ON THE ROBUSTNESS OF AUTHORSHIP ATTRIBUTION BASED ON CHARACTER *N*-GRAM FEATURES

*Efstathios Stamatatos**

ABSTRACT

A number of independent authorship attribution studies have demonstrated the effectiveness of character n*-gram features for representing the stylistic properties of text. However, the vast majority of these studies examined the simple case where the training and test corpora are similar in terms of genre, topic, and distribution of the texts. Hence, there are doubts whether such a simple and low-level representation is equally effective in realistic conditions where some of the above factors are not possible to remain stable. In this study, the robustness of authorship attribution based on character* n*-gram features is tested under cross-genre and cross-topic conditions. In addition, the distribution of texts over the candidate authors varies in training and test corpora to imitate real cases. Comparative results with another competitive text representation approach based on very frequent words show that character* n*-grams are better able to capture stylistic properties of text when there are significant differences among the training and test corpora. Moreover, a set of guidelines to tune an authorship attribution model according to the properties of training and test corpora is provided.*

* Assistant Professor, University of the Aegean.

I. INTRODUCTION

Authorship attribution is the line of research dealing with the identification of the author of a text under investigation given a set of candidate authors (e.g., suspects) and samples of known authorship for each one of them. Indeed, in many forensic examinations, part of the evidence refers to texts (e.g., notes, e-mail messages, SMS messages, written reports, etc.). The ability to verify that a text was written by one of the suspects could be crucial to support a case. During the last decades, significant progress has been achieved in the automation of this procedure by incorporating statistical and/or machine learning techniques (i.e., algorithms that can learn from data).[1] There is strong potential for this technology to be used as evidence in a judicial process, given that it provides effective results in well-designed experimental tests. So far, a primitive and controversial technique has been used in British courts.[2] In addition, Chaski discusses examples of the use of a semiautomated author identification method in U.S. courts.[3]

From the machine-learning point of view, authorship attribution can be viewed as a multiclass, single-label classification problem (i.e., there may be multiple suspect authors, one of whom must be selected) and can be studied

---

[1] *See* Patrick Juola, *Authorship Attribution*, 1 FOUND. & TRENDS IN INFO. RETRIEVAL 234, 235, 284–86 (2006); Moshe Koppel et al., *Computational Methods in Authorship Attribution*, 60 J. AM. SOC'Y FOR INFO. SCI. & TECH. 9, 10–13 (2009); Efstathios Stamatatos, *A Survey of Modern Authorship Attribution Methods*, 60 J. AM. SOC'Y FOR INFO. SCI. & TECH. 538, 538 (2009).

[2] R.A. Hardcastle, *CUSUM: A Credible Method for the Determination of Authorship?*, 37 J. FORENSIC SCI. SOC'Y 129, 137–38 (1997).

[3] *See* Carol E. Chaski, *Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations?*, INT'L J. DIGITAL EVIDENCE, Spring 2005, at 9, 10–11 (providing examples of cases in which the syntactic analysis method of authorship identification has been used in U.S. courts); Carol E. Chaski, *Empirical Evaluations of Language-Based Author Identification Techniques*, 8 FORENSIC LINGUISTICS 1, 1–2 (2001) (discussing the admissibility of FBI forensic stylistics methods in a federal district court case).

along the lines of other text categorization tasks.[4] However, there are some properties of authorship attribution that differentiate it from other text categorization tasks.[5] First, and perhaps most important, the stylistic choices of an author are far more difficult to capture and quantify in comparison to topic-related information. Stylistic information is usually based on very frequent patterns that are encountered in texts by the same author. On the other hand, it is preferable to focus on stylistic choices that are unconsciously made by the author and remain stable over the text length. To this end, a very large number of such features have been proposed, including measures about the length of words or sentences, vocabulary richness measures, function word frequencies, character *n*-gram[6] frequencies, and syntactic-related or even semantic-related measures.[7] In several independent studies, it has been demonstrated that function words (defined as the set of the most frequent words of the training set) and character *n*-grams are among the most effective stylometric features, though the combination of several feature types usually improves the performance of an attribution model.[8]

Practical applications of authorship attribution usually provide a limited number of samples of known authorship unevenly distributed over the candidate authors. Therefore, it is essential for the attribution model to be able to handle limited and imbalanced training sets.[9] Moreover, the availability of many samples for one candidate author does not necessarily increase the probability that the author is the true author of

---

[4] *See* Fabrizio Sebastiani, *Machine Learning in Automated Text Categorization*, ACM COMPUTING SURVEYS, Mar. 2002, at 5 (listing "author identification for literary texts of unknown or disputed authorship" as an application of text categorization).

[5] *See* Stamatatos, *supra* note 1, at 553.

[6] For example, the character 3-grams of the beginning of this footnote would be "For", "or ", "r e", " ex", etc.

[7] *See* Stamatatos, *supra* note 1, at 539–44.

[8] KIM LUYCKX, SCALABILITY ISSUES IN AUTHORSHIP ATTRIBUTION 124–26 (2010); Jack Grieve, *Quantitative Authorship Attribution: An Evaluation of Techniques*, 22 LITERARY & LINGUISTIC COMPUTING 251, 266–67 (2007).

[9] *See* Efstathios Stamatatos, *Author Identification Using Imbalanced and Limited Training Tests*, PROC. EIGHTEENTH INT'L WORKSHOP ON DATABASE & EXPERT SYS. APPLICATIONS: DEXA 2007, at 237, 237–41.

another text. This is in contrast to other text categorization tasks (e.g., thematic classification of texts) where well-represented classes have high prior probability.[10] In addition, in authorship attribution applications it is probable to have samples of known authorship on a certain thematic area (e.g., politics) while the unknown texts are on another thematic area (e.g., sports). The same can be said about the genre (e.g., known samples are scientific papers while the unknown texts are e-mail messages). In other words, in authorship attribution it is very likely to have heterogeneous training and test sets in terms of distribution of samples over the training authors, topic of texts, and genre of texts. Note that in text categorization research, it is usually assumed that the test set follows the properties of the training set.[11]

Most of the authorship attribution studies examine the simple case where the topic and genre are controlled in both the training and the test corpus.[12] While this differs from most practical applications, it aims at ensuring that the authorial style will be the crucial factor responsible for the differences among texts. In some cases, a variety of topics are covered but the

---

[10] *See* Stamatatos, *supra* note 1, at 540, 553.

[11] *See* Sebastiani, *supra* note 4, at 19.

[12] *See* Stamatatos, *supra* note 9 (addressing the problem of author identification); Moshe Koppel et al., *Authorship Attribution in the Wild*, 45 LANGUAGE RESOURCES & EVALUATION 83, 83–94 (2011) (explaining how similarity-based methods can be used with "high precision" to attribute authorship to a "set of known candidates [that is] extremely large (possibly many thousands) and might not even include the actual author"); Moshe Koppel et al., *Measuring Differentiability: Unmasking Pseudonymous Authors*, 8 J. MACHINE LEARNING RES. 1261, 1261–76 (2007) (presenting "a new learning-based method for adducing the 'depth of difference' between two example sets and offer[ing] evidence that this method solves the authorship verification problem with very high accuracy"); Efstathios Stamatatos et al., *Automatic Text Categorization in Terms of Genre and Author*, 26 COMPUTATIONAL LINGUISTICS 471, 471–95 (2000) (presenting "an approach to text categorization in terms of genre and author for Modern Greek"); Hans van Halteren et al., *New Machine Learning Methods Demonstrate the Existence of a Human Stylome*, 12 J. QUANTITATIVE LINGUISTICS 65, 65–77 (2005) (explaining how the ability to distinguish between writings of less experienced authors "implies that a stylome exists even in the general population").

same topics may be found in both the training and test set.[13] Although this setting makes sense in laboratory experiments, it is rarely the case in practical applications where usually the available texts of known authorship and the texts under investigation are completely different with respect to thematic area and genre. The control for topic and genre in training and test sets provide results that may overestimate the effectiveness of the examined models in more difficult (but realistic) cases. In a recent study,[14] the authors present a cross-genre authorship verification experiment where the well-known *unmasking* method[15] is applied on pairs of documents that belong to two different genres (e.g., prose works and theatrical plays) and the performance is considerably decreased in comparison to intragenre document pairs. In order for authorship attribution technology to be used as evidence in courts, more complicated tests should be performed to verify the robustness of this technology under realistic scenarios.

In this paper, an experimental authorship attribution study is presented where authorship attribution models based on character *n*-gram and word features are stress-tested under cross-topic and cross-genre conditions. In contrast to the vast majority of the published studies, the performed experiments better match the requirements of a realistic scenario of forensic applications where the available texts by the candidate authors (e.g., suspects) may belong to certain genres and discuss specific topics while the texts under investigation belong to other genres and are about completely different topics. We examine the case where the training set contains texts on a certain thematic area

---

[13] LUYCKX, *supra* note 8, at 96–99.

[14] Mike Kestemont et al., *Cross-Genre Authorship Verification Using Unmasking*, 93 ENG. STUD. 340, 340 (2012).

[15] *See generally* Koppel et al., *Measuring Differentiability*, *supra* note 12, at 1264 ("The intuitive idea of unmasking is to iteratively remove those features that are most useful for distinguishing between A and X and to gauge the speed with which cross-validation accuracy degrades as more features are removed. . . . [I]f A and X are by the same author, then whatever differences there are between them will be reflected in only a relatively small number of features, despite possible differences in theme, genre and the like.").

or genre while the test set includes texts on another thematic area or genre. Moreover, we make sure that the distribution of texts over the candidate authors differs in training and test sets, again to imitate realistic conditions. Two of the most successful stylometric features are tested: frequent words and character *n*-grams. Moreover, it is demonstrated that, when training and test corpora have significant differences, the most crucial decision concerns the appropriate selection of the representation dimensionality (i.e., number of features). Based on the experimental results, a set of general guidelines is provided to tune an attribution model according to specific properties of training and test corpora.

The next section compares the stylometric features we examine. Section III describes the corpus used in this study while Section IV includes the performed experiments. Finally, Section V summarizes the main conclusions and proposes future work directions.

II. FREQUENT WORDS VERSUS CHARACTER *N*-GRAMS

An intuitive way to quantify a text is based on frequencies of occurrence of words. For authorship attribution, as well as any style-based text categorization task, the most frequent words have proved to be the most useful features.[16] Interestingly, in topic-related text categorization, very frequent words (e.g., articles, prepositions, conjunctions, etc.) are usually excluded since they carry no semantic information. Hence, they are frequently called "stopwords" or function words. There are two main methods to define a set of such words to be used in an authorship attribution model: 1) using a predefined list of words belonging to specific closed-class parts of speech, such as articles, prepositions, etc.,[17] or 2) using the most frequent words

---

[16] Stamatatos, *supra* note 1, at 540.

[17] Shlomo Argamon et al., *Stylistic Text Classification Using Functional Lexical Features*, 58 J. AM. SOC'Y INFO. SCI. & TECH. 802, 803 (2007); *see also* Ahmed Abbasi & Hsinchun Chen, *Applying Authorship Analysis to Extremist Group Web Forum Messages*, IEEE INTELLIGENT SYS., Sept. 2005, at 67, 68 (focusing on the use of lexical, syntactic, structural, and content-specific features).

of the training corpus.[18] In the latter case, the top words with respect to their frequency correspond to function words. As we descend the ranked list, we encounter more and more nouns, verbs, and adjectives (possibly related with thematic choices). One disadvantage of lexical features is that they fail to capture any similarity in cases of noisy word forms (probably the result of errors in language use). For example, "stylometric" and "stilometric" are considered two different words. Another shortcoming is that in some languages, mostly East Asian ones, it is not easy to define what a word is.

Nowadays, character *n*-grams provide a standard approach to represent texts. Each text is considered as a mere sequence of characters. Then, all the overlapping sequences of *n* consecutive characters are extracted. For example, the character 3-grams of the beginning of this sentence would be "For," "or," "r e," "ex," etc. Character *n*-gram features have several important advantages: simplicity of measurement; language independence; tolerance to noise ("stylometric" and "stilometric" have many

Figure 1: An example of an online article and the extracted main text.

[18] J.F. Burrows, *Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information*, 7 LITERARY & LINGUISTIC COMPUTING 91, 91–109 (1992).

common character 3-grams); effectiveness in authorship attribution tasks, as has been proven in several studies and competitions;[19] and they require a high-dimensional representation based on information difficult to understand by humans, so deception attempts are less likely to be successful. On the other hand, the high dimensional representation requirement means that they can only be used in combination with certain classification algorithms able to support thousands of features. Furthermore, they capture small pieces of stylistic information, making the interpretation of the stylistic property of text very difficult if not impossible. Such an interpretation is crucial in case the authorship attribution technology is used as evidence in a judicial process.

Another common intuition is that character *n*-grams unavoidably capture thematic information in addition to the stylistic information. Under the assumption that all the available texts are on the same thematic area, this property of character *n*-grams can be viewed as an advantage since they provide a richer representation including preference of the authors on specific thematic-related choices of words or expressions (e.g., vehicle vs. automobile). However, when the available texts are not on the same thematic area, a topic-independent approach to represent texts, like the use of a few dozen function words, sounds more promising. In this paper we examine this assumption and show that, contrary to intuition, character *n*-grams are more robust features than frequent words when the thematic area or the genre of the texts is not controlled.

III. *THE GUARDIAN* CORPUS

The corpus used in this study is composed of texts published in *The Guardian* daily newspaper. The texts were downloaded using the publicly available API[20] and preprocessed to keep the unformatted main text.[21] An example is depicted in Table 1.

---

[19] *See* Grieve, *supra* note 8, at 259; Vlado Keselj et al., N-*Gram-Based Author Profiles for Authorship Attribution*, PROC. PAC. ASS'N FOR COMPUTATIONAL LINGUISTICS, 2003, at 255, 255–64; Stamatatos, *supra* note 1, at 538–56; Stamatatos, *supra* note 9, at 237–41.

[20] *Open Platform*, GUARDIAN, http://explorer.content.guardianapis.com/

Table 1: The Guardian corpus.

Author	Opinion articles				Book
	Politics	Society	World	UK	reviews
CB	12	4	11	14	16
GM	6	3	41	3	0
HY	8	6	35	5	3
JF	9	1	100	16	2
MK	7	0	36	3	2
MR	8	12	23	24	4
NC	30	2	9	7	5
PP	14	1	66	10	72
PT	17	36	12	5	4
RH	22	4	3	15	39
SH	100	5	5	6	2
WH	17	6	22	5	7
ZW	4	14	14	6	4
Total:	254	94	377	119	160

The majority of the corpus comprises opinion articles (comments). The newspaper describes the opinion articles using a set of tags indicating its subject. There are eight top-level tags (World, U.S., U.K., Belief, Culture, Life&Style, Politics, Society), each one of them having multiple subtags. It is possible (and very common) for an article to be described by multiple tags belonging to different main categories (e.g., a specific article may simultaneously belong to U.K., Politics, and Society). In order to have a clearer picture of the thematic area of the collected texts, we only used articles that belong to a single main category. Therefore, each article can be described by multiple tags, all of them belonging to a single main category. Moreover, articles coauthored by multiple authors were discarded.

In addition to opinion articles on several thematic areas, the presented corpus comprises a second text genre—book reviews. The book reviews are also described by a set of tags similar to the opinion articles. However, no thematic tag restriction was taken into account when collecting book reviews, since our main concern was to find texts of a specific genre that cover multiple

(last visited Mar. 2, 2013).

[21] Titles, names of authors, dates, tags, images, etc. were removed.

thematic areas. Note that since all texts come from the same newspaper, they are expected to have been edited according to the same rules, so any significant difference among the texts is not likely to be attributed to the editing process.

Table 1 shows details about *The Guardian* Corpus ("TGC"). It comprises texts from thirteen authors selected on the basis of having published texts in multiple thematic areas (Politics, Society, World, U.K.) and different genres (opinion articles and book reviews). At most 100 texts per author and category have been collected—all of them published within a decade (from 1999 to 2009). Note that the opinion article thematic areas can be divided into two pairs of low similarity, namely Politics-Society and World-U.K. In other words, the Politics texts are more likely to have some thematic similarities with World or U.K. texts than with the Society texts.

TGC provides texts on two different genres from the same set of authors. Moreover, one genre is divided into four thematic areas. Therefore, it can be used to examine authorship attribution models under cross-genre and cross-topic conditions.

IV. EXPERIMENTS

Two types of text representation features are examined—namely, words and character 3-grams. In both cases, the features are selected according to their total frequency of occurrence in the training corpus, a method proven to be suitable for authorship attribution tasks.[22] Let V be the vocabulary of the training corpus (the set of different words or character 3-grams) and $F = \{f_1, f_2, \ldots, f_i, \ldots, f_v\}$ be the set of features ordered in decreased frequency of occurrence in the training corpus. Given a predefined threshold t, the feature set F_t includes all the features with $f_i \geq t$. The higher the t, the lower the dimensionality of the representation and vice versa. Therefore, it is possible to examine different sizes of the feature

[22] John Houvardas & Efstathios Stamatatos, N-*Gram Feature Selection for Authorship Identification*, *in* ARTIFICIAL INTELLIGENCE: METHODOLOGY, SYSTEMS, AND APPLICATIONS 77, 82–84 (Jérôme Euzenat & John Domingue eds., 2006).

set by modifying t. In this study, the following frequency threshold values were used: 500, 300, 200, 100, 50, 30, 20, 10, 5, 3, 2, 1.

The well-known Support Vector Machine ("SVM") classifier[23] is used. It is a powerful classification model that can handle high dimensional and sparse data, and it is considered one of the best algorithms for text categorization tasks. The linear kernel (which is used to produce a linear boundary between the classes) is used since the dimensionality of the representation is usually high, including several hundreds or thousands of features.[24] There is no attempt to optimize the classification model by using different classification algorithms, since our aim is to highlight the capability of text representation features to remain robust in cross-topic and cross-genre conditions.

In each experiment, we follow the procedure described below:

- An attribution model is learned based on SVM and texts from a single topic category of TGC (e.g., Politics). At most, ten texts per author are used in the training phase. This provides an imbalanced training corpus.
- The learned classifier is applied to the texts of a category of TGC. Again, at most ten texts per author are used. If the selected category is Politics, that is the same as the topic category used in the training phase (intratopic attribution). The first ten texts are skipped, so there is no overlapping with the texts used in the training corpus. If the selected category is U.K., World, Society (cross-topic attribution) or Books (cross-genre attribution), then an imbalanced test corpus is compiled. Note that the distribution of the training corpus over the candidate authors is not necessarily the same with the corresponding distribution of the test corpus. This ensures that in case the attribution model favors the authors with the most training texts, it will produce many errors.

---

[23] *See* Corinna Cortes & Vladimir Vapnik, *Support-Vector Networks*, 20 MACHINE LEARNING 273, 274–75 (1995).

[24] *See* Thorsten Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, MACHINE LEARNING: ECML-98: 10TH EUR. CONF. ON MACHINE LEARNING, 1998, at 137.

A. Intratopic Attribution

In the first experiment, we examine the simplest (but unrealistic) scenario that all texts included in both training and test corpora belong to the same genre and the same thematic area. That way, the personal style of the author is more likely to be the most significant factor for discriminating between texts. Using TGC, the texts of the Politics thematic category were used for both training and test (recall, there is no overlap between training and test texts). The distribution of test texts over the candidate authors is unavoidably similar to the corresponding distribution of the training texts.

The classification accuracy results are shown in Figure 2 for models based on frequent words and character 3-grams with a varying number of features (acquired by the different values of the frequency threshold). As can be seen, the models based on character 3-grams are far more effective than models based on words and achieve perfect classification accuracy. Their performance seems to increase with the dimensionality of the representation. This indicates that even the most rare character *n*-grams carry information that help the classifier to discriminate between author choices. Since all the texts are on the same thematic area, these choices also include preferences of the authors on specific thematic-related words or phrases.

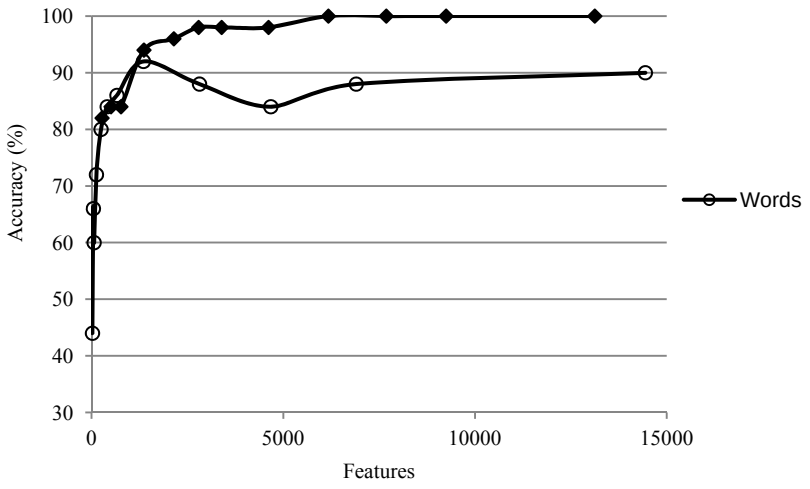

Figure 2: Performance of the intratopic attribution models (training on Politics, test on Politics).

As concerns models using word features, their performance constantly increases until about 1,500 features, then drops a little bit and then increases again. Hence, low-frequency words, probably associated with thematic-related choices, provide useful information to the classifier. In conclusion, when all the texts are controlled in terms of genre and topic, it seems that a very high dimensionality of the representation is a reliable option for both character *n*-gram and word features.

B. Cross-Topic Attribution

Next, and more interestingly, we examine the cross-topic scenario where the classifier is trained using the Politics texts and then applied to the other thematic categories (that is, Society, World, and U.K.) of the same genre. Recall that the test texts distribution over the candidate authors does not follow the corresponding distribution of the training texts. The results are shown in Figures 3, 4, and 5, respectively.

In all three cases, character 3-gram features are significantly more effective than words. When the topic of the test texts is distant with respect to training texts (i.e., Society), the performance steadily increases until about 3,500 features and then significantly drops. In the cases of thematic areas unrelated with the training texts (i.e., World and U.K.), there is a similar pattern but the performance does not drop so much when the dimensionality increases. This indicates that low frequency features found in the training corpus (usually associated with thematic information) should be avoided when the thematic area of the test corpus is distant with respect to the thematic area of the training corpus. On the other hand, these rare features are not so crucial when the thematic area of the test corpus is not specifically related to that of the training corpus. The best performance is acquired by different frequency thresholds. In the World texts the performance peak is at about 6,000 features while in the U.K. texts the peak is at about 2,500 features. Therefore, it seems that one very crucial decision in cross-topic attribution to achieve high performance is the appropriate selection of the number of features.

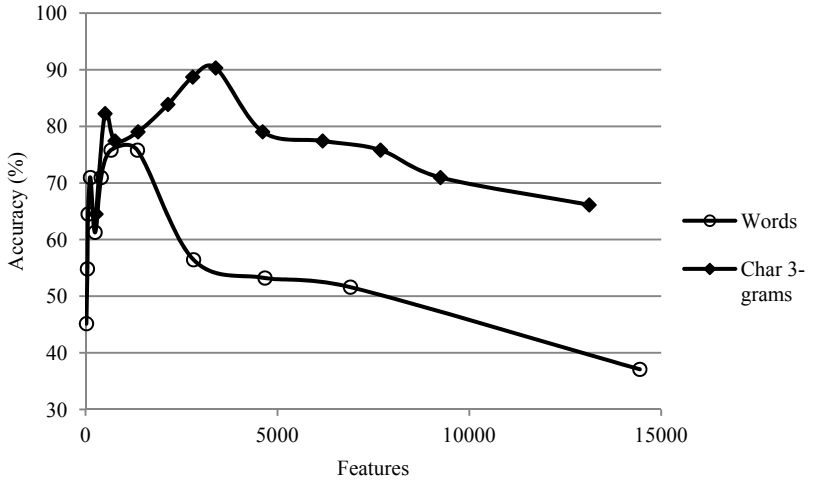

Figure 3: Performance of the cross-topic attribution models (training on Politics, test on Society).

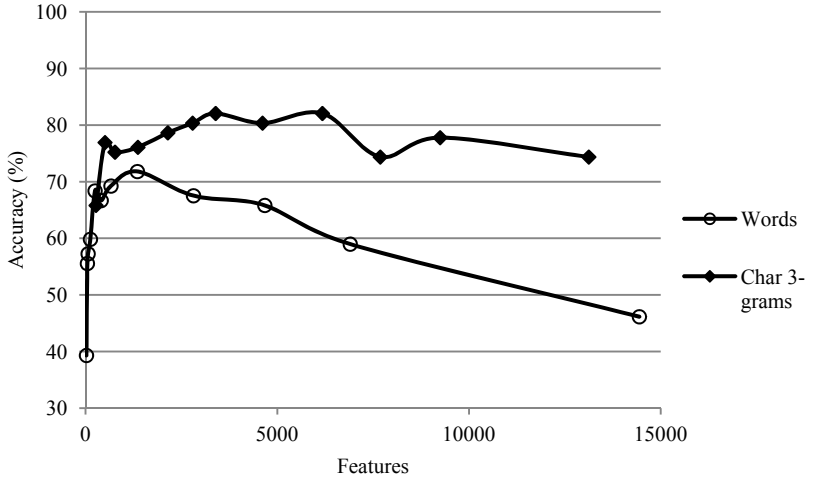

Figure 4: Performance of the cross-topic attribution models (training on Politics, test on World).

The performance of the models based on word features has similar characteristics. It steadily grows or remains practically stable until about 1,500 features and then drops significantly. The drop is much more abrupt in the case of Society texts

indicating that thematic-related words have a very negative effect when the test texts are about a topic distant from that of the training texts. In comparison to character *n*-grams, the word features are far more vulnerable by low frequency features in cross-topic conditions. Moreover, the models based on word features achieve their best performance with about 1,000 features (Society), 1,500 features (World), and 250 features (U.K.). Again, the appropriate selection of the dimensionality of the representation seems to be crucial. In comparison to character *n*-grams, word features need lower dimensionality to achieve good results in cross-topic attribution.

C. Cross-Genre Attribution

Finally, we applied the classifier learned on opinion articles about Politics to texts of another genre, book reviews. As with the cross-topic experiments, the test set is imbalanced but its distribution over the candidate authors does not follow that of the training texts. The classification accuracy results for attribution models based on word and character 3-gram features are shown in Figure 6.

Again, character *n*-gram representation seems to be far better than the word representation. The best achieved performance is lower than all the best performances for the three cross-topic experiments, indicating that cross-genre attribution is a more difficult case. However, the average performance of the cross-genre models is very close to the average performance of the cross-topic models. Another interesting point is that the best performance is achieved with considerably higher dimensionality (about 9,000 features) with respect to the best performance of the cross-topic attribution models. It seems that low frequency features, probably related to thematic information, are helpful in cross-genre conditions. Some of the book reviews included in the test corpus may refer to books about Politics. Hence, when text genre varies between training and test corpora, topic-related choices may assist the attribution model.

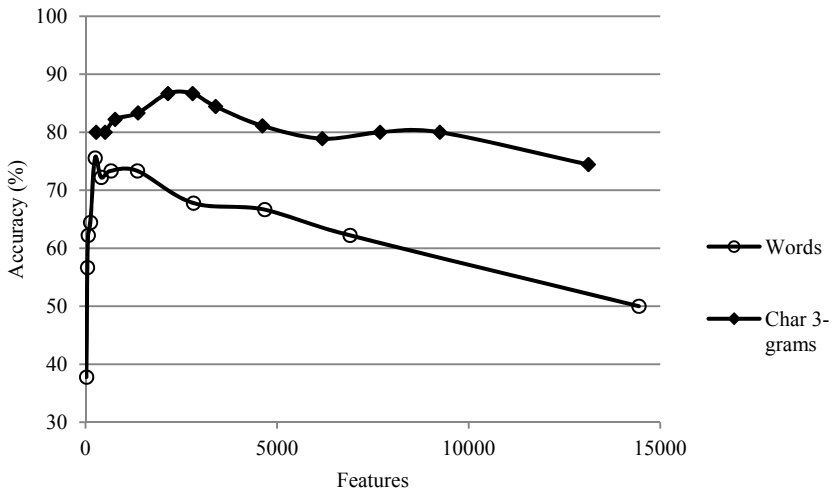

Figure 5: Performance of the cross-topic attribution models (training on Politics, test on UK).

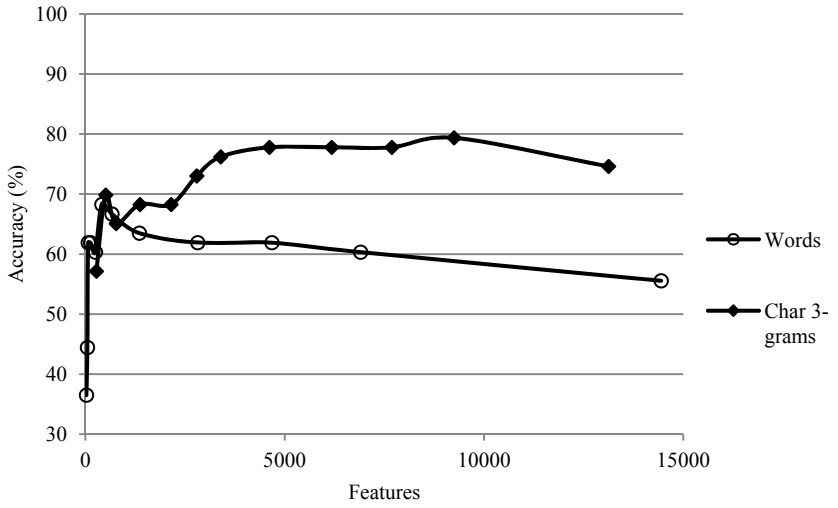

Figure 6: Performance of the cross-genre attribution models (training on Politics, test on Book reviews).

The models based on word features achieve their best performance at about 400 features, far lower than the character *n*-gram representation. However, the performance of the models based on more features does not drop dramatically as happens in cross-topic experiments. Again, this confirms the above conclusion about the usefulness of thematic-related information

in cross-genre attribution. On the other hand, the appropriate selection of the number of features is very important to achieve the best possible results.

V. DISCUSSION

One main conclusion of this study is that, in addition to the simple intratopic attribution, character *n*-grams produce models more effective and robust than those based on word features in both cross-topic and cross-genre conditions. In general, models based on words require fewer features to achieve their best results, but they are significantly inferior to the best models based on character *n*-grams. An authorship attribution model based on character 3-grams in combination with a SVM classifier with linear kernel, although simple, proves to be very effective and can be used as a baseline approach, with which every new or advanced model should be compared.

The simple scenario of intratopic (in combination with intragenre) attribution seems to be a relatively tractable problem for current technology. The performance based on both character *n*-grams and words is very high, and unlikely to be matched by human experts, even when there are multiple candidate authors and relatively short texts. However, taking into account only such cases, the accuracy of the attribution models may be overestimated.[25] The presented cross-topic and cross-genre experiments show that the performance is affected sometimes considerably when topic and genre of training and test texts are not controlled. On the other hand, in such difficult cases, if the models are fine-tuned to the appropriate dimensionality of the representation, then the classification results remain surprisingly high. Hence, in the general case of applying authorship attribution technology to real world applications, a one-model-fits-all approach is not adequate. According to the properties of the texts of known authorship and the texts under investigation, one should fine-tune the attribution models appropriately to maintain a high level of effectiveness.

---

[25] *See* LUYCKX, *supra* note 8, at 4; Kestemont et al., *supra* note 14, at 343.

Several observations from the performed experiments may be used as guidelines for tuning an attribution model:

- In intratopic attribution, a very high dimensionality of the representation is advisable. Surely, high frequency features are the most important. However, it seems that low frequency features also contribute to the discrimination ability of the model.
- In cross-topic attribution, if the topic is distant from the topic of the training texts (e.g., Politics vs. Society, World vs. U.K.), low frequency features should be avoided. Since they are closely related with nuances of thematic choices, they harm the effectiveness of the attribution models. The crucial decision is the appropriate selection of the representation dimensionality.
- In cross-topic attribution, if the topic is not specifically associated to the topic of the training texts (e.g., Politics vs. World), low frequency features are not so harmful. However, it is better to exclude them, and again there is a crucial decision about the appropriate selection of the representation dimensionality.
- In cross-genre attribution, a high representation dimensionality seems to be advisable, especially when topic similarities are likely to be found in training and test texts.

An interesting conclusion that can be drawn from this study is that cross-topic attribution where the topic of the training and test texts can be regarded as highly dissimilar (e.g., Politics vs. Society) may be more challenging than cross-genre attribution. Additionally, in cross-genre attribution, perhaps counterintuitively, models based on thousands of features (both character *n*-grams and words) are either better than or competitive with ones that use only a few hundreds of features.

Surely, more experiments are needed to verify all these conclusions. An interesting direction for future work is to explore the role of the candidate set size and how it affects the appropriate representation dimensionality. The combination of different feature types should also be examined since this approach usually improves the performance of the attribution models, as is exemplified by some of the most successful participant methods in the recently organized competitions on

authorship attribution.[26] Finally, a missing block in the authorship attribution research that is necessary to use this technology as evidence in court is the ability to explain the automatically derived decisions. In the case of attribution models based on low-level information, like character *n*-grams, that seem to be the most robust and effective approach, what is needed is a way to associate this highly dimensional information to some human interpretable high-level features.

[26] *See* Shlomo Argamon & Patrick Juola, *Overview of the International Authorship Identification Competition at PAN-2011* (Sept. 19–22, 2011), http://www.uni-weimar.de/medien/webis/research/events/pan-11/pan11-papers-final/pan11-authorship-identification/juola11-overview-of-the-authorship-identification-competition-at-pan.pdf; Patrick Juola, *An Overview of the Traditional Authorship Attribution Subtask Notebook for PAN at CLEF 2012* (Sept. 17–20, 2012), http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-papers-final/pan12-author-identification/juola12-overview-of-the-traditional-authorship-attribution-subtask.pdf.