UNIVERSALITY OF STYLISTIC TRAITS IN TEXTS

Efstathios Stamatatos

University of the Aegean

Karlovassi, Greece

e-mail: stamatatos@aegean.gr

Abstract The style of documents is an important property that can be used as discriminant factor in text mining applications. Among the great number of possible measures proposed to quantify writing style there are some features that can be characterized as universal, in the sense that they can be easily extracted from any kind of text in practically any natural language and provide accurate results when used in style-based text categorization tasks. In this paper we examine whether such universal stylometric features remain effective under difficult scenarios where the topic and/or genre of documents used in the training phase differ from that of the questioned documents. Based on a series of experiments in authorship attribution, we demonstrate that character n-gram features are reliable and effective given that the appropriate number of features is used. It is also shown that when the number of candidate authors increases, the representation dimensionality should also increase to improve classification results.

1 Introduction

Large amounts of electronic texts are produced daily, a great part of which is available online through Internet services. As a consequence, the need to handle textual information efficiently is now greater than ever. A large body of research in text mining attempts to develop methodologies and build tools performing text categorization and filtering, text clustering, text summarization, etc. (Weiss, et al., 2005) Documents can be described by several factors and properties. The most prevalent factor is their topic or theme. This can be used to build document taxonomies or filter texts according to their topic (Sebastiani, 2002). Another important property is the sentiment of texts, especially important when one attempts to handle the vast amount of opinionated texts available in social media (Pang & Lee, 2008).

Another factor that characterizes texts is their style. Although hard to define exactly what style is, there are two main aspects of style especially useful for distinguishing between texts:

- Functional style: this depends on the functional purpose of the texts, it is strongly associated with its form and the medium used to publish the texts as well as their genre and register. For example, we expect that all research papers share some stylistic choices no matter what the topic of the research is or who the author is. The same is true for e-mail messages, newspaper articles, blogs, etc. The task of attempting to exploit this type of writing style is called *genre identification* and has important applications in information retrieval and natural language processing (Meyer zu Eissen & Stein, 2004; Lim, et al., 2005; Santini, 2007; Kanaris & Stamatatos, 2009).
- Authorial style: this depends on the author of the texts and is composed of their personal use of language and idiosyncrasies. We expect that all texts written by the same author share some stylistic choices not affected by the topic and the genre of texts. A great number of research studies has been performed in *authorship attribution* (also called authorship identification) attempting to reveal the authors of anonymous documents or to solve disputed cases where several individuals claim the authorship of a certain document (Stamatatos, et al. 2000; Gamon, 2004; Luckx & Daelemans, 2008; Koppel, et al., 2007; Koppel, et al., 2011). Significant forensic applications are associated with this task (Abbasi & Chen, 2005). We also expect that groups of authors that share some demographics (i.e., age, gender, education, etc.) would have striking similarities in their personal writing style. This led to the recent development of the *author profiling* community that attempts to extract characteristics of authors from analysing their texts (Rangel, et al., 2013).

For any given document, the combination of its topic, functional and authorial style produces a unique blend that may be considered as a *document fingerprint* and can be used to identify cases of plagiarism or text reuse (i.e., when parts of one document have been used in another document) (Stamatatos, 2011). In such cases, stylistic inconsistencies may be especially useful to detect the suspicious parts of documents (Stamatatos, 2009b).

In order to be able to use writing style in the framework of a text mining application, we need to quantify it. The line of research dealing with the quantification of style is called *stylometry* and has a long history. The first pioneering studies in stylomtery, dating back to the 19th century, were based on the quite laborious task of manual counting word or letter frequencies in long documents (Mendenhall, 1887; Yule, 1944). Later on, the availability of computational systems and tools enabled researchers to use automated analysis and extract rich sets of features that can describe the style of texts. Examples of such features are vocabulary richness measures, frequencies of function words, character n-grams, part-of-speeches, etc. (Keselj, et al., 2003; Van Halteren, 2007; Luyck & Daelemans, 2008; Kanaris & Stamatatos, 2009).

In general, to achieve the best possible results, researchers use a combination of stylometric features that is suitable for a particular case (Van Halteren, 2007; Grieve, 2007). Many features are application-specific and can only be used when

the examined texts share some properties. For example, if it is known that all documents belong to a certain genre, then appropriate genre-specific measures can be defined (e.g., the use of greetings in e-mail messages) (de Vel, et al., 2001), if it is known that all documents are in the same thematic area, then appropriate topic-specific measures can be used (Zheng, et al., 2006). In addition, the language of documents may be an obstacle towards the application of sophisticated natural language processing tools able to extract syntactic or semantic-related stylometric features (Stamatatos, et al., 2000; Gamon, 2004; Argamon, et al., 2007). On the other hand, there are certain types of features that may be considered *universal* since they can be used in any case (i.e., practically all kinds of genres and natural languages).

In this paper, we examine two types of such universal stylometric features: function words and character n-grams. They can easily be extracted from every document and they have been successfully used in several style-based text categorization tasks, like authorship attribution, automatic genre identification, and plagiarism detection. Focusing on the authorship attribution task, we aim at examining the effectiveness of these features in difficult cases where there are differences in topic and genre of the documents under examination. It is demonstrated that one crucial decision concerns the representation dimensionality and it is possible to attain high accuracy results given that the appropriate number of features is used. Moreover, we show that the number of candidate authors also affects this decision, especially for character n-gram features.

The rest of the paper is organized as follows. Section 2 presents several approaches to the quantification of style. Section 3 discusses authorship attribution tasks and defines challenging scenarios for universal features while Section 4 presents experimental settings and results. Finally, Section 5 summarizes the main conclusions of this study.

2 Stylometry

There is not a consensus on the definition of style. As a consequence, a lot of different approaches have been reported in previous studies aiming at quantifying some textual properties considered to be associated with stylistic choices. An excellent review of early-stage stylometry is presented by Holmes (1998) while a more recent survey is given by Stamatatos (2009a). In this section, we describe the basic categories of stylometric features found in style-based text categorization tasks, mainly authorship attribution and genre identification.

The most commonly used type of information used in stylometric studies refers to lexical features. Text can be seen as a sequence of words grouped in sentences. Thus, word length, sentence length, number of unique words (hapax legomena), type/token ratio and other vocabulary richness measures were very popular in early-stage studies (Mendenhall, 1887; Yule, 1944). Another popular approach is to

use word frequencies, especially of very frequent closed-class words, like articles, prepositions, conjunctions, etc., also known as *function words*. Such words are very important since they are usually associated with certain syntactic structures so their frequency is an indirect measure of syntactic information. The set of function words may be predefined for all tasks (Abbasi & Chen, 2005; Argamon, et al., 2003), or specifically chosen for a given task (Mosteller & Wallace, 1964), or extracted automatically for any given task using the most frequent words of a training corpus (Burrows, 1992).

Another popular idea is to use character features. According to this approach, texts can be seen as strings of characters. Frequencies of letters, digits, punctuation marks, etc. belong to this category (de Vel, et al., 2001; Zheng, et al., 2006). Moreover, character sequences, like prefixes and suffices provide an indirect measure of lexical and syntactic information (Madigan, et al., 2005). A simple approach that has been proved very effective in many tasks is based on the set of most frequent sequences of characters, known as character n-grams (Keselj, et al., 2003; Grieve, 2007; Luyckx & Daelemans, 2008; Kanaris & Stamatatos, 2009). This method is able to capture many types of information (lexical, syntactic, formatting, etc.) and does not require complicated tools for extracting the relevant measures. Compression-based methods (using text compression algorithms as a means to measure stylistic homogeneity) also exploit information from character sequences (Benedetto, et al., 2003; Khmelev & Teahan, 2003).

In theory, syntactic structures and semantic forms provide more reliable stylistic information since they should not be affected be topical shifts and are used by the authors unconsciously (Luyckx & Daelemans, 2005; Gamon, 2004; van Halteren, 2007; Argamon, et al., 2007; Sidorov, et al., 2014). However, their use requires the availability of certain natural language processing tools than can analyse the documents within a task and provide accurate syntactic or semantic measures. Therefore, such features are language-dependent (they can be used only for languages where appropriate tools are available) and noisy (the tools make errors and the provided measures may not be 100% correct). The most popular features of this category are part-of-speech frequencies mainly because part-of-speech tagging is effectively performed by existing tools in many languages.

Another source of stylistic information is the layout or the presentation of the document. This is especially useful in genre identification since certain genres are strongly associated with specific document layouts (e.g., research papers may by multi-column with tables, graphs, etc.) In the case of web pages, such structural features can easily be extracted (e.g., HTML tag and meta-tag frequencies, image counts, use of JavaScript, number of links, etc.) and their use together with textual features increases the potential of the stylometric model (Meyer zu Eissen & Stein, 2004; Lim, et al., 2005; Santini, 2007; Kanaris & Stamatatos, 2009). However, they are not general-purpose features since the format of documents within a certain task may not provide such information.

To take advantage of certain properties of the available documents within a task, other application-specific features may be defined. Mainly, they attempt to

exploit the fact that all documents are matched for genre (e.g., the use of greetings and farewells in e-mail messages) (de Vel, et al., 2001), or topic (e.g., the use of certain topic-specific words, like deal or sale, in texts about computer sales) (Zheng, et al., 2006), or language (e.g., use of slang words in conversations) (Cristani, et al., 2012). This type of information is especially useful since it permits the stylometric model to be adapted to the properties of a specific set of documents.

In general, the combination of several types of features increases the effectiveness of the resulting model (Grieve, 2007; van Halteren, 2007; Kanaris & Stamatatos, 2009). Another idea is to attempt to use the most appropriate type of features according to the properties of a certain case (Seidman, 2013). Moreover, in the vast majority of published studies, for each feature a single measure is extracted from a text. Alternatively, distributional measures indicate how a certain feature varies within a text (Jair Escalante, et al., 2011). From another perspective, graph-based models have been proposed to capture dependencies between different features (Arun, et al., 2009).

3 Universality in Authorship Attribution

Authorship attribution may be viewed as a single-label multi-class text categorization problem. In general, we are given a set of candidate authors and for each one of them we get undisputed samples of their texts. This is the training corpus that can be used to build a model able to distinguish between the text samples of candidate authors. Then, any document of disputed authorship may be assigned to one of the candidate authors by using this model. There are three main tasks in authorship attribution:

- Closed-set attribution: where the true author of any disputed text is necessarily one of the candidate authors. This is the easiest case and most of the studies in authorship attribution follow this scenario (Stamatatos, et al., 2000; Keselj, et al., 2003; Grieve, 2007; Luyckx & Daelemans, 2008). It is appropriate for most forensic applications where police investigators are able to limit the number of suspects based on evidence about their knowledge of certain issues or their accessibility to certain resources.
- Open-set attribution: where the true author of a disputed text may not be included in the set of candidate authors. This is the most general scenario and resembles any case where it is not possible to limit the number of suspects. Previous work have shown that this task is more difficult when the set of candidate authors is small (2 or 3) rather than large (Koppel, et al. 2011).
- Author verification: where the set of candidate authors is singleton (Koppel, et al., 2007). Any authorship attribution problem, either closed-set of open-set, can be decomposed into a set of author verification cases. Therefore, the ability

to solve this problem is of crucial importance and there is increasing interest on this task recently (Koppel & Winter, 2014; Stamatatos, et al., 2014).

The vast majority of published studies in authorship attribution only consider the case where all texts in both training corpus (documents of known authorship) and evaluation corpus (documents of unknown or disputed authorship) are in the same thematic area and belong to the same genre. For many practical applications these assumptions seem reasonable. However, there are certain cases where such assumptions do not hold. For example, we can imagine one scholar attempting to verify the authenticity of a suicide note requiring the availability of other samples of suicide notes from all suspects (Chaski, 2005). It is therefore crucial, at least for certain applications, the stylometric method we use to remain effective even when the available documents are not matched for topic or genre. Thus, we define four scenarios for examining the robustness of an authorship attribution model:

- Same-topic same-genre: the simplest case, where the training texts we use to build the model and the disputed texts are in the same thematic area and belong to the same genre.
- Cross-topic same-genre: where there are differences in the topic of training and disputed texts while they all belong to the same genre.
- Same-topic cross-genre: where the training and disputed texts are in the same thematic area but differ in genre.
- Cross-topic cross-genre: where the training and disputed texts do not agree in topic and genre, certainly the most difficult case.

As already discussed, stylometric approaches based on function words and character n-grams may be considered as universal given that they can be easily applied to any kind of texts and practically all natural languages. Moreover they have produced competitive performance results in previously published studies (Keselj, et al., 2003; Koppel, et al., 2007; Grieve, 2007; Luyckx & Daelemans, 2008; Koppel, et al., 2011). However, it remains to be seen how much they are affected under cross-topic or cross-genre conditions. The remainder of this paper deals with this question.

4 Experiments

4.1 Corpus

The corpus used in this study is composed of texts published in *The Guardian* daily UK newspaper. The texts were downloaded using the publicly-available API[1]

[1] http://explorer.content.guardianapis.com/

Author	Opinion articles				Book reviews
	Politics	Society	World	UK	
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

Table 1. The corpus used in this study comprising documents in different topics and genres

and pre-processed so that to keep the unformatted main text (titles, name of authors, dates, tags, images, etc. were removed).

The opinion articles of this newspaper (comments) are described using a set of tags indicating their subject. There are 8 top-level tags (World, US, UK, Belief, Culture, Life&Style, Politics, Society) each one of them having multiple sub-tags. It is possible (and very frequent) an article to be described by tags belonging to different main categories (e.g. a specific article may belong to all UK, Politics, and Society). In order to have a clearer picture of the thematic area of the collected texts, in the presented corpus we only used articles that belong to a single main category. Therefore, each article can be described by multiple tags all of them having to belong to a single main category. Moreover, articles co-authored by multiple authors were discarded.

In addition to opinion articles on several thematic areas, this corpus comprises book reviews, a different genre. 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.

Table 1 shows details about this corpus. It comprises texts from 13 authors selected so that they have published texts in multiple thematic areas (*Politics*, *Society*, *World*, *UK*) 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 till 2009). Note that the opinion article thematic areas can be divided into two pairs of low similarity, namely *Politics*-*Society* and *World*-*UK*. In

other words, the *Politics* texts are more likely to have some thematic similarities with *World* or *UK* texts rather than with *Society* texts.

4.2 Experimental Settings

Unfortunately, it is not possible to examine all four scenarios mentioned at the end of Section 3 using the Guardian corpus. In particular, it is not possible to examine the last two (intra-topic cross-genre and cross-topic cross-genre) since there is limited information about the topic of the available book reviews. Therefore, we merge these two cross-genre scenarios into one. We focus on closed-set attribution as described in Section 3. In each author identification experiment all training texts come from a certain topic of opinion articles while all evaluation texts come either from the same topic (same-topic, same-genre), a different topic (cross-topic, same genre), or a different genre (cross-genre scenario). Each time, at most 10 training/evaluation texts per author are used. When the training and evaluation sets come from the same category (same-topic), training and evaluation texts are disjoint. Note that the training and evaluation texts are unevenly distributed among the candidate authors and this distribution varies according to topic or genre.

Two types of universal features are examined, namely words and character n-grams. In both cases, the features are selected according to their total frequency of occurrence in the training corpus. Let V be the vocabulary of the training corpus (the set of different words or character n-grams) and $F=\{f_1, f_2,\ldots, f_v\}$ be the frequency of occurrence of all possible features in the training corpus. Given a predefined threshold t, we include in the feature set all features with $f_i \geq t$. The higher the t, the lower the dimensionality and vice versa. Therefore, it is possible to examine different sizes of the feature 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. Note that for high values of frequency threshold and word features, we practically get function words only. As the frequency threshold gets lower more nouns, adjectives, and verbs are included in the list.

The well-known *support vector machines* classifier (Joachims, 1998) is used in the experiments. This model can handle high dimensional and sparse data, like the stylometric features we extract from texts, and it is considered one of the best algorithms for text categorization tasks. The linear kernel is used since the dimensionality of the representation is usually high (hundreds or thousands of features).

4.3 Results

A closed-set authorship attribution model was trained using opinion articles from the *Politics* topic category and then it was evaluated using texts from either the

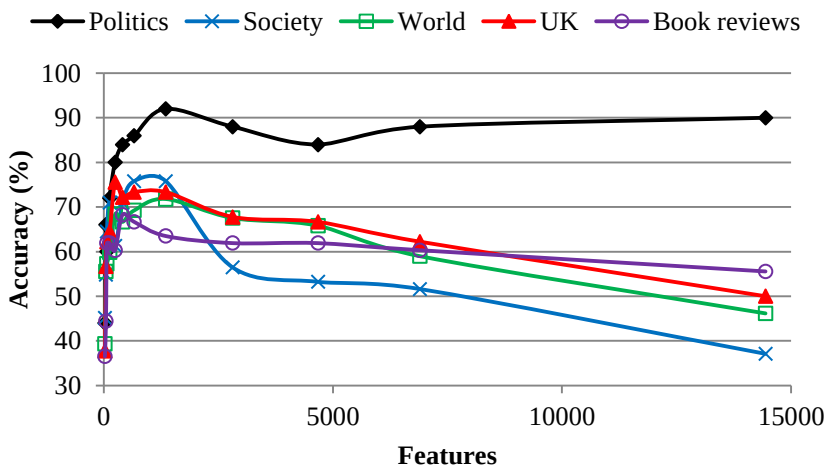

Fig. 1. Performance of the word-based attribution model trained on texts from Politics and evaluated with texts from the same topic (Politics), a different topic (Society, World, UK) or a different genre (Book reviews)

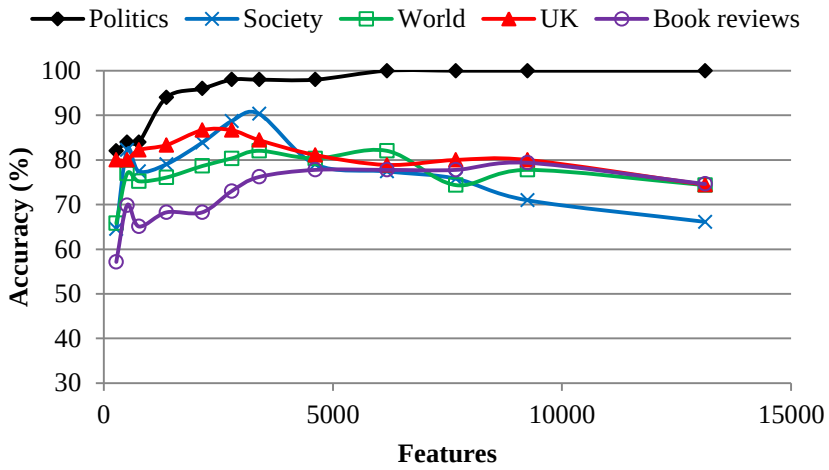

Fig. 2. Performance of the character n-gram attribution model trained on texts from Politics and evaluated with texts from the same topic (Politics), a different topic (Society, World, UK) or a different genre (Book reviews)

same genre and topic (*Politics*), the same genre but a different topic (*World* and *UK*), or a different genre (*Book reviews*). Figures 1 and 2 show the micro-average classification accuracy results of attribution models based on word and character

n-gram features, respectively, varying the representation dimensionality. It is obvious that character n-gram models are more effective than word-based models even in difficult cross-topic and cross-genre cases. In the simple case of same-topic, same-genre, character n-grams provide perfect classification results when the dimensionality is maximized. On the other hand, when topic or genre change, the appropriate selection of the dimensionality seems crucial for obtaining good performance. In those hard cases, performance drops after a certain point (about 3,000-4,000 features) when increasing the dimensionality. This drop is more dramatic for *Society* texts which may be considered as an 'opposite' topic with respect to *Politics*. This indicates that in cross-topic cases, when topics significantly differ, lower dimensionality is advisable, since low frequency features correspond to topic-specific information. The cross-genre case (*Book reviews*) seems not to be influenced so much by such topic-specific features. However, one should keep in mind that book reviews and opinion articles have many similarities and many of the book reviews included in the corpus talk about politics.

Word features produce the same general picture. The main difference is that performance drops much earlier, at about 500-1,500 features corresponding mainly to function words and some very frequent closed-class words (nouns, verbs, etc.) The inclusion of more topic-specific words harms the word-based attribution models especially in cross-topic conditions where the topic significantly differs with that of the training texts (e.g., *Politics* vs. *Society*).

Next, we repeat the above experiment with a varying number of candidate authors. In more detail, we tested candidate set sizes of 2, 3, 5, and 8. For each candidate set size, 30 repetitions were performed by selecting (without replacement) a subset of the 13 authors included in the corpus. Figures 3 and 4 show the classification performance (averaged over the repetitions) for word and character n-gram features, respectively, in a same-genre cross-topic scenario (training texts come from *Politics* topic while evaluation texts come from *World* topic) for varying candidate set sizes (the case where all 13 authors are included is also shown). Naturally, when the candidate set size grows larger classification performance drops (recall this is a closed-set classification task). Again, character n-gram models are more effective in comparison to the respective word-based models. An interesting point is that, in both cases, the most appropriate dimensionality seems to depend on the candidate set size. The larger the candidate set size, the larger the number of features corresponding to the best obtained results. For character n-gram features, the optimal point seems to start at about 2,000 features for 2 candidate authors and grows to about 6,000 features for 13 authors. For word features, the optimal point starts at about 250 features for 2 authors and increases to about 1,500 features for 13 candidate authors.

Figures 5 and 6 show the results of a similar experiment, where the training texts come from opinion articles on *Politics* and evaluation texts come from a different genre (*Book reviews*). Again, we note that the number of character n-gram features corresponding to the best results starts at about 3,000 for 2 authors and gradually increases with the candidate set size reaching about 9,000 features for

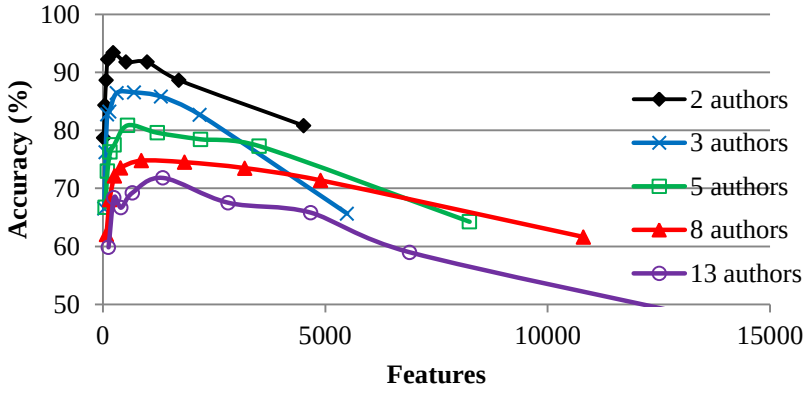

Fig. 3. Performance of the word-based attribution model trained on texts from one topic (Politics) and evaluated on texts from another topic (World) for a varying number of candidate authors

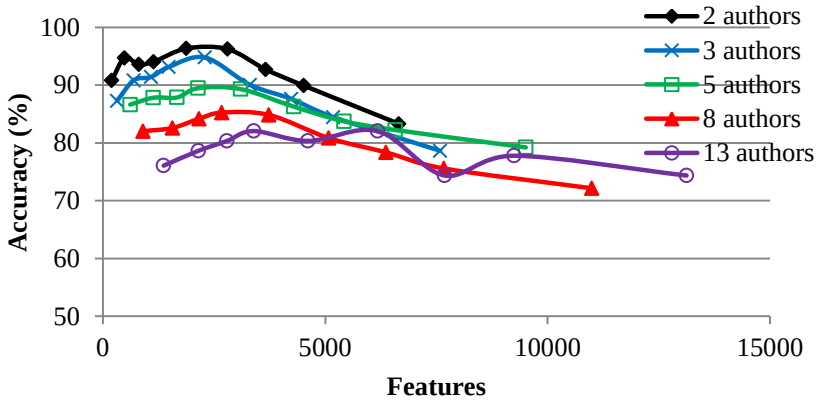

Fig. 4. Performance of the character n-gram attribution model trained on texts from one topic (Politics) and evaluated on texts from another topic (World) for a varying number of candidate authors

13 authors. In comparison to the previous cross-topic experiment, a higher number of features is required. This can be explained by the fact that many book reviews of this corpus talk about politics, therefore low-frequency topic-specific features are useful. On the other hand, it is noted that the number of word-based features that provide the best achieved performance does not vary that much. It starts at about 250 features for 2 candidate authors and reaches about 400 features for 13 authors. Since the most frequent words correspond to function words, these results indicate that function words are reliable features in cross-genre conditions. On the

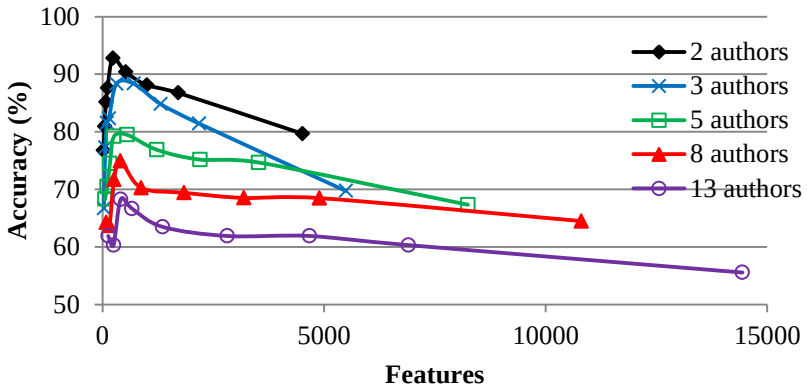

Fig. 5. Performance of the word-based attribution model trained on texts from one genre (opinion articles about Politics) and evaluated on texts from another genre (Book reviews) for a varying number of candidate authors

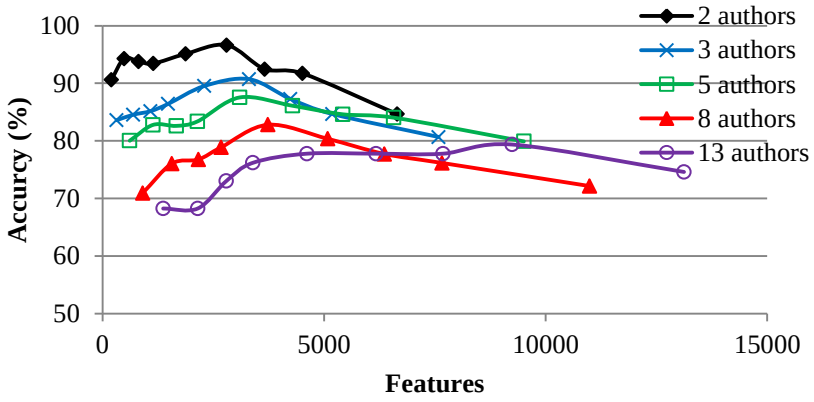

Fig. 6. Performance of the character n-gram attribution model trained on texts from one genre (opinion articles about Politics) and evaluated on texts from another genre (Book reviews) for a varying number of candidate authors

other hand, word features seem unable to exploit the topic similarities (captured by low frequency words) in cross-genre cases, in contrast to character n-gram features.

5 Conclusion

In this paper we applied some well-known stylometric approaches, that is, function words and character n-grams to authorship attribution tasks. Such features are universal since they are easily available for practically any type of text and natural language. Beyond the case where all documents are matched for topic and genre, we examined their effectiveness under more challenging scenarios where the training and evaluation documents talk about different topics or belong to different genres. It is demonstrated that the attribution models, especially the ones based on character n-gram features, can be surprisingly effective in cross-topic or cross-genre conditions.

Character n-gram features performed better than word-based features in all experiments. The attribution models based on character n-grams require considerably higher dimensionality and are able to take advantage of low-frequency features where there are topic similarities among texts. On the other hand, word-based models mainly exploit topic-independent function words while low-frequency words seem to harm their effectiveness when there are changes in topic or genre.

One crucial decision concerns the dimensionality of the representation. It was shown that changes in topic and/or genre as well as the number of candidate authors considerably affect the appropriate choice of the number of features in the attribution models. In the simple scenario where training and evaluation documents are matched for topic and genre, maximum dimensionality is advisable for character n-gram features. When the topic or genre changes, the representation dimensionality should be carefully defined taking into account the number of candidate authors. However, it is not yet clear how this could be done formally.

It is not claimed that the examined models are the best possible for authorship attribution tasks. Their advantage is that they are universal, in the sense that they can be used in any case and provide a robust and accurate model even under difficult scenarios. Any alternative stylometric model, either a new set of features or a combination of several types of features, should be compared with the discussed function word and character n-gram models to prove that it performs better than these baseline approaches.

References

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

Argamon S., Saric M., Stein S. (2003) Style mining of electronic messages for multiple authorship discrimination: First results. In: Proceedings of the 9th ACM SIGKDD, pp. 475-480

Argamon S., Whitelaw C., Chase P., 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

Arun R., Suresh V., Madhavan C.E.V. (2009) Stopword graphs and authorship attribution in text corpora. In: Proc. of the 3rd IEEE International Conference on Semantic Computing, pp. 192-196

Benedetto D., Caglioti E., Loreto V. (2002) Language trees and zipping. Physical Review Letters, 88(4), 048702

Burrows J.F. (1992) Not unless you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing, 7(2):91–109

Chaski C.E. (2005) Who's at the keyboard?: Authorship attribution in digital evidence investigations. International Journal of Digital Evidence, 4(1)

Cristani, M., Roffo G., Segalin C., Bazzani L., Vinciarelli A., Murino V. (2012) Conversationally-inspired stylometric features for authorship attribution in instant messaging. In: Proc. of the 20th ACM International Conference on Multimedia, pp. 1121-1124

de Vel O., Anderson A., Corney M., Mohay G. (2001) Mining e-mail content for author identification forensics. SIGMOD Record, 30(4):55-64

Gamon M. (2004) Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In: Proceedings of the 20th International Conference on Computational Linguistics, pp. 611-617

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

Holmes D.I. (1998) The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing, 13(3):111-117

Jair Escalante H., Solorio T., Montes-y-Gómez M. (2011) Local histograms of character n-grams for authorship attribution. In: Proc. of ACL, pp. 288-298

Joachims T. (1998) Text categorization with support vector machines: Learning with many relevant features. In Proc. of the 10th European Conference on Machine Learning, pp. 137-142

Kanaris I., Stamatatos E. (2009) Learning to recognize webpage genres. Information Processing and Management, 45(5):499-512

Khmelev D.V., Teahan W.J. (2003) A repetition based measure for verification of text collections and for text categorization. In: Proceedings of the 26th ACM SIGIR, pp. 104–110

Keselj V., Peng F., Cercone N., Thomas C. (2003) N-gram-based author profiles for authorship attribution. In: Proceedings of the Pacific Association for Computational Linguistics, pp. 255-264

Koppel M., Schler J., Argamon S. (2011) Authorship attribution in the wild. Language Resources and Evaluation, 45:83–94

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

Koppel M., Winter Y. (2014) Determining if two documents are by the same author. Journal of the American Society for Information Science and Technology, 65(1):178-187

Lim C.S., Lee K.J., Kim G.C. (2005) Multiple sets of features for automatic genre classification of web documents. Information Processing and Management, 41(5):1263-1276

Luyckx K., Daelemans W. (2008) Authorship attribution and verification with many authors and limited data. In: Proceedings of the Twenty-Second International Conference on Computational Linguistics, pp. 513-520

Luyckx K., Daelemans W. (2005) Shallow text analysis and machine learning for authorship attribution. In: Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands

Madigan D., Genkin A., Lewis D., Argamon S., Fradkin D., Ye L. (2005) Author identification on the large scale. In: Proceedings of CSNA-05

Mendenhall T. C. (1887) The characteristic curves of composition. Science, IX:237–49

Meyer zu Eissen S., Stein B. (2004) Genre classification of web pages: User study and feasibility analysis. In: Biundo S., Fruhwirth T., Palm G. (eds.) KI 2004: Advances in Artificial Intelligence, Springer, pp. 256-269

Mosteller, F., Wallace D.L. (1964) Inference and disputed authorship: The Federalist. Addison-Wesley.

Pang B., Lee L. (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2):1–135

Rangel F., Rosso P., Koppel M., Stamatatos E., Inches, G. (2013) Overview of the author profiling task at PAN 2013. In: Forner P., Navigli R., Tufis D. (eds.) Working Notes Papers of the CLEF 2013 Evaluation Labs

Santini M. (2007) Automatic identification of genre in webpages. Ph.D. Thesis, University of Brighton.

Seidman S. (2013) Authorship verification using the impostors method. In: Forner P., Navigli R., Tufis D. (eds). CLEF 2013 Evaluation Labs and Workshop –Working Notes Papers

Sebastiani F. (2002) Machine learning in automated text categorization. ACM Computing Surveys, 34(1)

Sidorov G., Velasquez F., Stamatatos E., Gelbukh A.F., Chanona-Hernández L. (2014) Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications, 41(3):853-860

Stamatatos E. (2011) Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology, 62(12): 2512-2527

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

Stamatatos E. (2009b) Intrinsic plagiarism detection using character n-gram profiles. In: Proc. of the 3rd Int. Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse

Stamatatos E., Daelemans W., Verhoeven B., Stein B., Potthast M., Juola P., Sánchez-Pérez M.A., Barrón-Cedeño A. (2014) Overview of the author identification task at PAN 2014. CLEF Working Notes, 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

Van Halteren H. (2007) Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing, 4(1):1-17

Weiss S.M., Indurkhya N., Zhang T., Damerau F. (2005) Text mining: Predictive methods for analyzing unstructured information, Springer

Yule G.U. (1944) The statistical study of literary vocabulary. Cambridge University Press

Zheng R., Li J., Chen H., Huang Z. (2006) A framework for authorship identification of online messages: Writing style features and classification techniques. Journal of the American Society of Information Science and Technology, 57(3):378-393