Author Verification Using Common N-Gram Profiles of Text Documents

Magdalena Jankowska and **Evangelos Milios** and **Vlado Kešelj**
Faculty of Computer Science, Dalhousie University
6050 University Avenue
Halifax, NS B3H 4R2, Canada
{jankowsk, eem, vlado}@cs.dal.ca

Abstract

Authorship verification is the problem of answering the question whether or not a sample text document was written by a specific person, given a few other documents known to be authored by them. We propose a proximity based method for one-class classification that applies the Common N-Gram (CNG) dissimilarity measure. The CNG dissimilarity (Kešelj et al., 2003) is based on the differences in the frequencies of n-grams of tokens (characters, words) that are most common in the considered documents. Our method utilizes the pairs of most dissimilar documents among documents of known authorship. We evaluate various variants of the method in the setting of a single classifier or an ensemble of classifiers, on a multilingual authorship verification corpus of the PAN 2013 Author Identification evaluation framework. Our method yields competitive results when compared to the results achieved by the participants of the PAN 2013 competition on the entire set, as well as separately on two subsets — English and Spanish ones — out of the three language subsets of the corpus.

1 Introduction

The task of computational detection of who wrote a given text is a widely studied linguistic and machine learning problem with applications in domains such as forensics, security, criminal and civil law, or literary research. The authorship verification problem is a type of such a computational authorship analysis task, in which, given a set of documents written by one author, and a sample document, we are asked whether or not this sample document was written by this given author. This is different from the more traditional problem of deciding who among a finite number of candidate authors for which we are given sample writings, wrote a document in question, and, albeit more difficult, is often considered to better reflect the real-life problems related to authorship detection (Koppel et al., 2012).

We describe our one-class proximity based classification method and evaluate it on the multilingual dataset of the Authorship Identification competition task of PAN 2013 (evaluation lab on uncovering plagiarism, authorship, and social software misuse) (Juola and Stamatatos, 2013).

During the competition, to which a variant of our method has been submitted (Jankowska et al., 2013), it yielded ranking 5th (joint) out of 18 with respect to the accuracy, and 1st rank out of 10 in the secondary ranking based on the area under the ROC curve (AUC), which evaluates the ordering of instances by the confidence score. In this paper we show some further experiments on how a different way of tuning the classifier parameters, using solely the training dataset of the competition, as well as an ensemble of classifiers based on our method, without any parameter tuning, leads to competitive accuracy results while still achieving high AUC values.

2 Related Work

The author analysis has been studied extensively in the context of the authorship attribution problem, in which there is a small set of candidate authors out of which the author of a questioned document is to

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 387–397, Dublin, Ireland, August 23-29 2014.

be selected. There are several papers (Stamatatos, 2009; Juola, 2008; Koppel et al., 2009) presenting excellent surveys of this area.

The two main categories (Stamatatos, 2009) of solutions for the problem are similarity based approaches, in which a classification is performed in a Neigherst Neighbour scheme, attributing a sample text to the author whose writing is most similar according to some measure, and machine-learning based approaches, in which each document by an author is treated as a data sample within a class, and a supervised classifier is trained on these data.

A more limited research has been performed on an open-set variant on this problem, in which it is possible that none of the candidate authors wrote a document in question, with authorship verification being the extreme case of an open-set problem with only one candidate. The "unmasking method" for authorship verification (Koppel and Schler, 2004) is successful for novel-length texts. This approach, similarly as our method, falls into a category of *intrinsic* methods (Juola and Stamatatos, 2013); it uses only the documents in question, without constructing classes of other authors. The ensemble of one-class classifiers (Halvani et al., 2013), which achieved high accuracy at the PAN 2013 Author Identification competition, is also an example of such an intrinsic method. It varies from our approach by using a different scheme of creating the dissimilarity between an unknown document and the known authorship set of texts, based on the Nearest Neighbour technique (Tax, 2001), as well as by a different distance measure and features used.

Another way of approaching the author verification problem is to cast it into a binary or multi-class classification, by creating a class or classes of other authors. The "imposters" method (Koppel and Winter, 2014) generates a very large set of texts by authors that did not write the questioned document, to transform the problem into a open-set author attribution problem with many candidates, handled by an ensemble-based similarity method (Koppel et al., 2011). A modified version of the imposters method (Seidman, 2013) achieved first ranking in the PAN 2013 Authorship Identification competition. The method (Veenman and Li, 2013), which achieved the highest accuracy on the English set in this competition, is also of such an *extrinsic* type; its first step is a careful selection of online documents similar to the ones in the problems. The method (Ghaeini, 2013), which produces competitive ordering of verification instances, uses weighted k-NN approach using classes of other authors created from other verification instances.

3 Methodology

The formulation of the authorship verification task for the Author Identification Task at PAN 2013 is the following: "Given a set of documents (no more than 10, possibly only one) by the same author, is an additional (out-of-set) document also by that author?" (Juola and Stamatatos, 2013).

We approach this task with an algorithm based on the idea of proximity based methods for one-class classification. In one-class classification framework, an object is classified as belonging or not belonging to a target class, while only sample examples of objects from the target class are available during the training phase. Our method resembles the idea of the k-centers algorithm for one-class classification (Ypma et al., 1998; Tax, 2001), with k being equal to the number of all training documents in the target set (i.e., written by the given author). The k-centers algorithm is suitable for cases when there are many data points from the target class; it uses equal radius sphere boundaries around the target data points and compares the sample document to the closest such centre. We propose a different classification condition, described below, utilizing the pairs of most dissimilar documents within the set of known documents.

Let $A = \{d_1, ..., d_k\}$, $k \geq 2$, be the input set of documents written by a given author, which we will call *known documents*. If only one known document is provided, we split it in half and treat these two chunks as two known documents. Let u be the input sample document, of which the authorship we are to verify, that is return the answer "Yes" or "No" to the posed question whether it was written by the given author.

Our algorithm calculates for each known document $d_i, i = 1, 2, ..., k$, the maximum dissimilarity between this document and all other known documents: $D^{max}(d_i, A)$, as well as the dissimilar-

ity between this document and the sample document u: $D(d_i, u)$, and finally the dissimilarity ratio $r(d_i, u, A) = \frac{D(d_i,u)}{D^{max}(d_i,A)}$ (and thus $r(d_i, u, A) < 1$ means that there exists a known document more dissimilar to d_i than u, while $r(d_i, u, A) > 1$ means that all the known documents are more similar to d_i than u). The average $M(u, A)$ of the dissimilarity ratio over all known documents $d_1, d_2, ..., d_k$ from A, is the subject of the thresholding: the sample u is classified as written by the same person as the known documents if and only if $M(u, A)$ is at most equal to a selected threshold θ. Notice that in this framework the dissimilarity between the documents does not need to be a metric distance, i.e., it does not need to fulfil the triangle inequality (as is the case for the dissimilarity measure we choose).

For the dissimilarity measure between documents we use the Common N-Gram (CNG) dissimilarity; proposed by Kešelj et al. (2003); this dissimilarity (or its variants) used in the Nearest Neighbour classification scheme (Common N-gram classifier) was successfully applied to authorship classification tasks (Kešelj et al., 2003; Juola, 2008; Stamatatos, 2007). The CNG dissimilarity is based on the differences in the usage frequencies of the most common n-grams of tokens (usually characters, but possibly other tokens) of the documents. Each document is represented by a *profile*: a sequence of the most common character n-grams (strings of characters of the given length n from the document) coupled with their frequencies (normalized by the length of the document). The dissimilarity between two documents of the profiles P_1 and P_2 is defined as follows:

$$D(P_1, P_2) = \sum_{x \in (P_1 \cup P_2)} \left(\frac{f_{P_1}(x) - f_{P_2}(x)}{\frac{f_{P_1}(x) + f_{P_2}(x)}{2}} \right)^2 \quad (1)$$

where x is a character n-gram from the union of two profiles, and $f_{P_i}(x)$ is the normalized frequency of the n-gram x in the the profile P_i, $i = 1, 2$ ($f_{P_i}(x) = 0$ whenever x does not appear in the profile P_i). The parameters of the dissimilarity are the length of the n-grams n and the length of the profile L. As our method is based on the ratios of dissimilarities between documents, we take care that the documents in a given problem are always represented by profiles of the same length. We experiment with two ways of selecting the length of the profiles. In the dynamic-length variant, the length of profiles is selected separately for each problem, based on the number of n-grams in the documents in the given instance (parametrized as a fraction f of all n-grams of the document that contains the least number of them). In the fixed-length variant, we use a selected fixed length L of profiles. For a one-class classifier we need to select two parameters defining the features used for dissimilarity (length of the n-grams n, and either the fixed length L of a profile, or the fraction f defining the profile length), and the parameter θ (for classifying by thresholding the average dissimilarity ratio M).

We linearly scale the measure M to represent it as a confidence score in the range from 0 (the highest confidence in the answer "No") to 1 (the highest confidence in the answer "Yes"), with the answer "Yes" given if and only if the confidence score is at least 0.5. The value of M equal to θ is transformed to the score 0.5, values greater than θ to the scores between 0 and 0.5, and values less than θ to the scores between 0.5 and 1 (a cutoff of 0.1 is applied, , i.e. all values of $M(u, A) < \theta - \textit{cutoff}$ are mapped to the score 1, and all values of $M(u, A) > \theta + \textit{cutoff}$ are mapped to the score 0).

4 Training and test datasets

We leverage the evaluation framework of the PAN 2013 competition task of Author Identification (Juola and Stamatatos, 2013), the datasets of which were carefully created for authorship verification, with effort made to match within each problem instance the texts by the same genre, register, theme and time of writing. The dataset consists of English, Greek and Spanish subsets. In each instance, the number of documents of known authorship is not greater than 10 (possibly only one). The dataset is divided into the training set `pan13-ai-train` and the test set `pan13-ai-test`. The training set was made available for the participants before the competition; the test set was used to evaluate the submissions and subsequently published (PAN, 2013).

To enrich the training dataset for our competition submission, we also compiled ourselves two additional datasets using existing sets for other authorship identification tasks. `mod-pan12-aa-EN` is

an English author verification set compiled from the fiction corpus for the Traditional Authorship Attribution sub task of the PAN 2012 competition (PAN, 2012; Juola, 2012). `mod-Bpc-GR` is a Greek author verification set compiled from the Greek dataset of journal articles (Stamatatos et al., 2000). It is important to note that these sets are different from the competition dataset in that we did not attempt to match the theme or time of writing of the texts.

Table 1 presents characteristics of the datasets.

	`pan13-ai-train`			
	total	English	Spanish	Greek
number of problems	35	10	5	20
mean of the known document number per problem	4.4	3.2	2.4	5.5
mean length of documents in words	1226	1038	653	1362
genre		textbooks	editorials, fiction	articles
	`pan13-ai-test`			
	total	English	Spanish	Greek
number of problems	85	30	25	30
mean of the known document number per problem	4.1	4.2	3.0	4.9
mean length of documents in words	1163	1043	890	1423
genre		textbooks	editorials, fiction	articles
	`mod-pan12-aa-EN` total: English			
number of problems	22			
mean of the known document number per problem	2.0			
mean length of documents in words	4799			
genre	fiction			
	`mod-Bpc-GR` total: Greek			
number of problems	76			
mean of the known document number per problem	2.5			
mean length of documents in words	1120			
genre	articles			

Table 1: Characteristics of datasets used in our authorship verification experiments.

5 Evaluation measures

In our experiments we use two measures of evaluation, based on the measures proposed for the PAN 2013 competition. The accuracy is the fraction of all problems that have been answered correctly. The AUC measure is the area under the ROC curve based on the confidence scores. It is the nature of applications of authorship verification, such as forensics, that makes the confidence score and not only the binary answer, an important aspect of a solution (Gollub et al., 2013).

For our method accuracy is equivalent to the measure that was used in the competition for the main evaluation. This measure is F_1, defined based on the fact that in the competition it was allowed to withdraw an answer (i.e., use an "I do not know" option). Precision and recall were defined as follows: $recall = \frac{\#correct\_answers}{\#problems}$, $precision = \frac{\#correct\_answers}{\#answers}$, and F_1 is the harmonic mean of precision and recall. For any method that, as our method, provides the answer "Yes" or "No" for all problem instances, the accuracy and F_1 are equivalent.

6 Types of classifiers

A single classifier of our method requires two parameters defining the features to be used to represent a document (the length of an n-gram and the length of a profile), as well as a selection of the threshold for the dissimilarity for the classification decision. We tune and evaluate four version of such single classifiers. Combining many such one-class classifiers, each using different combination of features defining parameters, into one ensemble, allows to remove or mitigate the parameter tuning. We describe the creation and the evaluation of four types of ensembles.

Table 2 reports the considered space for feature defining parameters. On a training set, for a given combination of feature defining parameters (n,L) or (n,f), we use the accuracy at the optimal threshold (a threshold θ that maximizes the accuracy), as a measure of performance for these parameters.

Parameters	
n	length of n-grams
L	# of n-grams: profile length (fixed-length)
f	fraction of n-grams for profile length (dynamic-length)
θ	threshold for classification
θ_{2+}	threshold for classification if at least 2 known documents are given
θ_1	threshold for classification if only one known document is given

Space of considered parameters	
n for character n-grams	$\{3, 4, ..., 9, 10\}$
n for word n-grams	$\{1, 2, 3\}$
L	$\{200, 500, 1000, 1500, 2000, 2500, 3000\}$
f	$\{0.2, 0.3, ..., 0.9, 1\}$

single classifiers		English	Spanish	Greek
vD1	n	6	7	10
	f	0.75		
	θ	1.02	1.005	1.002
vF1	n	6		7
	L	2000		2000
	θ_{2+}	1.02		1.008
	θ_1	1.06		1.04
vF2	n	7	3	9
	L	3000	2000	3000
	θ_{2+}	1.014	1.014	0.997
	θ_1	1.056	1.126	1.060
vD2	n	7	3	9
	f	0.8	0.6	0.8
	θ_{2+}	1.013	1.00530207	0.9966
	θ_1	1.053	1.089	1.059

ensembles		English	Spanish	Greek
eC	type	character		
	(n, L)	all in the considered space		
	θ	1		
eW	type	word		
	(n, L)	all in the considered space		
	θ	1		
eCW	type	character, word		
	(n, L)	all in the considered space		
	θ	1		
eCW	type	character, word		
	(n, L)	selected based on training data		
		(61)	(75)	(43)
	θ	1		

Table 2: Parameters for four variants of single one-class classifiers and four ensembles of one-class classifiers based on our method.

6.1 Single classifiers

For single character n-gram classifiers, we tuned the parameters for each language separately on training data, by selecting feature defining parameters based on their performance, and selecting the thresholds

to correspond to the optimal thresholds. Table 2 reports the parameters of four variants of single classifiers. We include our two submissions to the PAN 2013 Authorship Identification competition: the final submission `vF1` and the preliminary submission `vD1`. The other two classifiers were tuned and tested after the competition.

Our preliminary submission `vD1` (Table 2) is tuned on `pan13-ai-train`, with f chosen ad-hoc. This is the only classifier among the reported variants that does not use a preprocessing of truncation of all documents in a given problem instance to the length of the shortest document, which tend to increase the accuracy for cases of a significant difference in the length of documents.

For tuning of parameters of the final submission `vF1` (Table 2) we use not only `pan13-ai-train`, but also additional training sets `mod-pan12-aa-EN` and `mod-Bpc-GR`. We also introduce two threshold values: one for cases when there are at least two known documents, and another one for the cases when there is only one known document (which has to be divided in two). The intuition behind this double threshold approach is that when there is only one known document, the two halves of it can be more similar to each other than in other cases. After the parameters are selected based on subsets of training sets with only these problems that contain at least two known documents, the additional threshold is selected based on the optimal threshold on a modified "1-only" training set, from the problem of which all known documents except of a random single one is removed. For Spanish, with only three training instances with more than one known document, we use the same parameters as for English.

For tuning of `vF2` and `vD2` (Table 2) we use only competition training data, without the additional corpora used for `vF1`. Feature parameters are selected based on the performance on the subsets containing at least two known documents, and on the "1-only" modified sets (which allows us to use the Spanish training set for tuning the Spanish classifiers).

6.2 Ensembles of classifiers

We test ensembles of single one-class classifiers based on our method, with the ensemble combining answers of the classifiers, and each classifier using different set of features. An important advantage of an ensemble is the alleviation of the problem of tuning the parameters. Each classifier uses a different combination of parameters n and L defining the features. And as many classifiers are used, instead of tuning the threshold of a single classifier based on some training data, the threshold of each classifier is set to some fixed value, with 1 being a natural choice, as it corresponds to checking whether or not the unknown document is (on average) less similar to each given known document than the author's document that is most dissimilar to this given known document.

We test majority voting and voting weighted by the confidence scores of single classifiers. For each ensemble we combine answers of the classifiers in order to obtain the confidence score of the ensemble. For majority voting the confidence score of the ensemble is the ratio of the number of classifiers that output "Yes" to the total number of classifiers, the confidence score of the weighted voting is the average of the confidence scores of the single classifiers.

We experiment with n-grams being characters (utf8-encoded) and words (converted to uppercase). Table 2 summarize the ensembles. The ensemble `eC` is of all character n-gram classifiers in our space of considered parameters n and L; `eW` is of all word n-gram classifiers; `eCW` is of all classifiers of `eC` and `eW`. These ensembles do not use any training data. We also create a classifier `eCW_sel` (Table 2), which is a subset of the classifiers of `eCW`, selected based on the performance of the single classifiers on the training data of the competition. For each language separately, we remove classifiers that on the training data achieved lowest accuracies at their respective optimal thresholds, while keeping at least half of the character based classifiers and at least half of the word based classifiers. (For Spanish, `eCW_sel` and `eCW` differ just by one classifier: the only one that on the small Spanish training set has the optimal accuracy less than 1.)

7 Results

The accuracy and the area under the ROC curve (AUC) values achieved by the variants of our method on the PAN 2013 Author Identification test dataset are presented in Table 3. The table states also the

best PAN 2013 competition results of other participants[1] (that is the results of these participants that achieved the highest accuracy or AUC on any (sub)set). There were 17 other participants for which there are accuracy (or F_1) results, 9 of which submitted also confidence scores evaluated by AUC.

		PAN 2013 Author Identification test dataset							
		F_1 = accuracy except for Ghaeini,2013				AUC			
		all	English	Spanish	Greek	all	English	Spanish	Greek
single classifiers									
vD1		0.718	0.733	0.760	0.667	0.790	0.837	0.846	0.718
vF1		0.682	0.733	0.720	0.600	0.793	0.839	0.859	0.711
vD2		0.729	0.767	0.760	0.667	0.805	0.850	**0.936**	0.704
vF2		0.753	0.767	**0.880**	0.633	**0.810**	0.844	0.885	0.664
ensembles of classifiers									
eC	majority	0.729	0.800	0.840	0.567	0.754	0.777	0.833	0.620
	weight	0.729	**0.833**	0.800	0.567	0.764	0.830	0.859	0.582
eW	majority	0.718	0.733	0.720	0.700	0.763	0.830	0.805	0.700
	weight	0.741	0.767	0.760	0.700	0.822	**0.886**	0.853	0.782
eCW	majority	**0.800**	**0.833**	0.840	0.733	0.755	0.817	0.821	0.633
	weight	0.741	0.800	0.840	0.600	0.780	0.842	0.853	0.622
eCW_sel	majority	**0.800**	**0.833**	0.840	0.733	0.778	0.826	0.814	0.682
	weight	0.788	0.800	0.840	0.733	0.805	0.857	0.853	0.687
boxed values: best competition results of other PAN 2013 Author Identification participants									
Seidman,2013		0.753	0.800	0.600	**0.833**	0.735	0.792	0.583	**0.824**
Veenman and Li,2013		–	0.800	–	–	–	–	–	–
Halvani et al.,2013		0.718	0.700	0.840	0.633	–	–	–	–
Ghaeini,2013		0.606	0.691	0.667	0.461	0.729	0.837	0.926	0.527

Table 3: Area under the ROC curve (AUC) and F_1 (which is equal to accuracy for all algorithms except for (Ghaeini, 2013)) on the test dataset of PAN 2013 Author Identification competition task. Results of variants of our method compared with competition results of those among other competition participants that achieved the highest value of any evaluation measure on any (sub)set. The highest result in any category is bold; the highest result by other competition participants in any category is boxed.

All variants of our method perform better on the English and Spanish subset than on the Greek one, both in terms of the accuracy and in terms of AUC. On the Greek subset they are all outperformed by other competition participant(s). This is most likely due to the fact that the Greek subset was created in a way that makes it especially difficult for algorithms that are based on CNG character-based dissimilarity (Juola and Stamatatos, 2013), by using a variant of CNG dissimilarity for the character 3-grams in order to select difficult cases. This particularity of the set may also be the reason why the ensemble eC of character n-gram classifiers performed worse than other methods on this set.

The variants of our method are competitive in terms of the ordering of the verification instances according to the confidence score as measured by AUC. During the competition, our final submission vF1 achieved the first ranking according to the AUC on the entire set, the highest AUC on the English subset, and the second-highest AUC values on the Spanish and Greek subset, out of 10 participants that submit-

[1]The results of our methods are on the published competition dataset. The results by other participants are the published competition results. The actual competition evaluation set for Spanish may have some text in a different encoding then the published set; our final submission method vF1 yielded on it a different result than on the published dataset.

ted confidence scores. All variants of our method perform better than any other competition participant on the entire set. On the English subset the single classifiers and the ensembles with weighted voting have AUC above 0.8, and out of those only eC has AUC lower than the best result by other participants. On the Spanish subset all variants of our method achieved AUC above 0.8, with vD2 achieving AUC higher than the best competition result on this subset.

In terms of overall accuracy on the entire set, the ensembles combining character and word based classifiers: eCW with majority voting and eCW_sel with both types of voting, achieve accuracy higher then the best overall accuracy in the competition. They also match or surpass the best competition accuracy on the English subset, and match the best competition accuracy on the Spanish subset. The highest accuracy on the English subset was achieved by eC with weighted voting, eCW with majority voting, and eCW_sel with majority voting (higher than the best competition result). vF2 yields on the Spanish subset accuracy higher than the best competition result.

For the ensembles of classifiers, on the English and Spanish subsets, the AUC for voting weighted by the confidence scores are higher than the AUC for the majority voting, but not so on the Greek subset. This is consistent with the fact that on the Greek subset the confidence scores for single classifier variants yield worse ordering (AUC) than on other sets. Creation of eCW_sel by removing from the ensemble eCW the classifiers that perform worst on the training data improves the Greek results, and slightly the English results.

We tested the statistical significance of accuracy differences between all pairs of accuracies reported in Table 3 by the exact binomial McNemar's test (Dietterich, 1998). Only few of these differences are statistically significant. On the entire set these are: the difference between the accuracy of eCW with majority voting and of eC with majority voting, vD1 and vF1, as well as the difference between the accuracies of eCW_sel with weighted voting and of vF1. On the Greek subset, this is the difference between the accuracies of the submission (Seidman, 2013) and the lower accuracy of eC with weighted voting.

		English mod-pan12-aa-EN		Greek mod-Bpc-GR	
		accuracy	AUC	accuracy	AUC
vD1		0.545	0.649	0.605	0.661
vD2		0.727	0.826	0.566	0.698
vF2		**0.773**	**0.843**	0.618	0.709
eC	majority	0.636	**0.843**	0.658	0.694
	weighted	0.682	0.806	0.671	0.703
eW	majority	0.636	0.674	**0.750**	**0.757**
	weighted	0.727	0.736	0.737	0.749
eCW	majority	0.636	0.785	0.737	0.725
	weighted	0.682	0.818	0.711	0.719
eCW_sel	majority	0.636	0.789	**0.750**	0.742
	weighted	0.682	0.826	0.737	0.737

Table 4: Accuracy and area under ROC curve (AUC) of our method on other English and Greek datasets. The sets were compiled by ourselves for the purpose of enriching training domain for other variant of our classifier. The highest result in any category is bold.

The datasets mod-pan12-aa-EN and mod-Bpc-GR were compiled by ourselves from other authorship attribution sets for the purpose of enriching the training corpora for our final submission vF1. The comparison between results on the English and Greek subsets of vF1 with the results of vF2 (for which these additional sets were not used), shows that vF2 achieved better results on English data. while vF1 has higher AUC on Greek data.

Though these additional sets were not created specifically for authorship verification evaluation, we

examine the results of our methods on these sets (with the exception of vF1, which is tuned on them). We present the results in Table 4. vD1 performs poorly on mod-pan12-aa-EN. This is in part due to the fact that in this set the documents in a given problem instance can differ significantly with respect to the length, and the variant vD1 does not use the preprocessing of truncation all files withing a problem to the same length. The variants vD2 and vF2 (which apply this truncation) yielded accuracy and AUC similar in value to the ones achieved on the PAN 2013 English subset. The ensembles containing character n-gram classifiers yielded similar AUC on mod-pan12-aa-EN as on the PAN2013 English subset, close in value to 0.8. But their accuracies are distinctly lower than the results on the English competition subset, with values below 0.7 (for each such an ensemble, vast majority of the misclassified instances are false negatives: cases classified as not written by the same person when in fact they are). For mod-Bpc-GR the single classifiers (with parameters tuned on the competition Greek subset) perform rather poorly, with results similar but lower in values than the results yielded on the competition Greek test set. The ensembles containing word n-gram based classifiers perform better than the ensembles containing only the character n-gram classifiers, yielding both AUC and accuracy in the range of 0.71 – 0.75.

8 Future Work

It will be of interest to investigate the relation between the performance of our method and the number and the length of the considered texts. An interesting direction indicated by results of our experiments is also the analysis of the role of word n-grams and character n-grams for authorship verification depending on the genre of the texts, and on the topical similarity between the documents.

9 Conclusions

We present our proximity based one-class classification method for authorship verification. The method uses for each document of known authorship the most dissimilar document of the same author, and examines how much more or less similar is the questioned document. We use Common N-Gram dissimilarity based on differences in frequencies of character and word n-grams.

We evaluate our method on the set of PAN 2013 Authorship Identification competition. One variant of our method was submitted to the competition. The ordering by scores indicating the confidence that the documents were written by the same person, yielded by our method, and evaluated by area under ROC curve (AUC), is competitive with respect to other participants of the competition, overall, and on the English and Spanish subsets. On the entire set, AUC by each variant of our method is higher than the best result by other participants. In terms of accuracy, the method also performs better on the English and Spanish subsets of the dataset, and worse on the Greek one. An ensemble combining character based classifiers and word based classifiers yields the best accuracy, surpassing the best competition result on the entire set and on the English subset, while matching the best competition result on the Spanish subset.

As all proximity based one-class classification algorithms, our method relies on a selected threshold on the proximity between the questioned text and the set of documents of known authorship. Additionally, a single classifier requires two parameters defining the features representing documents. Ensembles of classifiers allow to alleviate the parameter tuning, by using many classifiers for many combinations of feature defining parameters, with a threshold fixed to 1 (a natural, albeit arbitrary, value).

Acknowledgements

This research was funded by a contract from the Boeing Company, Killam Predoctoral Scholarship, and a Collaborative Research and Development grant from the Natural Sciences and Engineering Research Council of Canada.

References

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

M.R. Ghaeini. 2013. Intrinsic Author Identification Using Modified Weighted KNN - Notebook for PAN at CLEF 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, September.

Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco M. Rangel Pardo, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Recent trends in digital text forensics and its evaluation - plagiarism detection, author identification, and author profiling. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *CLEF*, volume 8138 of *Lecture Notes in Computer Science*, pages 282–302. Springer.

Oren Halvani, Martin Steinebach, and Ralf Zimmermann. 2013. Authorship Verification via k-Nearest Neighbor Estimation - Notebook for PAN at CLEF 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, September.

Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios. 2013. Proximity Based One-class Classification with Common N-Gram Dissimilarity for Authorship Verification Task - Notebook for PAN at CLEF 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, September.

Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, September.

Patrick Juola. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Patrick Juola. 2012. An overview of the traditional authorship attribution subtask. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, pages 255–264, Dalhousie University, Halifax, Nova Scotia, Canada, August.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the 21st International Conference on Machine Learning*, ICML '04, page 489–495, Banf, Alberta, Canada, July. ACM.

Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, March.

Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Yaron Winter. 2012. The "Fundamental Problem" of Authorship Attribution. *English Studies*, 93(3):284–291.

PAN. 2012. Dataset of PAN 2012, Author Identification task. `http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/authorship.html`. Accessed on Apr 2, 2013.

PAN. 2013. Dataset of PAN 2013, Author Identification task. `http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-identification.html`. Accessed on Oct 8, 2013.

Shachar Seidman. 2013. Authorship Verification Using the Impostors Method - Notebook for PAN at CLEF 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, September.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495, December.

Efstathios Stamatatos. 2007. Author identification using imbalanced and limited training texts. In *Proceeding of the 18th International Workshop on Database and Expert Systems Applications, DEXA'07*, pages 237–241, Regensburg, Germany, September.

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

David Tax. 2001. *One Class Classification. Concept-learning in the absence of counter-examples.* Ph.D. thesis, Delft University of Technology, June.

Cor J. Veenman and Zhenshi Li. 2013. Authorship Verification with Compression Features. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, September.

Alexander Ypma, Er Ypma, and Robert P.W. Duin. 1998. Support objects for domain approximation. In *Proceedings of International Conference on Artificial Neural Networks*, pages 2–4, Skovde, Sweden, September. Springer.