The Use of Orthogonal Similarity Relations in the Prediction of Authorship

Upendra Sapkota[1], Thamar Solorio[1], Manuel Montes-y-Gómez[2], and Paolo Rosso[3]

[1] University of Alabama at Birmingham, Birmingham, AL 35294, USA,
{upendra, solorio}@cis.uab.edu,
[2] Instituto Nacional de Astrofísica, Optica y Electrónica , Puebla, Mexico,
mmontesg@ccc.inaoep.mx,
[3] NLE Lab - ELiRF, Universitat Politècnica de València, Valencia, Spain,
prosso@dsic.upv.es

Abstract. Recent work on Authorship Attribution (AA) proposes the use of meta characteristics to train author models. The meta characteristics are orthogonal sets of similarity relations between the features from the different candidate authors. In that approach, the features are grouped and processed separately according to the type of information they encode, the so called linguistic modalities. For instance, the syntactic, stylistic and semantic features are each considered different modalities as they represent different aspects of the texts. The assumption is that the independent extraction of meta characteristics results in more informative feature vectors, that in turn result in higher accuracies. In this paper we set out to the task of studying the empirical value of this modality specific process. We experimented with different ways of generating the meta characteristics on different data sets with different numbers of authors and genres. Our results show that by extracting the meta characteristics from splitting features by their linguistic dimension we achieve consistent improvement of prediction accuracy.

1 Introduction and Background

Authorship Attribution (AA) is the task of identifying the author of a given anonymous text, or a text whose authorship is in doubt. Although the authorship attribution task is often solved as a multi-class, single-label text categorization task, the purpose of AA is to model each author's writing style rather than modeling thematic content of the available documents, as in the case of the typical text classification task. There are many relevant applications of AA in Forensic Linguistics. For instance, AA can help fight spam filtering [26], cyber bullying, and other forms of cyber crime (e.g., identifying authors of malicious code, or potential pedophiles). Other applications include plagiarism detection [22], author recognition of a given program [7], and web information management.

As described in the Stamatatos survey [23], there are two main frameworks that have been successfully used in the relevant literature: the profile-based approach, and the standard machine learning one. Both of them assume the availability of some number of documents with known authorship that can be used to build the models. In a profile based approach, all documents from the same author in the training set are concatenated. Then profiles are created for each author by extracting several features from

these merged files. These approaches rely mostly on low level features, such as character n-grams. To predict authorship of a new document, a similarity score between the new document and each author profile needs to be computed. The document will then be assigned to the author whose profile yields the highest similarity score. Because the similarity between the test document and the profiles is computed independently for each author, this approach allows to use profile-specific features. Typically the features are selected based on their frequency of appearance in the profile. Examples of a profile based approach include [10, 20, 11].

In contrast, machine learning approaches to AA use a feature vector representation where each single document from the training set is represented individually by the same set of features. The feature vectors are then used to train a machine learning algorithm. These feature vectors are usually a varied combination of lexical, character, and syntactic features such as average word length, average sentence length, content words, function words, word n-grams, character n-grams, and parts-of-speech (POS) n-grams. Recent approaches have reported good prediction performance for this task using Support Vector Machines [6], memory based learners [13, 14], and Probabilistic Context Free Grammars [17].

In a recent work, Solorio *et al.* [19] proposed an AA approach that explicitly exploits the differences in the nature of the features representing the documents to generate informative meta features. The key assumption in their work is that by breaking down the document representation into a set of orthogonal dimensions[4], meaningful similarity patterns among authors could emerge. Then these similarity patterns can be exploited by the machine learning algorithms to boost authorship prediction accuracy. This approach is loosely related to well known machine learning approaches, such as the co-training algorithm by Blum and Mitchell [3] where two classifiers are trained on different views of the data. However, the goal of having different views of the data in Solorio *et al.*'s work is to extract disjoint similarity relations among the instances from different classes and not to train classifiers on disjoint subsets of features.

In this paper we set out to investigate the value of extracting the meta features following the framework proposed by [19]. The main contribution of this paper is the empirical evidence gathered that shows we can model the writeprint of authors by combining standard lexical and stylistic features with modality specific similarity relations among the writing preferences of different authors. Although the idea of these meta features was proposed by previous work, the empirical evaluation was done on a single corpus and with a single train/test partition of the data. Moreover, the authors in that paper left an important question unanswered: *Is the notion of linguistic modalities really needed?* In other words, similarity relations from disjoint sets of features seem to help boost prediction accuracy. Do we need to partition the feature set by their linguistic nature, or is it sufficient to just partition this set randomly? Because the implications of these questions are relevant to explore a more general application of this approach, we consider necessary to search for answers and report our findings. This study presents the first statistically significant results supporting the need for linguistic modalities in several datasets using a cross validation setting. We also report on results of experi-

[4] We use the term *orthogonal* loosely in this paper to refer to sets of features that are coming from different linguistic dimensions and that are disjoint from one another.

ments that allow for a direct comparison with state-of-the-art AA approaches. New in this paper is also a study of the individual modalities that are being used. In sum, we aim to provide a better understanding of the value of adding the meta characteristics to the representation of documents in the AA task.

2 Document Representation

Following the formulation in [19], we exploit the notion of linguistic modalities, where each linguistic modality refers to a set of features representing different aspects of the text. For instance, features related to syntax are considered a different modality from features related to semantics. Therefore, rather than representing each instance directly by a feature vector $\boldsymbol{x}$, we represent it by a set of M smaller feature vectors $\{\boldsymbol{x_1}, \boldsymbol{x_2}..., \boldsymbol{x_m}\}$, where $m = |M|$, the number of modalities, and $\boldsymbol{x_i}$ is the feature vector in modality i. The combination (union) of these smaller vectors (sub-vectors) forms the single feature vector representation of the instance in the standard scenario. We call this set of vectors first level features (FLF) following the same convention as in [19]. After the extraction of FLF we proceed with the generation of modality specific meta features (MSMF) as follows:

1. The first step in this meta feature extraction is the unsupervised clustering of all the feature vectors in the training set belonging to the same modality. We do this for each modality in the training set, which results in k clusters per modality, i.e., $m \times k$ total clusters. Because the training instances are clustered by modality, each modality will have its own clustering solution.
2. For each ith clustering solution, we compute the centroid of each cluster $c_{i,j}$ by averaging the feature vectors belonging to that cluster.
3. For each document, from the training and testing sets, we compute its similarity to the centroids of each cluster using the cosine similarity function. These similarity scores are the meta features.

As meta features are calculated on a modality basis, each modality gives us as many meta features as the number of clusters in that modality. Each sub-vector of FLF, $\boldsymbol{x_i}$, has a corresponding meta feature vector $\boldsymbol{x'_i}$ with the length equal to the number of clusters k in the given modality. We use same k for each modality, therefore, $m \times k$ meta features are extracted from m modalities. All FLF and meta features are joined (concatenated) into a single feature vector that is used to train a machine learning algorithm. In Figure 1 we show a graphical representation of the computation of MSMF. In that figure, $\{c_{sty,1}..c_{sty,k}\}$ are k clusters formed from the stylistic feature vectors, and likewise, $\{c_{syn,1}..c_{syn,k}\}$ are k clusters formed from the syntactic feature vectors. The stylistic meta feature vectors $\boldsymbol{x'_{sty}}$ = $\{x'_{sty,1}..x'_{sty,k}\}$ and syntactic meta feature vectors $\boldsymbol{x'_{syn}}$ = $\{x'_{syn,1}..x'_{syn,k}\}$ are formed after computing the cosine similarity of the document instance to the cluster centroid on their own modality. $\boldsymbol{x_{p(sty)}}$ and $\boldsymbol{x_{p(syn)}}$ are vector representations of each instance in the *Stylistic* and *Syntactic* modality, respectively.

In this work we consider four different types of feature groups –stylistic, lexical, perplexity values from character level n-gram language models, and syntactic features, for a total of four modalities ($m = 4$). It is worth noting that the notion of linguistic

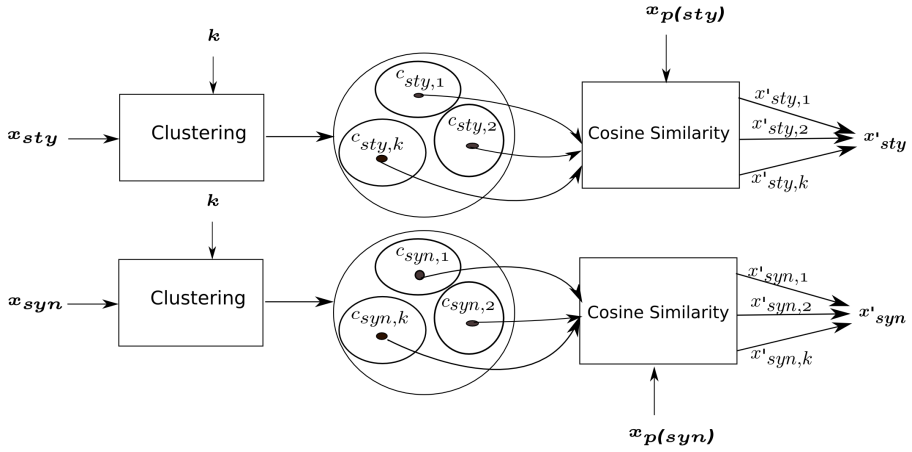

Fig. 1. Diagram showing the computation of modality specific meta features from two modalities: *Stylistic (sty)* and *Syntactic (syn)*.

modalities as used in this and previous work has a connection with the notion of linguistic dimensions defined by Biber's work on genre analysis [2]. The contrast is at the level of abstraction. Biber's dimensions define a set of features common at a discourse level, while in this work linguistic modalities refer to different lower levels of analysis.

Note that since no class information is used during the clustering process, the MSMF approach is clearly different from other well studied methods for reducing data dimensionality [1, 18, 4]. The goal of the clustering step, in our AA framework, is to generate new meaningful features. The clustering allows us to generate similarity patterns from the posts of different authors on individual modalities. Some authors use the emoticons in a similar way, while some share the use of punctuations marks. We believe that the encoding of these similarities in the meta features complements the information provided by the first level features to the machine learning algorithm.

2.1 Features

The FLF features used in our work are refinement (addition and modification) of [19]. The final list of features is shown in Table 1. The first column shows how these features are categorized by the type of information they are extracting from the document. The first modality (*Stylistic*) tries to capture writing choices that reflect authors preferences and thus contains features related to the use of punctuation marks, length of sentences, and use of contractions, among others. The *Syntactic* modality focuses on the grammatical patterns of the authors. It includes n-grams from POS tags and bag of syntactic relations. In the *Semantic* modality the goal is to capture the topic/author correlation, as well as the information related to word choices for each author. This modality then uses the standard bag of words representation used in text classification tasks where

stop words are removed from the documents before generating the feature vectors. The last modality (*Perplexity*) contains perplexity values from language models. We train one language model per author. We expect that perplexity scores will be lower for the documents belonging to the corresponding author's model, similar to the intuition in [17] of using probabilistic context free grammars. We trained character-level 4-gram language models for each author in the training set. Then we compute the perplexity values for each document in a leave-one-out setting.

There are some differences between the final feature set used in our work and that used by Solorio *et al.* (2011). We added new features in the *Stylistic* modality: total number of sentences, percentage of words without a vowel, number of balanced parenthesis, and number of tokens containing at least one capital letter. We also modified the feature for the use of quotations in the same *Stylistic* modality. Instead of having a binary feature we use here the total number of quotations. Because several datasets are coming from social media, we thought vowel-less words would be a common feature and might improve the performance. The features such as number of sentences have been successfully used in previous research. The goal is to distinguish authors that produce long and wordy documents from those that tend to be more succinct. For the *Perplexity* modality, we used higher order language models, 4-grams, instead of 3-grams. Character 4-grams have been successfully used for AA tasks and in our case, we believe that 4-grams allow us to better capture not only patterns from the endings of the words but also the lemmas of the words as well as patterns about the use of functional words. Features that are different in this paper are highlighted in Table 1.

Table 1. First level features used in the representation of documents for the AA task. The '$^+$' after a feature indicates new features not present in previous work. The '*' indicates a modified feature.

Modality	**First Level Features (FLF)**
Stylistic	Total number of sentences$^+$
	Average number of tokens per sentence
	Percentage of words without vowel$^+$
	Average number of punctuations per sentence
	Percentage of contractions
	Total number of balanced parenthesis$^+$
	Percentage of two consecutive punctuation marks
	Percentage of three consecutive punctuation marks
	Total number of alphabetic characters
	Average number of tokens with at least a capitalized letter per sentence$^+$
	Toal number of sentence initial words with first letter capitalized
	Total number of quotations*
Syntactic	Top 1000 POS tag unigrams
	Top 1000 POS tag bigrams
	Top 1000 POS tag trigrams
	Top 1000 Grammatical relations from the dependency parses
Semantic	Top 1000 bag-of-words
Perplexity	All the perplexity values from character 4-grams*

3 Data Sets

We tried to consider a collection of test sets with varied challenging characteristics to provide a more comprehensive data that will also allow us to benchmark our results. We selected three collections used by state-of-the-art approaches to AA. Table 2 shows several statistics of the different data sets. The first collection is from Solorio *et al.* (2011) [19]. This collection consists of five datasets with a different number of authors taken from forums of The Chronicle of Higher Education (CHE). As shown in Table 2, this data set contains very short documents ($\sim$6 sentences per post), which imposes an interesting challenge for the AA task. Another important characteristic of this data set is the imbalanced distribution of documents per author. In the data sets with 20, 50 and 100, there are some authors that are heavily represented and some for which only a few documents are available. This setting is closer to what one would expect to see in real world scenarios, since we cannot control how much each user interacts on the forum. However the nice characteristic about this collection is that all posts are coming from the same topic. We expect this will reduce the chances of having a strong topic/author correlation that will be reflected in the value of the *Semantic* modality.

Another collection is from Raghavan *et al.* (2010) [17]. They collected five datasets from material downloaded from the Internet. Four of them contain news articles on topics related to Business, Travel, Football, and Cricket. The fifth data set contains poems from the Project Gutenberg website[5]. We chose Raghavan *et al.*'s collection because it contains data from different topics and different genres, and because we can do a direct comparison with their results. The datasets in this collection have a varied number of authors ranging from 3 to 6.

The last collection is the CCAT topic class, a subset of the Reuters Corpus Volume 1 [12]. This collection was not gathered for the goal of doing authorship analysis studies. But the common use of this data set in previous studies provides a unique opportunity to benchmark our results. Previous work has reported results for AA with 10 and 50 authors [21, 16, 6] and we follow this lead to experiment as well with 10 and 50 authors.

We do not expect to have a single best approach that outperforms all other results in such a diverse collection of benchmark data. The goal is to study whether the benefit of using linguistic modality framework generalizes to different datasets, and to try to tease apart how different modalities have varied performance accross collections.

4 Experiments

In this paper we report results that are the overall average accuracy from 5-fold cross-validation, along with the statistical significance of our results. But to provide a one to one reference for comparison, we also performed experiments with the fixed train/test partitions used in state-of-the-art systems whenever this information was available.

We used support vector machines (SVMs) implemented in Weka [27] with default parameters as the underlying classifier. For the *Syntactic* modality, the POS tags were generated by the Stanford tagger [25]. We used the Stanford parser to generate the

[5] `http://www.gutenberg.org/wiki/Main_Page`

Table 2. Some statistics, including distribution of the documents across authors, from the collections we used in our AA experiments. Figures shown in Columns 3 and 4 are averages over the entire collection. **min** shows the minimum number of documents for any single author and **max** shows the maximum number of documents belonging to a single author.

dataset	#auth	#words/doc	# sent/doc	min #docs	max #docs	#docs
CHE	5	75.88	6.26	434	693	2,889
CHE	10	78.24	6.82	321	914	5,579
CHE	20	84.60	7.20	173	1,369	9,779
CHE	50	79.27	6.86	33	2,369	15,543
CHE	100	79.89	6.89	6	2,369	16,171
Football	3	877.00	44.00	31	34	97
Business	6	827.00	40.00	25	30	175
Travel	4	908.00	40.00	37	45	172
Cricket	4	978.00	50.00	30	48	158
Poetry	6	271.00	13.00	19	56	200
CCAT	10	507.24	21.07	100	100	1,000
CCAT	50	505.65	21.54	100	100	5,000

dependency parsers [15]. The SRILM toolkit [24] was used to train the character 4-gram language models. The clustering of the FLF on modality basis was done using CLUTO [9]'s $vculster$ clustering program with parameter $clmethod = rbr$.

We performed a set of experiments designed to answer the question posed in Section 1: *Is the notion of linguistic modalities really needed?* To answer this question we run experiments where instead of fragmenting the FLF by linguistic modality, we randomly generate m subsets of FLF, simulating "random modality meta features" (RMMF). Then we compare results of this approach against generating meta features by linguistic modality, as described in Section 2. For the sake of completeness we also show figures for using only the FLF to train the machine learning algorithm.

The results for all the datasets are presented in Table 3. In 10 out of the 12 datasets the approach using the modality specific meta features in combination with first level features (MSMF+FLF) yields the best results. This was expected since previous work showed this combination to be the best setting in the CHE dataset. The results on additional datasets show the consistency of the approach.

The table also highlights the results that are statistically significant with a 95% confidence level using a two-tailed t-test. All the results, with the exception of the Cricket dataset, show the differences between randomly splitting the feature set and using the notion of linguistic modality to be statistically significant. From these results we can conclude that the boost in accuracy results from the discriminative power of the orthogonal similarities extracted, and not because of a simple decomposition of the feature vector into disjoint subvectors.

There is also a notable difference between the margin of increase in accuracy among the different collections. It seems the CHE collection benefits the most out of the MSMF+FLF setting. There are two possible reasons for this, document length and number of authors. On average, the CHE posts have a length of 80 words and ≈ 7 sentences, while all other datasets have as much as 8 times more tokens and sentences (see Table 2)

Table 3. Accuracies on 12 datasets from 3 different collections used in previous work for AA. Each column shows accuracies for different feature sets: FLF, Randomized Modality Meta Features combined with FLF (RMMF+FLF), and MSMF combined with FLF (MSMF+FLF). Statistical significance between RMMF+FLF and MSMF+FLF, using a two-tailed t-test, is marked with *. Similarly, differences between FLF and MSMF+FLF that are statistically significant are noted with $^\flat$. Gain is given by 100(Col5 - Col3)/Col3.

dataset	#auth	FLF	RMMF+FLF	MSMF+FLF	Gain(%)
CHE	5	72.24	71.86	**79.00**$^{*\flat}$	+9.36
CHE	10	71.04	70.56	**76.07**$^{*\flat}$	+7.07
CHE	20	65.94	64.35	**71.79**$^{*\flat}$	+8.88
CHE	50	57.92	56.31	**65.09**$^{*\flat}$	+12.39
CHE	100	55.53	52.10	**63.50**$^{*\flat}$	+14.35
Football	3	89.30	88.35	**92.75***	+3.41
Business	6	**86.66**	83.05	86.29*	-0.66
Travel	4	83.84	81.46	**86.70***	+3.41
Cricket	4	**96.20**	93.23	95.59	-0.63
Poetry	6	64.27	63.05	**78.29**$^{*\flat}$	+21.81
CCAT	10	83.50	80.5	**84.20***	+0.83
CCAT	50	74.42	69.06	**76.12**$^{*\flat}$	+2.28

per document. It is possible that for documents like the ones on the CHE collection, the FLF do not carry enough information for accurate classification due to the short length of these posts, and therefore there is much more to gain from the meta features. This is also supported by the larger increase in accuracy with the MSMF+FLF approach for the Poetry dataset, which is the one with shorter documents from Raghavan *et al.*'s collection.

But another possible reason why the MSMF+FLF yields higher gains in accuracy could be the number of authors. In a small pool of authors the potential relationships that can emerge from comparing writing styles is limited. In a sufficiently large pool of authors it is clear that there are many more possible combinations, and it is more likely that some of the authors will share writing styles in specific dimensions with different authors. The meta features in such a setting will then carry new and more discriminative value than in the setting with a small number of authors.

To allow a comparison with state-of-the-art approaches we run additional experiments using the same train/test partitions as those reported on recent work. It should be noted that the figures we report for each data set and existing approaches are the highest accuracies we found on those papers. These results are shown in Table 4. It is clear that the MSMF+FLF approach is competitive across the different collections, reaching very similar results to those reported earlier, and in some cases outperforming previous approaches.

5 Analysis of Results

The previous section presented interesting results on different AA tasks where the notion of linguistic modalities and a framework designed to exploit similarity scores in

Table 4. Benchmark comparison with recent AA approaches using the same collections and on the same train/test partitions. The numbers in parenthesis show our results from the 5-fold cross-validation setting. For each dataset, bold figure represents the best performance.

dataset	#auth	MSMF+FLF (5fcv)	Benchmark Comparison
CHE	5	74.30 (79.00)	**75.47** [19]
CHE	10	**77.96** (76.07)	77.38 [19]
CHE	20	**72.48** (71.79)	71.42 [19]
CHE	50	**67.00** (65.09)	63.79 [19]
CHE	100	**63.61** (63.50)	62.10 [19]
Football	3	91.11 (92.75)	**93.34** CNG-WPI [5] 91.11 PCFG-E [17]
Business	6	86.66 (86.29)	**91.11** PCFG-E [17] 80.00 CNG-WPI [5]
Travel	4	90.00 (86.70)	**91.67** PCFG-E [17] 73.33 CNG-WPI [5]
Cricket	4	91.11 (95.59)	**95.00** PCFG-E [17] 90.00 CNG-WPI [5]
Poetry	6	63.63 (78.29)	**87.27** PCFG-E [17] 85.45 CNG-WPI [5]
CCAT	10	78.80 (84.20)	**86.4** BOLH Diffusion Kernel [6] 79.40 Char n-grams SVM [21] 78.00 STM-Asymmetric cross [16] 73.60 CNG-WPI [5]
CCAT	50	69.48 (76.12)	**74.04** Char n-grams SVM [8]

a modality specific way yields higher prediction rates than simply using the first level features. We run and analysed an additional set of experiments to explore how much these linguistic modalities are contributing to the models.

Table 5 shows the results on training a SVM using a single modality at a time. In this set of experiments we used the single train/test partition as that used in previous work, and in the results reported in Table 4. These results clearly show that the different characteristics of the data sets have a notable effect on the usefulness of the different linguistic modalities. The *Stylistic* modality has a considerable contribution to the final classification for all the CHE datasets, and is the one with one of the lowest accuracies for all other data sets. This was somewhat expected as several of the stylistic features in that set were crafted in [19] with web forum data as the focus. If we go back to the description of the features (see Table 1), we can easily identify some of the stylistic features that are most likely to not carry any discriminative value for the other data sets, such as the ones related to punctuation marks and capitalization information, as these other data sets have a very uniform pattern for them.

For the *Semantic* modality we have a different result. In the CHE collection, this modality was the second best one in accuracy, but in the CCAT collection this modality reached the highest accuracy. We believe this is a good indication that in this data set there is a stronger correlation between the topic of the document and the authors. Similarly, there seems to be a strong author/topic effect in the Raghavan et al.'s collection,

Table 5. Comparison on accuracies obtained by individual modalities on various datasets. For each dataset, the bold figure indicates the accuracy of the best modality obtained by the best feature set (one of the three feature sets: FLF, MSMF, and MSMF+FLF).

Dataset	#Author	Feature Set	Modality			
			Semantic	***Perplexity***	***Syntactic***	***Stylistic***
CHE	5	MSMF	38.02	23.95	34.89	54.86
		FLF	45.86	16.36	36.42	59.44
		MSMF+FLF	46.40	41.09	36.87	**65.11**
CHE	10	MSMF	30.75	40.73	23.02	60.70
		FLF	45.86	16.36	36.42	60.70
		MSMF+FLF	46.40	40.64	36.87	**65.10**
CHE	20	MSMF	31.46	14.73	22.17	56.05
		FLF	39.01	14.06	30.13	52.92
		MSMF+FLF	39.83	14.88	29.67	**60.42**
CHE	50	MSMF	30.50	15.57	19.68	51.45
		FLF	35.36	15.34	25.38	45.88
		MSMF+FLF	37.07	15.54	25.35	**54.04**
CHE	100	MSMF	31.21	14.84	20.72	50.93
		FLF	32.46	14.84	23.70	45.27
		MSMF+FLF	32.87	15.02	24.20	**52.09**
Football	3	MSMF	80.00	75.55	60.00	44.44
		FLF	86.66	**91.11**	82.22	64.44
		MSMF+FLF	86.66	77.77	82.22	73.33
Business	6	MSMF	73.33	77.77	40.00	32.22
		FLF	80.00	63.33	73.33	57.77
		MSMF+FLF	80.00	**83.33**	73.33	53.33
Travel	4	MSMF	80.00	**86.66**	36.66	35.00
		FLF	76.66	76.66	81.66	43.33
		MSMF+FLF	76.66	85.00	81.66	46.66
Cricket	4	MSMF	73.33	91.66	58.33	63.33
		FLF	80.00	61.66	90.00	66.66
		MSMF+FLF	80.00	**91.66**	91.66	80.00
Poetry	6	MSMF	40.00	**80.00**	27.27	18.18
		FLF	34.54	52.72	40.00	18.18
		MSMF+FLF	34.54	78.18	43.63	20.00
CCAT	10	MSMF	68.40	68.60	28.80	24.60
		FLF	74.80	51.00	74.00	33.20
		MSMF+FLF	**76.00**	69.60	73.60	31.80
CCAT	50	MSMF	57.92	56.92	15.96	13.28
		FLF	62.76	34.00	55.20	10.96
		MSMF+FLF	**66.08**	57.56	55.36	14.60

as the results from this modality are among the highest ones. This could also explain why their PCFG-E approach gave the best results in those data sets as the use of lexical features that carry the semantic content could help boost accuracy of their system.

The results on the *Syntactic* modality seem to indicate a correlation with document length. This modality yields some of the lowest results for those data sets with shorter

documents. Overall, these features seem to have a limited contribution to identify authors for the CHE collection, the one with the shortest documents. In the Raghavan et al.'s collection, the data sets with longer documents have higher accuracies when the SVM is trained on only these features. The datasets Football, Business, Travel and Cricket yield accuracies higher than 70% when using the *Syntactic* modality. But for the Poetry dataset this same modality reached an accuracy of 40% in the best case. This latter dataset has an average of 250 words per document, while the former datasets have between 800 and a little over 950 words (see Table 2). Another plausible explanation for these results can be the genre of the datasets. In the CHE datasets there could be a lot of noise in the parser output because of the spontaneity and casual writing style. Although since it is a forum tied to academe, the level of noise from typos, emoticons, abbreviations and slang is not as high as in a typical web forum. In the Poetry dataset it is possible that the format from the prose in there can cause the syntactic analysers to break and output noisy tags and parses.

The same document length effect can be observed in the *Perplexity* modality. Overall, higher accuracies can be seen for datasets with longer documents (CCAT collection and Raghavan et al.'s Football, Business, Cricket, and Travel data sets). Since we are using character 4-grams, there will be some semantic content included here. It is likely too that this is also playing a role in reaching better results for collections that were not deliberately controlled for topic.

In summary, the differences in accuracy reached by the individual modalities indicate that the genre of the documents should guide the selection of features for building the models. For the MSMF approach studied here this conclusion motivates the need for a more sophisticated way to combine the features from the individual modalities. It is possible that higher overall results could be attained if we allow the more discriminative modalities to have a higher weight than other less meaningful modalities in the final author models. A framework like this must be adaptable to the peculiarities of the target datasets and this could be reached with the help of a validation set where such parameters could be fine tuned.

6 Conclusions and Future Work

In this paper we set out to the task of investigating the empirical value of extracting orthogonal similarity patterns in authors writing style to improve AA accuracy. Most approaches rely on finding distinguishable markers in each author's writing style to perform the task, whereas this approach explicitly exploits the notion that authors share writing patterns across specific linguistic dimensions. This idea has been explored by previous work, but without a comprehensive evaluation across different datasets, a one to one comparison with state-of-the-art approaches, and without a necessary comparison with a random generation of linguistic dimensions. The sets of experiments reported here resolve those remaining questions. Our findings show this is a competitive AA framework that seems to be especially useful for datasets with larger number of authors and shorter documents.

The findings from this work also underscore the need for a better modelling of the genre for the AA task. Our results show that significant differences can be attained

by the linguistic modalities depending on the nature of the documents. Therefore, a promising line for future work concerns the investigation of an adaptable model where the meta features from different linguistic dimensions will have different weights on the final decisions to reflect their discriminative value.

One of the trends identified in our experimental results refers to observing higher gains in accuracy when adding the modality specific meta features when the number of candidate authors increases. This trend was consistent for all but one of the CHE datasets and both of the datasets from the CCAT collection. It indicates the possibility that the differences in the writeprint of authors reach the ceiling of their discriminative power as the number of candidate authors increases. At the same time, the richer set of authors allows to extract more powerful similarity coefficients when following the modality specific framework. Further experiments are needed to support this claim and we plan to focus on this in the coming months.

It is possible that the notion of orthogonal similarity patterns could be useful in other classification tasks beyond authorship analysis. One potential task is genre classification. Clearly there must be several similarities between different genres across different dimensions. Some genres share stylistic features, consider data from social media, while some others share a different set of modalities, *Semantic* or *Syntactic*. It is then possible that a MSMF approach could yield competitive results.

Acknowledgements

This research was partially supported by ONR grant N00014-12-1-0217 and by NSF award 1254108. It was also supported in part by the CONACYT grant 134186 and by the European Commission as part of the WIQ-EI project (project no. 269180) within the FP7 People Programme.

References

1. Baker, L.D., McCallum, A.: Distributional clustering of words for text classification. In: SIGIR 98: Proceedings of the 21st Annual International ACM SIGIR. pp. 96–103. ACM, Melbourne, Australia (August 1998)
2. Biber, D.: The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. Computers and the Humanities 26, 331–345 (1993)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the 1998 Conference on Computational Learning Theory (1998)
4. Dhillon, I.S., Mallela, S., Kumar, R.: A divisive information-theoretic feature clsutering algorithm for text classification. Journal of Machine Learning Research 3, 1265–1287 (2003)
5. Escalante, H.J., Montes-y Gomez, M., Solorio, T.: Weighted profile intersection measure for profile-based authorship attribution. In: Batyrshin, I., Sidorov, G. (eds.) Proceedings of the 10th Mexican International Conference on Artificial Intelligence, MICAI 2011. LNCS, vol. 7094, pp. 232–243. Puebla, Mexico (2011)
6. Escalante, H.J., Solorio, T., Montes-y Gomez, M.: 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. Association for Computational Linguistics, Portland, Oregon, USA (June 2011)

7. Hayes, J.H.: Authorship attribution: A principal component and linear discriminant analysis of the consistent programmer hypothesis. I. J. Comput. Appl. pp. 79–99 (2008)
8. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: Euzenat, J., Domingue, J. (eds.) AIMSA 2006. LNAI, vol. 4183, pp. 77–86 (2006)
9. Karypis, G.: CLUTO - a clustering toolkit. Tech. Rep. #02-017 (Nov 2003)
10. Keselj, V., Peng, F., Cercone, N., Thomas, C.: N-gram based author profiles for authorship attribution. In: Proceedings of the Pacific Association for Computational Linguistics. pp. 255–264 (2003)
11. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language Resources and Evaluation 45, 83–94 (2011)
12. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research 5, 361–397 (2004)
13. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). pp. 513–520. Manchester, UK (August 2008)
14. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Literary and Linguistic Computing pp. 1–21 (August 2010)
15. Marneffe, M.D., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: LREC 2006 (2006)
16. Plakias, S., Stamatatos, E.: Tensor space models for authorship attribution. In: Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications. LNCS, vol. 5138, pp. 239–249. Syros, Greece (2008)
17. Raghavan, S., Kovashka, A., Mooney, R.: Authorship attribution using probabilistic context-free grammars. In: Proceedings of the ACL 2010 Conference Short Papers. pp. 38–42. Association for Computational Linguistics, Uppsala, Sweden (July 2010)
18. Slonim, N., Tishby, N.: The power of word clusters for text classification. In: 23rd European Colloquium on Information Retrieval Research (ECIR) (2001)
19. Solorio, T., Pillay, S., Raghavan, S., Montes-y Gómez, M.: Generating metafeatures for authorship attribution on web forum posts. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, IJCNLP 2011. pp. 156–164. AFNLP, Chiang Mai, Thailand (November 2011)
20. Stamatatos, E.: Author identification using imbalanced and limited training texts. In: 18th International Workshop on Database and Expert Systems Applications, DEXA '07. pp. 237 –241 (Sept 2007)
21. Stamatatos, E.: Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Managemement 44, 790–799 (2008)
22. Stamatatos, E.: Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology 62(12), 2512–2527 (2011)
23. Stamatatos, E.: A survey on modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
24. Stolcke, A.: SRILM - an extensible language modeling toolkit. pp. 901–904 (2002)
25. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. pp. 173–180. NAACL '03 (2003)
26. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Multi-topic e-mail authorship attribution forensics. In: Proceedings of the Workshop on Data Mining for Security Applications, 8th ACM Conference on Computer Security (2001)
27. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kauffmann, 2nd edn. (2005)