Modality Specific Meta Features for Authorship Attribution in Web Forum Posts

Thamar Solorio and **Sangita Pillay**
The University of Alabama at Birmingham
1300 University Boulevard
Birmingham, AL 35294, USA
{solorio,rsangita}@cis.uab.edu

Sindhu Raghavan
The University of Texas at Austin
1 University Station C0500
Austin, TX 78712, USA
sindhu@cis.uab.edu

Manuel Montes y Gómez
The University of Alabama at Birmingham
1300 University Boulevard
Birmingham, AL 35294, USA
National Institute of Astrophysics, Optics, and Electronics
Luis Enrique Erro No. 1
Tonantzintla, Puebla, Mexico
mmontesg@ccc.inaoep.mx

Abstract

This paper presents a new method for Authorship Attribution (AA) on online forum posts. The idea behind the method is to generate meta features that capture modality specific similarity relations among texts from different authors. Each modality represents a particular linguistic dimension (syntactic, lexical, stylistic). To evaluate this approach we measure prediction accuracy on data from an online forum with up to 100 candidate authors. We also compare our results with a state of the art approach that has shown to perform well across different genres. We have found the meta features to be especially helpful in the online forum domain, where the documents are very short, showing this to be a very promising direction for AA on a realistic web forum scenario.

1 Introduction

Authorship attribution (AA) refers to the task of analyzing a document to identify the potential author who wrote the text. Earlier work on this problem involved gathering statistics about the frequency of words with specific length, together with other stylistic characteristics extracted from written samples that were in most cases an entire book or volume (Mendenhall, 1887; Mosteller and Wallace, 1964). Current approaches to AA relay on casting this problem as a text classification task, where instead of aiming to do a thematic classification of documents the goal is to have the models learn the distinguishable characteristics in the written work of authors. The focus of analysis on more recent work has also shifted from book-length pieces to documents with length ranging from a couple of blocks (Hirst and Feiguina, 2007) to samples with at most 140 characters (Layton et al., 2010).

AA can help settle disputes over the original creators of a given piece of text. But other practical applications include using AA for building a prosecution case against an online abuser. This is an important application, especially when we consider the raising trends in cyber-bullying and other electronic forms of teen violence[1].

Achieving good accuracy in AA on spontaneous online data is, however, far more challenging than the typical scenario for AA. One of the major complicating factors involves the limited amount of training data. In the typical scenario, we may have an entire document (several pages long), or even an entire book, while in the case of online data from social media we will have very short texts that are a couple of sentences long. Another challenge related to online data from social media is the number of potential authors that the model will need to learn. Consider aiming to do AA for data of online web forums, which is the goal in our work. In this case the potential author of a given post is one out of the thousands of registered users in that forum. In contrast, the majority of the text classification problems have a small number of classes. Lastly, we have to consider the problems with processing spontaneous written

[1] http://cyberbullying.us/index.php

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 156–164,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

language, and in particular, from informal interactions of web forum posts. The fact that the written fragments are informal is not by itself a complicating factor. We can argue that because they are informal they allow the writer to express more freely and thus they may contain more revealing information. But if we are to use syntactic analyzers to extract features for our learning model, as previous work in AA has done, then these informal and spontaneous samples can cause the analyzers to break, or output very noisy information.

This paper presents a new approach for AA on web forum data that generates informative meta features that can help discriminate the posts from different authors. The meta features in our work are derived from clustering of the feature vectors. However, different from distributional clustering and related approaches such as (Baker and McCallum, 1998; Slonim and Tishby, 2001; Dhillon et al., 2003), we do not cluster the features, but the instances in an unsupervised fashion. The goal of the meta features is to encode high level relations of similarities among posts from different authors, and not to reduce the feature set or find semantically related features as in the work listed above. Moreover, another difference and contribution of our work is the idea of generating modality specific meta features. This modality specific framework allows to reach higher AA accuracy on short texts than that achieved by the standard approach using only first level features and competitive state of the art approaches.

The question we aim to answer here is whether generating these metafeatures, which contribute to the computational cost, are indeed helpful in the scenario of AA on web forum posts. Our experiments are done on a much smaller scale than that of the real scenario, with data sets of up to 100 authors and in a closed-class setting. However, they represent the best results reported so far under similar conditions and thus they show promise to scale up. The results we present show that the modality specific meta features are indeed helpful for short online data, and outperform accuracy of previous work.

The next section reviews related work on AA. Then in Section 3 we discuss our approach to generating meta features from clustering the instances using different "views" of the posts. A discussion of the first level features is presented in Section 4. Section 5 presents the data sets used in our experiments. The evaluation of our approach is outlined in Section 6, where we also discuss our baseline system and results. The last section summarizes our findings and outlines our research goals for the immediate future.

2 Related Work

Authorship Attribution (AA) and related author analysis tasks, such as plagiarism detection and author profiling, have received a lot of attention recently, but most of the evaluation sets have a small number of authors. Here we highlight previous work that involves a large number of authors (at least 50) and refer the reader to the survey by (Stamatatos, 2009).

Luyckx and Daelemans studied the impact on accuracy of the number of potential authors and the size of the training data per author (Luyckx and Daelemans, 2010). They measured classification accuracy of a memory-based learner on three datasets with up to 145 candidate authors for one of them. The features used in their experiments include lexical features, such as word and lemma n-grams, type/token ratio, and readability measures; the syntactic features include Parts-of-Speech (POS), grammatical relations, chunk n-grams, and tokens with POS attached. They also used character n-grams, features that have been found to work well for AA (Peng et al., 2004; Plakias and Stamatatos, 2008). As expected, accuracy reported for 145 authors (1̃2%) was considerably lower than that achieved when the number of authors was smaller. An important characteristic of Luyckx and Daelemans work is that the 145 author set has only one document per author. In the experimental setup they partitioned each document into 10 fragments and used 9 of these fragments for training their model while testing on the remaining one. We believe that training a model on pieces of the same document used for testing is not exactly the task of AA, at least not in a realistic scenario. However, we recognize that the limited training data is an important constraint.

(Layton et al., 2010) shows results from using the Source Code Authorship Profile (SCAP) method on Twitter data where the microblogs are restricted to a maximum of 140 characters. The SCAP method, as developed by (Frantzeskou et al., 2007), determines authorship by measuring the overlap in character n-grams from the text document to the concatenated documents of each author. On a set with 50 authors, the SCAP method reached an accuracy of 55%, although when the @username was included in the text their accuracy increased to little over 70%. As the authors suggest, expecting to have this @username information from an author that wants to remain anonymous is not realistic.

Koppel *et al.* (2011) present a study of AA using blog data crawled from blogger.com. The approach used by them is based on computing cosine similarity from vectors of character 4-grams, and the number of candidate authors is by far the largest reported in the literature: for some experiments they trained on 10,000 authors and tested on 1,000. Precision and Recall in this setting were reported at 87.9% and 28.2%, respectively. However, we should also note that in these experiments text was also fragmented into snippets, and similar to what Luyckx and Daelemans did, the similarity model uses fragments of the same source text to predict authorship. In our opinion, the task resembles more a data provenance problem than an AA one. Moreover, because the blog data used in this work was not controlled for topic, and given that they used character 4-grams as features in a similarity based approach, we speculate that in the Koppel *et al.* setting there is more risk to bias the task from AA to a semantic or topic categorization and the only way to disentangle the two is by controlling for topic variation.

In another interesting recent work on AA, Probabilistic Context-Free Grammars (PCFGs) were proposed for AA (Raghavan et al., 2010). The number of authors in the evaluation data sets was rather small (3 to 6) but it included different domains, such as poetry, football, business, travel, and cricket, and all the data was harvested from the Internet. Raghavan et al. trained a PCFG for each author independently and authorship on the test data was assigned by taking the highest likelihood score from these grammars. To overcome the data sparseness problem, they mixed treebanked data from the Wall Street Journal (WSJ). They also enriched this mostly syntactic models with lexical information by combining the output with a bag-of-words Maximum Entropy classifier and a word-based n-gram language model. In their case, the combined model performed better than the baseline and the other machine learning approaches for most of the datasets. What is very interesting from this previous work is the fact that the same inexpensive PCFG-based approach worked reasonably well on all the data sets tested. In our experiments we evaluate this approach on the web forum data and the results show this to be a competitive method, even though the number of potential authors increased by a large margin and the documents are shorter than those used in (Raghavan et al., 2010).

3 Modality Specific Meta Features for Authorship Attribution

The standard formulation of text classification considers having a set of labeled examples, l, where each document is represented by a feature vector $\mathbf{x} \in R^n$ and their corresponding labels y, where $y_i \in \{0, 1\}$ in a binary classification. The feature vectors and their true class values are then input to a learning algorithm that will then build a model to predict the class of new instances. In contrast, we extract a set of smaller feature vectors that are then the basis for generating meta features, or more concretely, meta feature vectors.

Our approach starts with the extraction of first-level features to generate a feature vector representation for each instance. However, in our framework instead of having a single feature vector for $\mathbf{x}$, we generate m smaller vectors that contain complementary types of features, or views, describing the instances. We call these different views multimodal because they represent different characteristics of the text. More formally, an instance $\mathbf{x}$ is now represented as $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$ where each $\mathbf{x}_i$ is a vector with $|\mathbf{x}_i|$ features in modality i. Note that $\textbf{union}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m) = \mathbf{x}$ and $\textbf{intersection}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m) = \emptyset$ since we are only generating sub vectors (or complementary views) from the original feature set.

The generation of meta features uses these m different vectors to produce m clustering solutions for the training data with k clusters each. That means that we end up with different arrangements of the training instances into clusters, one arrangement per modality. Note that since clustering is performed per modality, k may be different in each clustering solution. From each cluster c_k in each of the m clustering solutions, we compute a centroid by averaging all the feature vectors in that cluster.

$$\textbf{centroid}_{m_i} = \frac{1}{\mid c_{m_i} \mid} \sum_{x_j \in c_{m_i}} \mathbf{x}_j \qquad (1)$$

where i above ranges from 1 to k, the number of clusters. We then measure the *similarity* of each instance to these centroids using the cosine function. These $m \times k$ similarity values are then used as the meta features, x', and we compute them for training and testing instances. Thus, as a result of this step each instance $\mathbf{x}$ is now represented by m tuples of vectors, the first level feature vectors $\langle \mathbf{x}_{i_1}, ..., \mathbf{x}_{i_{|x_i|}} \rangle$ and the newly generated meta feature vectors $\langle \mathbf{x'}_{i_1}, ..., \mathbf{x'}_{i_k} \rangle$ for each modality i.

In our problem of AA, we consider four types of first level features: stylistic features, lexical features, syntactic features, and perplexity values from character 3-gram language models. That is, in these experiments $m = 4$. Therefore, in our problem we have $x = \{\mathbf{x}_{sty}, \mathbf{x}_{lex}, \mathbf{x}_{ppl}, \mathbf{x}_{syn}\}$, where $\mathbf{x}_{sty}$ refers to the feature vector containing only stylistic features, $\mathbf{x}_{lex}$ is the vector for lexical features, $\mathbf{x}_{ppl}$ is the feature vector for perplexity values, and lastly, $\mathbf{x}_{syn}$ is the vector of syntactic features. Section 4 describes the features we are using in more detail.

To summarize, our MSMF approach is different from previous machine learning approaches to AA in that it has an intermediate step where we generate meta features from clustering the training instances per modality. Thus, all the vectors $\mathbf{x}_{sty}$ in the training data are input to a k-means clustering algorithm. Similarly, the set of vectors $\mathbf{x}_{lex}$, $\mathbf{x}_{ppl}$, and $\mathbf{x}_{syn}$ are clustered separately.

We are proposing to generate new meta features from clustering the data that can better represent posts from each author, but more important, the relation, i.e. closeness, to posts from other authors. It should be noted that no class information is used during clustering as the idea is to uncover regularities across the posts from authors on individual modalities as a result of the clustering. New in this work as well is the idea of a multimodal clustering, where each feature modality is clustered separately. Our assumption is that generating clusters by looking at feature sets separately will allow contrasting authors' characteristics in a subdimensional space without the risk of blurring differences, or similarities, across authors that can occur when clustering the entire feature vector at once. For instance, one author may have a similar style on the use of emoticons to a subset of authors while sharing similar syntactic characteristics to a very different subset of authors. This information, we hope, will be captured by the metafeatures, and will yield higher classification accuracy than the first level features by themselves.

4 First Level Features

The previous section motivated and described the use of the meta features. This section describes the first level features, where by first level features we refer to features computed directly from the documents.

Table 1 shows a list of the features used arranged by modality. For the *stylistic* modality we crafted a list of features tuned for written interactions in social networks. Thus, we use percentages of non-alphanumeric characters that are commonly used in emoticons. We also include percentages of capitalized words, use of quotations, and use of signature, that we believe allow writers more freedom to express their unique writing style. The *lexical* modality is the standard bag of words representation used in text classification that has also been commonly used in previous AA work (Argamon and Levitan, 2005; Zhao and Zobel, 2005). In the modality noted as *perplexity* in Table 1 we use perplexity values as computed by character 3-gram language models. We use the training data to train one language model per author and each model generates a perplexity, or cross entropy, value per instance. For training the language models and computing perplexity values we used the SRI-LM toolkit (Stolcke, 2002). Frequencies of character n-grams have also been successfully used to build author profiles (Keselj et al., 2003). However, to the best of our knowledge, this is the first work exploiting character-based language models for AA, although, Raghavan *et al.*'s work on PCFGs is closely related to this. Lastly, in the *syntactic* modality we have unigrams, bigrams, and trigrams of POS tags, and typed dependency relations extracted using the Stanford parser (Marneffe et al., 2006), that have been used before in AA.

5 Data Sets

To test our approach we downloaded posts from the Chronicle of Higher Education (CHE). Because our focus is on evaluating the use of metafeatures for the problem of AA in web forum posts, we need to control potential confounding characteristics in the data. Therefore, for our evaluation we downloaded posts from a single topic and generated 5 data sets with a different number

Modality	Features
Stylistic	Total number of words
	Average number of words per sentence
	Binary feature indicating use of quotations
	Binary feature indicating use of signature
	Percentage of all caps words
	Percentage of non-alphanumeric characters
	Percentage of sentence initial words with first letter capitalized
	Percentage of digits
	Number of new lines in the text
	Average number of punctuations (!?.;:,) per sentence
	Percentage of contractions (won't, can't)
	Percentage of two or more consecutive non-alphanumeric characters
Lexical	Bag of words (freq. of unigrams)
Perplexity	Perplexity values from character 3-grams
Syntactic	Part-of-Speech (POS) tags
	Dependency relations
	Chunks (unigram freq.)

Table 1: Feature breakdown by modality

of authors each. Table 2 shows some statistics on these data sets.

Because the forum is related to higher education it is expected that users of this forum will be more conscious about their writing and grammar. This is one of the reasons why we decided to start working on this data as a first cut on the problem of AA on online forums. However, it is still a spontaneous and informal setting. Table 3 shows some excerpts from the forum that show this to be a middle ground between carefully edited and written text and typical social media samples.

> "OH MY GOD. ARE YOU THE STUPIDEST MAN... WAIT STUPIDEST PERSON ON THE FACE OF THIS EARTH?
> "We've been married for xx years and you STILL can't figure it out? Look, here's a quarter. Why don't you call someone and see if they will help you PULL YOUR HEAD OUT OF YOUR A$$."

Table 3: Excerpts from the CHE forums

Our datasets are available to the research community by contacting the authors[2].

6 Empirical Evaluation

For all our experiments we chose a fixed partition of training and testing for all collections. We randomly divided each data set into 80% training and 20% testing. We are presenting results of using Support Vector Machines (SVMs) (Schölkopf and Smola, 2002) as the underlying learner as implemented in WEKA (Witten and Frank, 2005). We report classification accuracy as the evaluation metric.

6.1 Baseline Experiments for AA

The first set of experiments we present are aimed at establishing a good baseline for our approach. Following the baselines presented in previous work, we measure prediction performance for AA on the CHE collection using a bag-of-words approach.

Authors	Baseline
5	51.30
10	44.59
20	36.58
50	29.20
100	27.95

Table 4: Baseline accuracy using SVMs and bag of words for the CHE data set

The results are shown in Table 4. The baselines chosen are strong. Especially for the data set with 100 authors, where SVMs reached an accuracy of close to 30%, much higher than a simple majority predictor (1/100), but also considerably higher than that reported for datasets with a similar number of authors (Luyckx and Daelemans, 2010). As expected, accuracy drops as the number of potential authors increases, with the 100 authors data set posing the greatest challenge for the classifier.

6.2 First Level Features (FLF) for AA

Before we evaluate the meta features approach we want to assess the value of the first level features for this problem. The features described in Section

[2] Because the CHE data set exceeds the 10MB limit we were unable to upload it as supplementary material.

Dataset	Authors	Total Number of Posts	Avg Number of Posts per Author	Avg Number of Words per Author
1	5	2,889	577.8	39,664
2	10	5,579	557.9	40,953
3	20	9,779	488.95	35,838
4	50	15,543	310.86	21,502
5	100	16,171	161.71	11,322

Table 2: Summary of the CHE data set

4 were tailored to the web forum domain, therefore we expect them to be valuable for learning to discriminate the writeprint of each author. We used SVMs as the underlying algorithm. The results are shown in Table 5. In all cases the FLF outperformed the baseline results.

Authors	FLF Accuracy
5	69.21
10	70.81
20	67.06
50	60.12
100	57.78

Table 5: Results using First Level Features (FLF) and SVMs for the CHE data set

These results are higher than what has been reported on AA for a similar number of authors. The FLF have shown to be competitive and in some cases the improvement in accuracy over the baseline reaches 100%. In most cases accuracy decreased with a larger number of potential authors, although, for the data set with 10 authors accuracy was a little bit higher than with 5 authors. Moreover, the drop in accuracy is not as pronounced as in the baseline system, suggesting that BOWs are not sufficient to solve this task and that a combination of features, such as those included in our FLF are more appropriate for this problem.

6.3 Using Modality Specific Meta Features (MSMF)

After establishing the baseline performance in our data set, and the performance of using only FLF, we want to evaluate the idea of generating meta features that are modality specific. As described in Section 3, we cluster each of the four types of feature vectors in the training data set separately. Because we use a k-means clustering algorithm, implemented in CLUTO, the first step is to choose the number of clusters. Determining the optimal number of clusters is challenging and beyond the scope of this exploratory work. But it is still an important parameter in our solution since the value of k determines the number of meta features generated per modality. The role of these meta features is to extract relations among the posts of different authors on a given modality. A reasonable assumption is then to set k as a function of the number of authors. We experimented setting k =number of authors $\times n$, with values of $n = 1, 3, 5, 10, 15$. For example, for the data set with 5 authors we experimented with values of $k = 5, 15, 25, 50, 75$.

Authors	K	MSMF	FLF	MSMF+FLF
5	1×5	45.04	69.21	73.39
	3×5	50.95		74.6
	5×5	53.91		74.08
	10×5	62.60		75.47
	15×5	65.04		**76.17**
10	1×10	37.47	70.81	77.38
	3×10	47.29		75.85
	5×10	50.09		76.3
	10×10	61.16		**77.38**
	15×10	59.54		76.84
20	1×20	35.35	67.06	70.81
	3×20	40.03		71.22
	5×20	43.78		71.37
	10×20	48.40		**71.42**
	15×20	49.58		70.96
50	1×50	32.77	60.12	63.20
	3×50	37.66		62.75
	5×50	39.83		63.72
	10×50	43.50		**63.79**
	15×50	44.53		63.33
100	1×100	33.15	57.78	60.41
	3×100	40.11		60.95
	5×100	42.02		61.17
	10×100	42.52		**62.10**
	15×100	43.34		59.54

Table 6: Accuracy results for AA on the CHE collection when using modality specific metafeatures (MSMF), first level features (FLF) and the combination of both (MSMF+FLF)

Table 6 summarizes our results showing accuracy values. For comparison purposes we include in this table results from using only first level features (FLF), only modality specific meta features (MSMF), and the combination of both (MSMF+FLF). The results show several consistent trends across all 5 data sets. First, meta features by themselves are always outperformed by the first level features. This is not surprising since the meta features aggregate posts from different authors depending on similarity and thus predicting authorship only on these features does not work as well as the standard approach of using first level features. However, these meta features do outperform the bag of words baseline results (compare column MSMF in Table 6 with results shown in Table 4), underscoring the fact that the CHE data set is harder than the typical text classification task where the lexical features by themselves can solve the problem with very high accuracy. Moreover, the combination of first level features and meta features (MSMF+FLF) consistently achieves higher accuracy than any of the other two alternatives, this is the second trend and the most relevant to our work. These results show that the modality specific meta features are important and yield improvements of up to 10% in accuracy over the standard approach of using only FLF, and of more than 100% in accuracy over a strong bag of words baseline. Third, with respect to the value of k, the results show that for all values chosen, the MSMF outperforms the baseline results, and that using the combined set of features (MSMF+FLF) will yield a higher accuracy than that of using only the first level features. However, it does seem that higher values of k result in higher accuracy, suggesting that trying to find more clusters, and therefore finer-grained clusters in the data is resulting in the extraction of more meaningful relations among the posts of different authors. The results also show that the best k overall was $k = 10\times$ number of authors. For larger k values only the data set with 5 authors reached better results. Overall, it is interesting to see as well that both types of features yield classifiers that are less affected by the larger number of authors, as the drop in accuracy seems to be less pronounced than in the baseline system (see Table 4).

Our previous results show that the meta features contribute to a better prediction of authorship. But what about the multi modal framework? In order to assess if generating modality specific meta features is helpful we performed additional experiments where all instances are represented by a single vector that concatenates all modality feature vectors. The rest of the meta features approach remains unchanged, the vectors are clustered using k-means clustering and we generate metafeatures for each instance. The results are shown in Table 7 and for all cases we chose k = number of authors$\times 10$. The results under AMMF are the meta features generated without separate processing per modality, AMMF+FLF shows results of combining first level features with "all modalities together" meta features. As we speculated, there is a considerable gain in accuracy from the independent processing per modality. The gain in accuracy of MSMF over AMMF ranges from 73% to $\sim 250\%$. This gain possibly comes from the ability to aggregate feature vectors that semantically represent the same type of information, which can be difficult to maintain when all modalities are grouped together. Both approaches improve accuracy when they are combined with FLF, but again the combination using modality specific meta features (MSMF+FLF) yields the best results. However, the gain in accuracy observed when going from AMMF+FLF (Column 4 in Table 7) to MSMF+FLF (Column 5 in Table 7) is not as large as that observed when going from using only AMMF (Column 2 in Table 7) to MSMF (Column 3 in Table 7). This is expected since both approaches share the same set of FLF, which we know are by themselves very powerful. Further experiments and analysis are needed to better characterize the advantages of the MSMF approach. We plan to leave this for future work.

6.4 Benchmark Comparisons

To explore further the intuition that our approach is a good alternative for AA on web forum data we performed additional experiments where we evaluated the PCFG-based approaches in (Raghavan et al., 2010). In their experiments they have three systems: one is the standard PCFG approach, noted as PCFG in Table 8, the second version uses treebanked data from the WSJ mixed with the original CHE data. This interpolated version is called PCFG-I in that table. We followed the same approach of training the parser on the first 10 sections of the WSJ. For the interpolation, we added Section 20 of the WSJ and replicated the original author's data twice. The third version, noted as PCFG-E, is the combination of the PCFG with the bag-of-words MaxEnt model, and an n-gram language model. The results in Table 8 show that the best accuracy in the CHE collection is achieved by our method in all four data sets. For comparison purposes we also included here the baseline results

Authors	**AMMF**	**MSMF**(% gain)	**AMMF+FLF**	**MSMF+FLF**(% gain)
5	36.00	62.60 (73%)	71.47	**75.47**(5%)
10	19.63	61.16 (211%)	70.99	**77.38**(9%)
20	19.63	48.40 (146%)	69.06	**71.42**(3%)
50	12.44	43.50 (249%)	61.65	**63.79**(3%)
100	15.07	42.52 (182%)	59.16	**62.10**(4%)

Table 7: Accuracy comparison on CHE data set between generating modality specific meta features (MSMF) and meta features with all modalities together (AMMF), and the combination of each with first level features (FLF).

Approach	**5 Authors**	**10 Authors**	**50 Authors**	**100 Authors**
SVM	51.3	44.59	29.20	27.95
PCFG	62.95	58.46	31.41	29.77
PCFG-I	64.17	61.26	46.02	44.43
PCFG-E	64.00	55.85	36.11	34.72
Our Approach	**75.47**	**77.38**	**63.79**	**62.10**

Table 8: Benchmark comparison of AA accuracy on the CHE collection

shown in Table 4. The baseline results are consistently outperformed by all of the PCFG-based approaches, showing yet again PCFGs to be robust to different genres but more important, to scale up well to a larger number of authors. However, the results were considerably lower than those of our method. These results support our hypothesis that the modality specific meta features are appropriate for online forum data where the documents are short, the number of potential authors is larger, the stylistic features are more discriminative, and there are less restrictions with respect to standards of writing. Another interesting finding from these experiments is the fact that the PCFG-I method always reached higher accuracies than the ensemble in the CHE collection. In Raghavan *et al.*'s collection, the ensemble (PCFG-E) was the most accurate model. We believe this difference is because of the fact that the CHE collection is single topic, having a more semantically uniform collection prevented the lexical-based components, such as bag of words and n-gram language models, used in the ensemble to boost accuracy.

7 Concluding Remarks

Following recommendations from previous work in AA, we have gathered a single topic evaluation data set of web forum posts with up to 100 candidate authors. The main contribution of this work is the use of modality specific meta features generated by an unsupervised approach. Previous work has used distributional clustering to aggregate features that share the same relation with the class (Baker and McCallum, 1998; Slonim and Tishby, 2001; Dhillon et al., 2003). Our proposed framework is different, we generate meta features from similarity metrics between centroids from an unsupervised clustering of instances and the instances themselves. The additional cost in clustering instances shows to be valuable for AA as we can gain up to 100% improvements in accuracy over strong baselines. Further analysis of results also showed that treating each modality separately to generate the meta features is also important and can yield gains of close to 10% in accuracy over the standard approach of using only first level features. To the best of our knowledge, this is by far the best result reported for AA in a task having up to 100 authors. The framework is general enough that it can be extended to other classification problems where instances can be represented using different modalities.

The experimental evaluation presented here shows that a relatively inexpensive approach based on PCFGs can scale up to a larger number of authors, even if the documents are only a couple of sentences long. However, this syntactically driven approach is outperformed by our proposed modality specific meta features framework.

The results are very promising although we recognize that this is not yet a real world scenario for web forum data, so we are currently gathering data sets with a larger number of authors. We also want to evaluate this work on different data sets to analyze the robustness and suitability of this method. Lastly, we want to study the effect of having more than one topic in the data set as in the work of (Schein et al., 2010).

Acknowledgements

This work was partially supported by a UAB faculty development grant to the first author. It was also supported in part by the CONACYT-Mexico (project no. 134186) and by the European Commission as part of the WIQ-EI project (project no. 269180) within the FP7 People Programme.

References

S. Argamon and S. Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*.

L. Douglas Baker and Andrew McCallum. 1998. Distributional clustering of words for text classification. In *SIGIR 98: Proceedings of the 21st Annual International ACM SIGIR*, pages 96–103, Melbourne, Australia, August. ACM.

Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. 2003. A divisive information-theoretic feature clsutering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287.

G. Frantzeskou, E. Stamatatos, S. Gritzalis, and C. E. Chaski. 2007. Identifying authorship by byte-level n-grams: The source code author profile (SCAP). *Journal of Digital Evidence*, 6(1).

Graeme Hirst and Ol'ga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination. *Literary and Linguistic Computing*, 22(4):405–417, October.

V. Keselj, F. Peng, N. Cercone, and C. Thomas. 2003. N-gram based author profiles for authorship attribution. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 255–264.

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

Robert Layton, Paul Watters, and Richard Dazeley. 2010. Authorship attribution for twitter in 140 characters or less. In *Second Cybercrime and Trustworthy Computing Workshop, CTC 2010*, pages 1–8, Ballart, VIC, Australia, July.

Kim Luyckx and Walter Daelemans. 2010. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, pages 1–21, August.

M.C. De Marneffe, Bill Maccartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

T.C. Mendenhall. 1887. The characterstic curves of composition. *Science*, IX:237–249.

F. Mosteller and D. L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

F. Peng, D. Shuurmans, and S. Wang. 2004. Augmenting naive Bayes classifiers with statistical language models. *Information Retrieval Journal*, 7(1):317–345.

S. Plakias and E. Stamatatos. 2008. Tensor space models for authorship attribution. In *Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications*, volume 5138 of *LNCS*, pages 239–249, Syros, Greece.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42, Uppsala, Sweden, July. Association for Computational Linguistics.

Andrew I. Schein, Johnnie F. Caver, Randale J. Honaker, and Craig H. Martell. 2010. Author attribution evaluation with novel topic cross-validation. In *The 2010 International Conference on Knowledge Discovery and Information Retrieval*, Valencia, Spain, October.

Bernhard Schölkopf and Alexander J. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press.

Noam Slonim and Naftali Tishby. 2001. The power of word clusters for text classification. In *23rd European Colloquium on Information Retrieval Research (ECIR)*.

Efstathios Stamatatos. 2009. A survey on modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. pages 901–904.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kauffmann, 2nd edition.

Y. Zhao and J. Zobel. 2005. Effective and scalable authorship attribution using function words. In *Proceedings of 2nd Asian Information Retrieval Symposium*, volume 3689 of *LNCS*, pages 174–189, Jeju Island, Korea.