA methods are affected by genre. Given that half of the available genres are conversational, it is also possible to study how an AA method can handle conversational/non-conversational genres. On the

Genre	Mode	Conversational	Avg. Text length
blog	text	no	383
email	text	no	298
essay	text	no	519
chat	text	yes	644
discussion	speech	yes	1,256
interview	speech	yes	605

Table 3
Genre properties of the corpus used in this study.

other hand, as any controlled corpus, it provides a limited number of samples and authors. Thus, it is not possible to test how a method scales up to hundreds of authors and thousands of text samples.

This corpus has been used by a couple of AA studies. First, Goldstein-Stewart et al. (2009) attempted to perform person identification experiments. For example, they reported results for predicting a person in one genre given information of other genres. However, in each experiment the evaluation text for one author concatenates all six texts by that author covering all available topics in the evaluation genre. Similarly, to predict a person in one topic given information on the rest of the topics, each evaluation text concatenates six texts by that author in all available genres on the evaluation topic. This setup is not realistic in AA applications since it requires the availability of a specific set of topics in the evaluation corpus in the case of cross-genre attribution or the availability of a specific set of genres in the evaluation corpus in the case of cross-topic attribution.

Another study by Sapkota et al. (2014) examined the effectiveness of several AA methods in cross-topic conditions. In more detail, they examined single cross-topic (i.e., where the training corpus comprises texts on one topic and the test corpus comprises texts on another topic) and multiple cross-topic conditions (i.e., where the training corpus comprises texts on several topics and the test corpus includes texts on another topic). The latter case is directly comparable with the cross-topic attribution results presented in this study.

Experiments

Experimental Setup

In order to study the effect of the proposed text distortion techniques, we use two existing AA methods and combine each one of them with a pre-processing step where a certain text distortion technique is applied. Both of the used AA methods are language-independent and based on character-level information so they do not require sophisticated (e.g. syntactic or semantic) text analysis that would be impossible in the distorted text. In more detail, we use the following AA methods:

- C3G-SVM: This method represents texts using the most frequent character 3-grams of the training corpus and then applies a SVM classifier. This simple method has produced very robust results in previous AA studies including cross-topic conditions (Stamatatos et al., 2014; Sapkota et al., 2014) and it is usually a hard-to-beat baseline. This approach follows the *instance-based paradigm* (Stamatatos, 2009) where each text sample has its own representation. In this study we use as features all character 3-grams that appear at least 5 times in the training corpus and we use a SVM classifier with a linear kernel and C=1. These settings are in accordance with previous AA studies (Stamatatos, 2013; Sapkota et al., 2015).

- PPM5: This method introduced by Teahan and Harper (2003) is based on *prediction by partial matching*, a text compression technique. First, it extracts a compression model from the texts by one author and then calculates the cross-entropy of applying the compression model to a test text. Finally, the candidate author whose model is able to better compress the test text is selected. This approach follows the *profile-based paradigm* (Stamatatos, 2009) where all available training texts by a candidate author are concatenated. It is a robust AA method as it has been demonstrated in a recent study by Potthast et al. (2016) where PPM5 achieved the best overall results on three benchmark corpora in comparison to several reproduced AA methods. In the current paper, we used a PPM-C compression model of order 5 in accordance with previous AA studies (Teahan & Harper, 2003; Rocha et al., 2017).

In each experiment the baseline refers to the case where the original AA method, without any text distortion technique, is used. In addition, this is compared with the case where a text distortion technique (either DV-MA, DV-SA, DV-EX, or DV-L2) is first applied to all texts (in both training and test corpora) as a pre-processing step and, then, the AA method (either C3G-SVM or PPM5) is applied in the distorted texts.

The proposed text distortion methods need a list of the most frequent words of the language. In this study, we used the most frequent words in English as represented in BNC. In all experiments the considered values of parameter k are {0,100,200,...,500,1000,1500,...,5000}. This range covers the case where mainly function words are considered ($k <= 500$) and cases where more topic-related words are included ($k > 500$). In each experiment, we report average classification accuracy. Note that since the corpus is balanced over authors, topics, and genres micro-average accuracy is equal to macro-average accuracy.

Cross-topic Attribution

Here we focus on cross-topic attribution where we assume that all texts are in the same genre while training and test texts differ in topic. Similar to Sapkota et al. (2014), we perform leave-one-topic-out cross-validation where all texts on a specific topic (within a certain genre) are included in the test corpus and all remaining texts on the remaining topics (in that genre) are included in the training corpus. This is repeated six times so that all available topics to serve exactly one time as the evaluation topic.

Figure 1 shows the performance of C3G-SVM baseline and the proposed AA models for varying parameter k values in all six genres. As can be seen, the performance curves follow the same pattern in all cases: there is a peak of accuracy when k is around 100-200 and then it decreases until it reaches baseline (blog, email, discussion) or it crosses baseline performance (essay, chat, interview). It is obvious that the text distortion models significantly help this AA method to become more effective in cross-topic attribution conditions. The fact that a low value of k seems to be the best choice means that a masked version of text where only function word occurrences remain intact and all other words are masked is the most

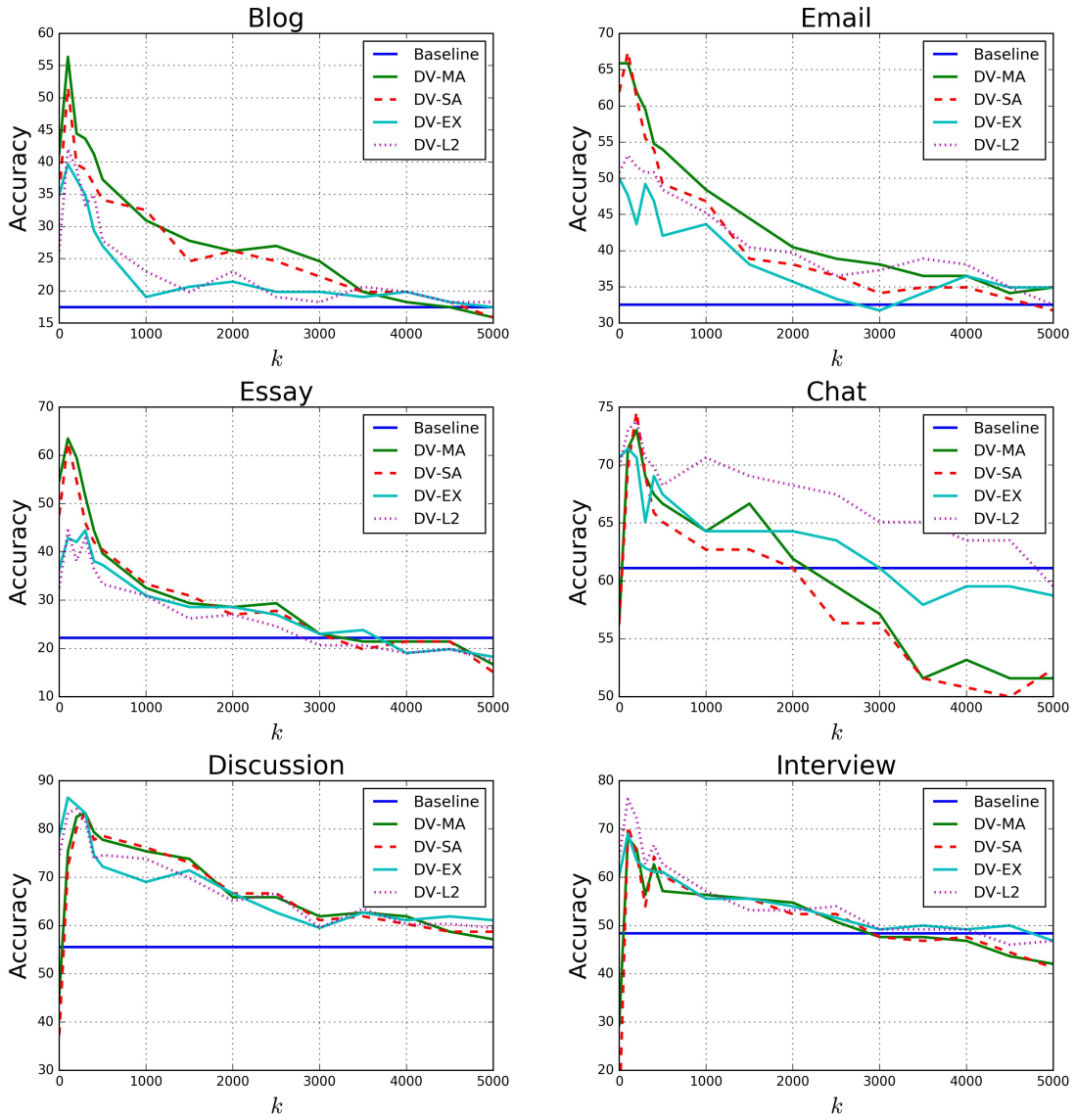

Figure 1. Cross-topic attribution performance of the proposed text distortion and baseline models for the C3G-SVM approach.

appropriate in cross-topic attribution. In other words, if all topic-relevant words are masked, the performance is enhanced significantly. As concerns the performance of individual text distorion models, it seems that DV-MA and DV-SA are more effective in non-conversational genres (blog, email, essay) while DV-EX and DV-L2 are better choices in conversational genres (chat, discussion, interview).

Figure 2 shows the respective results for PPM5 baseline and text distortion models. With the exception of blogs, the performance curves follow a basic pattern: they peak at very low k values (0-100) where they surpass baseline and then they suddenly decrease and become worse than the baseline. In the case of three genres (essay, chat, interview), there is a second

lower peak in high values of k (1,000-3,000) while in the rest of the cases the performance of text distortion models remains below the baseline while k increases. These curves mean that the performance of PPM5 in cross-topic attribution conditions is significantly enhanced when all (k=0) or almost all (k=100) word occurrences are masked and the text is stripped of any topic-relevant information. The second peak in performance observed in some genres when 500$< k <$3,500 could mean that PPM5 models can take advantage of author's traits to use words that are not strictly associated with certain topics (e.g. *decisions, opinion, aspects*). Again, with the exception of blogs, DV-MA and DV-SA models seem to be more effective in non-conversational genres while DV-EX and DV-L2 are better in conversational genres. Since texts in conversational genres of the corpus are relatively longer than the non-conversational ones, this could mean that DV-MA and DV-SA are better choices when limited text is available. Another likely explanation is that affix information (captured by DV-EX and DV-L2 models) is more important in conversational genres.

In order to compare the performance of the proposed models with the AA methods used by Sapkota et al. (2014) in the same corpus with the same experimental setup, we have to automatically determine the value of parameter k. To this direction, we perform grid search and leave-one-topic-out cross-validation in the training corpus. For example, in the blog genre, when the Church topic is left out for evaluation, we perform leave-one-topic-out cross-validation in the training corpus (consisting of blog texts on the rest of the topics) and examine $k \in$ {0,100,200,...,500,1000,1500,...,5000} to find the most appropriate value of k for this case. This k value is then used in the experiment that predicts the authors of blog texts about Church. Note that this process attempts to tune k for *another topic* rather than for the Church topic specifically, since texts on that topic are excluded (to obtain a fair comparison with other methods). We repeat this process for all 6 topics of blogs and for all 36 cases of all genres. The accuracy results of the proposed models with the automatically extracted values of k are given in Table 4. Each column presents the average cross-topic attribution performance for the corresponding genre and the last column is the overall average performance for all genres. The performance of baseline models as well results reported by Sapkota et al. (2014) are also shown. In addition, we applied the typed *n*-grams approach

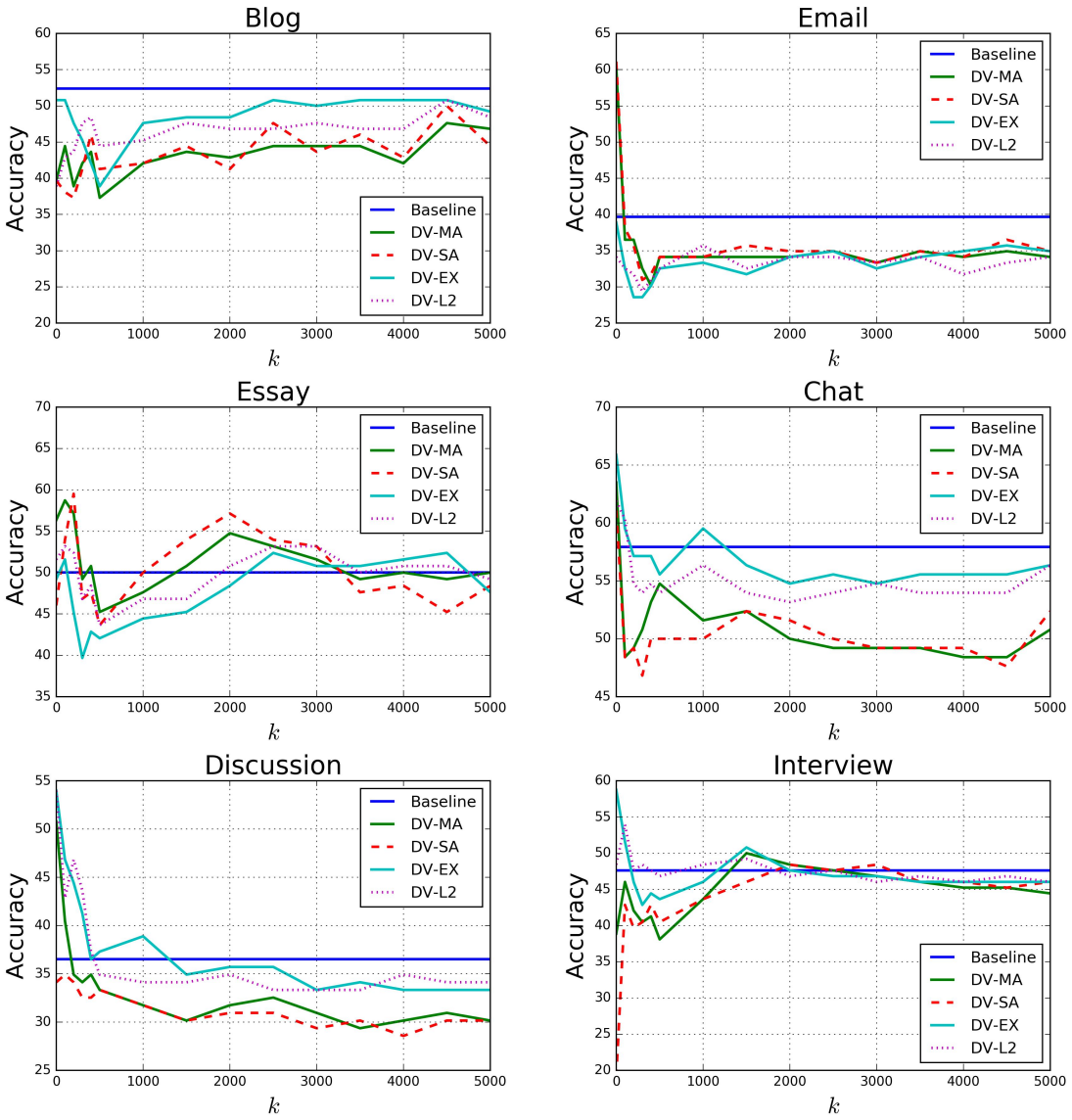

Figure 2. Cross-topic attribution performance of the proposed text distortion and baseline models for the PPM5 approach.

described in Sapkota et al. (2015) focused on *affix+punct* *n*-grams that essentially contain character *n*-grams that are word prefixes or suffixes, or they include punctuation marks.

The best result (average performance over all genres) of the methods examined by Sapkota et al. (2014) was 45.05%. This is the best performance for that specific corpus found in relevant literature with similar experimental settings. Typed *n*-grams were not able to improve this result while the PPM5 baseline model presented in the current study is slightly better (47.35%). These relatively low accuracy results demonstrate the difficulty of the cross-attribution task. Note also that PPM5 baseline is more robust than C3G-SVM baseline since its performance is more balanced over the genres. The proposed models that mask

Method	Blog	Email	Essay	Chat	Disc.	Interv.	Average
(From Sapkota et al. (2014))							
Lexical Features	13.15	11.87	12.64	33.02	25.26	20.95	19.48
Stylistic Features	15.67	33.12	23.36	37.62	24.33	23.49	26.27
Stopwords	20.43	24.97	22.06	33.49	32.59	38.89	28.74
Character n-grams	33.41	36.53	36.66	57.46	49.91	56.35	45.05
(Produced in the current study)							
Typed *n*-grams	19.84	36.51	26.19	65.87	56.35	53.17	42.99
C3G-SVM(Baseline)	17.46	32.54	22.22	61.11	55.56	48.41	39.55
C3G-SVM(DV-MA)	43.65	**65.87**	**60.32**	71.43	80.16	67.46	**64.81**
C3G-SVM(DV-SA)	51.59	61.90	54.76	71.43	78.57	68.25	64.42
C3G-SVM(DV-EX)	39.68	49.21	42.86	72.22	**84.13**	67.46	59.26
C3G-SVM(DV-L2)	37.30	51.59	45.24	**74.60**	82.54	**74.60**	60.98
PPM5(Baseline)	**52.38**	39.68	50.00	57.94	36.51	47.62	47.35
PPM5(DV-MA)	45.24	60.32	57.14	63.49	50.79	42.86	53.31
PPM5(DV-SA)	48.41	61.11	59.52	61.11	33.33	44.44	51.32
PPM5(DV-EX)	48.41	38.89	50.00	65.87	53.17	58.73	52.51
PPM5(DV-L2)	44.44	30.16	49.21	61.90	53.97	49.21	48.15

Table 4

Average performance results (per genre and overall) for cross-topic attribution.

topic-relevant information enhance the attribution accuracy considerably. All four text distortion models achieve results better than the Character *n*-gram approach by Sapkota et al. (2014) as well as PPM5 baseline. The best PPM5 model reaches 53.31% while the best obtained result for C3G-SVM models is 64.81%. In both cases, DV-MA models achieve the highest average accuracy over all genres.

As concerns performance in individual genres, with the exception of blogs, where PPM5 baseline obtained the best result, the presented text distortion models enhance the attribution performance by a large margin. It is noticeable that DV-MA and DV-SA are more effective in non-conversational genres (of shorter texts) while DV-EX and DV-L2 are more accurate in conversational genres (of longer texts).

A non-parametric Friedman test and a Nemenyi post-hoc test were performed to examine the statistical significance of differences in performance of the examined models and the results are depicted in Figure 3 (Demšar, 2006). For any pair of models the difference in their performance is statistically significant in case their scores differ more than the critical distance (CD) that was obtained for $p < 0.05$. There is no significant difference among the four C3G-SVM models based on text distortion. On the other hand, all but one (DV-EX)

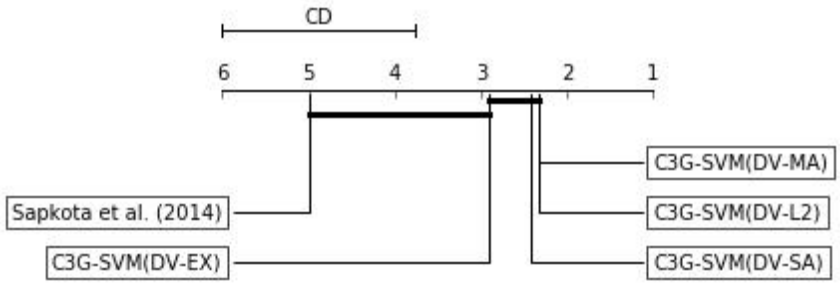

Figure 3. Results of the Nemenyi post-hoc test for C3G-SVM and baseline models in cross-topic AA.

distortion techniques are significantly better than the best model reported in previous studies for that corpus (the character *n*-gram model of Sapkota et al. (2014)). In contrast, none of the four PPM5 models (not depicted in Figure 3) was found significantly better than that baseline.

Cross-genre Attribution

Here we assume that all texts are on the same topic while training and test texts differ in genre. Given that the surface style of text is affected by both the personal style of author and genre, cross-genre attribution is more challenging than cross-topic attribution. Moreover, the proposed text distortion methods are designed to mask mainly topic information and it is unclear whether they are going to be effective in cross-genre conditions. In each experiment, we use the texts on a specific topic (e.g. Church) and perform leave-one-genre-out cross-validation. Again, C3G-SVM and PPM5 models are examined and the results are demonstrated in Figure 4. Since the performance curves for each topic are quite similar, we present the average performance over all topics. As can be seen, the performance of C3G-SVM is improved when k is set to high values (greater than 2,500). This practically means that when only the occurrences of rare words are masked the proposed models achieve to enhance cross-genre attribution performance. Recall from the previous section that for cross-topic attribution very low values of k were appropriate in most of the cases. This clearly shows that cross-topic and cross-genre attribution have different characteristics and should be handled accordingly.

Since all (both training and test) texts are on the same topic, a likely explanation of the improved performance for large values of k in this experiment is that each author approaches a

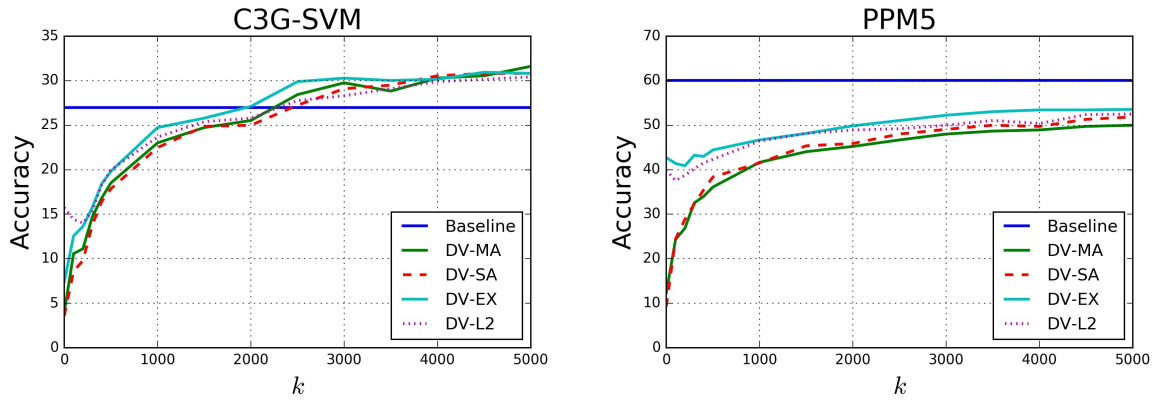

Figure 4. Cross-genre attribution performance of C3G-SVM and PPM5 models.

specific topic in a different way and this information is captured when some not-so-frequent words are retained in the text. For example, an author may think that the way the Catholic Church is presented by the media is important and they will include such comments in their texts of any genre. Thus, words related with this aspect (e.g., *media, journalists, broadcasting*) become useful clues of authorship.

In the case of PPM5, the proposed text distortion models were not able to reach the baseline performance which was exceptionally high (accuracy around 60%) given the difficulty of the task. This verifies that PPM5 is a very robust AA approach (Potthast et al., 2016; Rocha et al., 2017) and it is not confused too much by genre variations. All C3G-SVM models are much worse than PPM5 models in cross-genre attribution. Despite the relative improvement provided by applying text distortion methods, C3G-SVM is not competitive to PPM5 for this task.

As concerns, individual text distortion techniques, DV-EX is the most effective in cross-genre attribution. This model (together with DV-L2) is less affected by low values of k when the PPM5 approach is used. Moreover, DV-SA seems to be slightly better than DV-MA using PPM5 and slightly worse than DV-MA using C3G-SVM. This means that token length information is not so important in PPM5.

Cross-topic-and-genre Attribution

So far, all texts, both training and test, share either topic or genre. To make things even more challenging, we examine the case where training and test texts differ in both topic and

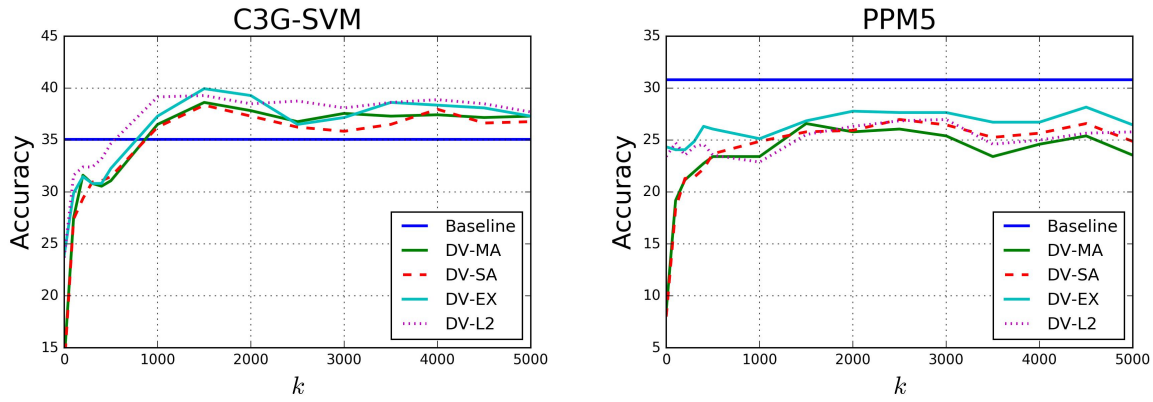

Figure 5. Cross-topic-and-genre attribution performance of C3G-SVM and PPM5 models.

genre. Using the controlled corpus, we performed leave-one-topic-and-one-genre-out cross-validation. This practically means that, in each fold, the training corpus comprises 25 texts per author (for each of 5 topics and 5 genres) while the test corpus comprises one text per author (on the remaining topic and in the remaining genre). The average performance results of the examined models for varying k values are depicted in Figure 5. Note that the performance curves have a similar form to the ones from cross-genre attribution (see Figure 4). This means that the genre factor is more significant than topic in AA. In the case of C3G-SVM attribution models, the proposed text distortion techniques achieve an enhanced performance when k>1,000. In comparison to cross-genre attribution, lower values of k are sufficient to surpass the baseline performance indicating that the topic factor is now more important.

With respect to the PPM5 attribution models, the proposed text distortion techniques were not able to reach the performance of baseline. The performance curves follow the same pattern and they seem to improve when k increases, similar to the cross-genre attribution case. Again, DV-EX and DV-L2 models are not significantly affected when k is set to very low values. Note that all C3G-SVM models (including the baseline) are significantly better than the PPM5 baseline. This means that the PPM5 approach, although robust in either cross-topic or cross-genre conditions, it fails in the most challenging case of cross-topic-and-genre attribution. Given this fact, in the following experiments we focus on C3G-SVM models.

Similar vs. Distant Genres

In the performed cross-topic-and-genre attribution experiments, the training corpus consists of a mix of genres some of them may be quite similar to the genre of the test corpus. For example, when the evaluation genre is Discussion, the training corpus includes Interview that has certain similarities with Discussion. To study how the existence of similar genres in the training corpus affect performance we conducted additional experiments taking into account similarity among genres.

First, we examine the case where the training and test genres are similar. To this end, we performed leave-one-topic-and-one-genre-out cross-validation in the set of conversational genres alone as well as the set of non-conversational genres alone. For example, we examined the case where Discussion texts about the Church topic form the test corpus while the training corpus consists of texts on the rest of 5 topics belonging in Chat and Interviews (i.e., excluding non-conversational genres). Similarly, when non-conversational genres are considered, Chat, Discussion, and Interview are excluded.

Figure 6 shows the performance of C3G-SVM models when either only conversational or non-conversational genres are considered. Note that the baseline model for non-conversational genres is far more accurate in comparison to the baseline model for conversational genres despite the fact that conversational texts are longer in this corpus. This means that this specific set of non-conversational genres is far more homogeneous in comparison to the specific set of conversational genres. Moreover, recall that two of the conversational genres in this corpus (discussion and interview) were obtained from speech transcripts. In both cases, the performance of the text distortion models start surpassing the baseline when k reaches 500. In conversational genres all text distortion models maintain a superior performance when k increases. On the other hand, in non-conversational genres, only DV-EX surpasses baseline for k greater than 2,500.

Just to make things even harder for attribution models, we finally examine the case where the training corpus comprises only conversational genres and the test corpus comprises only non-conversational genres and vice-versa. Again, topic of training and test corpora is different. This experiment is really challenging since it is a cross-topic-and-genre case where

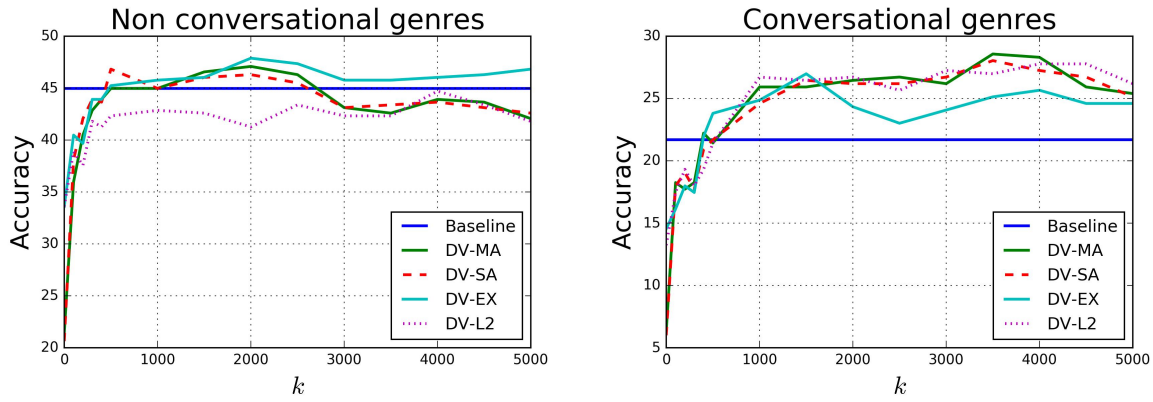

Figure 6. Cross-topic-and-genre attribution performance of the C3G-SVM models when similar (either non-conversational or conversational) genres are considered.

training and test genres have very limited (if any) similarity. We perform leave-one-topic-out cross-validation and in each fold all three genres of the test corpus (disjoint with those of the training corpus) are considered. The results of this experiment are depicted in Figure 7. The very low accuracy of baseline models in both cases (around 15%) indicates the extreme difficulty of the task. Recall that random guessing would obtain an accuracy of around 5%.

When the training corpus consists of conversational genres, the performance of the proposed text distortion models is much better than the baseline for large values of k. This is in tune with the previous results of cross-topic-and-genre experiments. On the other hand, when training corpus includes only non-conversational genres, text distortion techniques cannot help the attribution model to reach the baseline. Most probably, this can be explained by the homogeneity of non-conversational genres in the used corpus (recall the high accuracy of non-conversational genres in Figure 6). The great degree of similarities between blogs, emails, and essays in this corpus makes the classification model to overfit these genre properties and, then, when distant genres are considered, it is not effective any more. Obviously, the text distortion methods were not able to mask some of misleading information and failed to provide more robust attribution models. This problem is not present when training corpus consists of conversational genres. In that case, the classification accuracy of the best attribution models (using text distortion) reaches around 25% no matter if the test corpus comprises another conversational genre (see Figure 6) or non-conversational genres (see Figure 7).

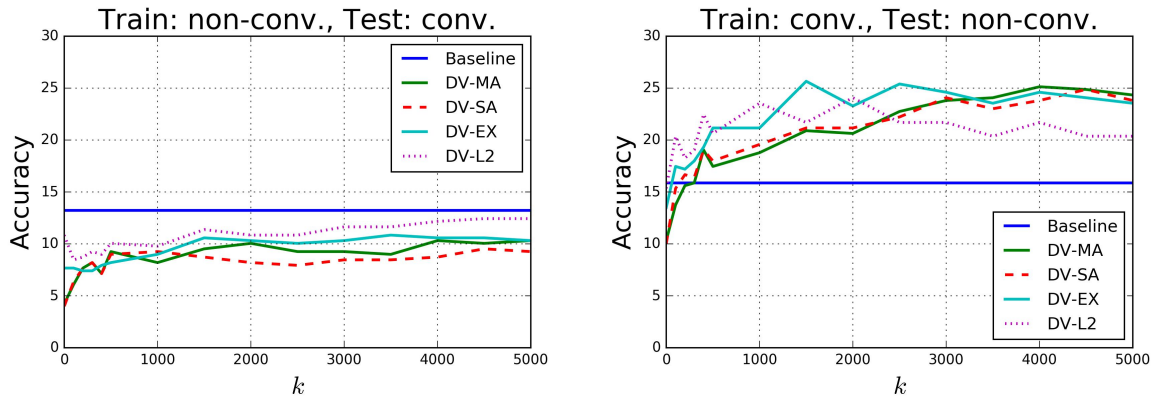

Figure 7. Cross-topic-and-genre attribution performance of the C3G-SVM models when distant genres are considered.

Conclusions

The focus of this paper is on a set of text distortion methods attempting to mask topic-related information and maintain the textual structure that is more likely to be associated with the personal style of authors. The examined methods are easy-to-use since they do not require complicated resources and essentially they are language-independent. They can be considered as a pre-processing step and can be combined with many existing AA methods. It has been demonstrated how two AA methods can take advantage of the proposed text distortion techniques. Naturally, it is not possible to apply any sophisticated (syntactic or semantic) text analysis on the distorted texts. However, the most successful methods in this field so far are based on low-level (character-level) information, like the ones used in this study.

Using a corpus that controls very specifically topic, gender and author demographics, it was possible to perform cross-domain experiments focusing on specific factors that affect attribution effectiveness and ensure that the effect of most other factors is diminished. The main findings from the conducted experiments are following:

- In cross-topic attribution, the proposed methods significantly enhance the performance of both examined AA methods. Very low values of k (0-300) are appropriate for this task meaning that only the most frequent words (mainly function words) should not be masked.

- In cross-genre attribution, the proposed methods enhance only the performance of

C3G-SVM. In contrast to the previous case, high values of k (>2,000) produced the best results. Given that all texts are on the same topic, topic-specific words provide useful clues of authorship that are not affected by genre differences. However, the proposed techniques failed to reach the performance of the PPM5 baseline that was surprisingly high given the difficulty of the task.

- In the most challenging, yet realistic, cross-topic-and-genre attribution, the form of the performance curves is similar to the ones in cross-genre experiments indicating that genre is a more significant factor than topic in AA. This sounds reasonable since genre affects surface style and is intermingled with the personal style of authors. C3G-SVM models are enhanced by using text distortion with k set to high values (>1,000) while PPM5 models are not competitive for this task.

- Another important factor in cross-topic-and-genre attribution is the degree of similarity between training and test genres. When training genres are very similar and quite distant to the test genres, the attribution model is confused and the text distortion methods are not helpful. If, on the other hand, the training genres are not very homogeneous, the text distortion methods enhance the performance of attribution models.

- As concerns the individual text distortion models, DV-MA seems to be the best choice in cross-topic attribution, especially for non-conversational genres (although there is no statistically significant difference with the rest of distortion models). The performance of DV-SA is quite similar to DV-MA meaning that token length information is not so important, or at least C3G-SVM and PPM5 fail to take advantage of it. In cross-genre and cross-topic-and-genre attribution, DV-EX is more effective in most of the cases. Finally, the performance of DV-L2 resembles that of DV-EX and actually the former is superior in cross-topic attribution of conversational genres.

In general, it seems that the proposed text distortion techniques are not equally effective for any AA method. C3G-SVM is usually enhanced while PPM5 is only enhanced for cross-topic attribution. It remains to be seen how other existing AA methods behave and, more interestingly, whether new AA methods, specifically designed to take advantage of the

more topic-neutral form of distorted views of text, will emerge. Moreover, it is interesting to study whether the proposed text distortion techniques are equally effective in other family of natural languages. It is likely that in highly inflected languages a more sophisticated tokenization procedure should be adopted where topic-neutral morphemes would be segmented (Kestemont, 2014). Another promising research direction is the application of text distortion to other style-based text categorization tasks, including cross-domain author profiling (where topic similarities in training and test corpora should not be taken for granted) and text genre detection.

References

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

Almishari, M., & Tsudik, G. (2012). Exploring linkability of user reviews. In *Computer Security, ESORICS 2012* (pp. 307–324).

Argamon, S., Whitelaw, C., Chase, P. J., Hota, S. R., Garg, N., & Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, *58*(6), 802–822.

Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. In *6th JADT* (pp. 29–37).

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

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

de Vel, O. Y., Anderson, A., Corney, M., & Mohay, G. M. (2001). Mining email content for author identification forensics. *SIGMOD Record*, *30*(4), 55–64.

Escalante, H. J., Solorio, T., & Montes-y Gomez, M. (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* (pp. 288–298).

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of COLING 2004* (pp. 611–617).

Goldstein-Stewart, J., Winder, R., & Sabin, R. (2009). Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL, EACL* (pp. 336–344).

Granados, A., Camacho, D., & de Borja Rodríguez, F. (2012). Is the contextual information relevant in text clustering by compression? *Expert Systems with Applications*, *39*(10), 8537–8546.

Granados, A., Cebrián, M., Camacho, D., & de Borja Rodríguez, F. (2011). Reducing the loss

of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering*, *23*(7), 1090–1102.

Granados, A., Martínez, R., Camacho, D., & de Borja Rodríguez, F. (2014). Improving NCD accuracy by combining document segmentation and document distortion. *Knowledge and Information Systems*, *41*(1), 223–245.

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

Hedegaard, S., & Simonsen, J. G. (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* (pp. 65–70).

Jankowska, M., Milios, E., & Keselj, V. (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* (pp. 387–397).

Johnson, R. L., & Eisler, M. E. (2012). The importance of the first and last letter in words during sentence reading. *Acta Psychologica*, *141*, 336âĂŞ-351.

Jordan, T. R., Thomas, S. M., Patching, G. R., & Scott-Brown, K. C. (2003). Assessing the importance of letter pairs in initial, exterior, and interior positions in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 883–893.

Juola, P. (2008). Authorship Attribution. *Foundations and Trends in Information Retrieval*, *1*, 234–334.

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

Kestemont, M. (2014). Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd workshop on computational linguistics for literature* (pp. 59–66).

Kestemont, M., Luyckx, K., Daelemans, W., & Crombez, T. (2012). Cross-genre authorship verification using unmasking. *English Studies*, *93*(3), 340–356.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*,

60(1), 9–26.

Koppel, M., Schler, J., & Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, *45*(1), 83-94.

Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, *8*, 1261–1276.

Koppel, M., & Seidman, S. (2013). Automatically identifying pseudepigraphic texts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1449–1454).

Koppel, M., & Winter, Y. (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.

Lambers, M., & Veenman, C. J. (2009). Forensic authorship attribution using compression distances to prototypes. In Z. J. Geradts, K. Y. Franke, & C. J. Veenman (Eds.), *Computational forensics: Third international workshop* (Vol. 5718, p. 13-24).

Luyckx, K., & Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd international conference on computational linguistics* (pp. 513–520).

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005). Author identification on the large scale. In *Proceedings of the Meeting of the Classification Society of North America.*

Menon, R., & Choi, Y. (2011). Domain independent authorship attribution without domain adaptation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 309–315).

Mikros, G. K., & Argiri, E. K. (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.*

Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J. M., … Hagen, M. (2016). Who wrote the web? revisiting influential author identification research applicable to information retrieval. In N. Ferro et al. (Eds.), *Advances in Information*

Retrieval: 38th European Conference on IR Research, ECIR (pp. 393–407).

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., ... Stamatatos, E. (2017). Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, *12*(1), 5-33.

Sapkota, U., Bethard, S., Montes, M., & Solorio, T. (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* (pp. 93–102).

Sapkota, U., Solorio, T., Montes, M., & Bethard, S. (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)* (pp. 2226–2235).

Sapkota, U., Solorio, T., Montes, M., Bethard, S., & Rosso, P. (2014). Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1228–1237).

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

Schein, A. I., Caver, J. F., Honaker, R. J., & Martell, C. H. (2010). Author attribution evaluation with novel topic cross-validation. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval - Volume 1* (pp. 206–215).

Schwartz, R., Tsur, O., Rappoport, A., & Koppel, M. (2013). Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1880–1891).

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship attribution with topic models. *Computational Linguistics*, *40*(2), 269–310.

Stamatatos, E. (2007). Author identification using imbalanced and limited training texts. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications* (pp. 237–241).

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

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

Stamatatos, E. (2017). Authorship attribution using text distortion. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL).*

Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B. (2015). Overview of the author identification task at PAN 2015. In *Working Notes of Conference and Labs of the Evaluation forum, CLEF*.

Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., … Barrón-Cedeño, A. (2014). Overview of the author identification task at PAN 2014. In *Working Notes for CLEF 2014 Conference* (pp. 877–897).

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*(4), 471–495.

Stover, J. A., Winter, Y., Koppel, M., & Kestemont, M. (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.

Teahan, W. J., & Harper, D. J. (2003). Using compression-based language models for text categorization. In W. B. Croft & J. Lafferty (Eds.), *Language modeling for information retrieval* (pp. 141–165).

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