Sub-Profiling by Linguistic Dimensions to Solve the Authorship Attribution Task

Notebook for PAN at CLEF 2012

Upendra Sapkota and Thamar Solorio

University of Alabama at Birmingham
Department of Computer and Information Sciences
upendra@uab.edu and solorio@cis.uab.edu

Abstract In this paper, we describe a modified version of the profile-based approach for the Authorship Attribution (AA) task of the PAN 2012 challenge. Our PAN system for AA utilizes the concept of linguistic modalities[1] on profile-based (PB) approaches. We concatenate all the training documents from the same author and build author-specific sub-profiles, one per linguistic modality. Then instead of using all the different types of features to compute the similarity of a test document against an author's profile in a single step, we compute several similarity scores using one set of features (modality) at a time. Each modality will assign the test document to the author whose sub-profile has the highest cosine similarity in that modality. Final classification decisions are based on the combination of decisions from each modality using majority voting. We achieved competitive results on PAN 2012, with encouraging results on the closed-class authorship attribution.

Keywords: authorship attribution, profile-based approach, linguistic modality

1 Introduction

Authorship attribution is the task of identifying the author of an unseen document. AA has a long history with multiple application areas that include spam filtering [20], cyber bullying, plagiarism detection [18], author recognition of a given program [5], and web information management. There are two flavors of AA tasks: closed-class and open-class. Most of the research has focused on the closed-class authorship identification task. In the closed class AA task, an anonymous test document should be assigned to one of a known group of candidate authors. In contrast, in the open class problem, an anonymous test document might belong to an author not included in the list of known candidates. The closed-class problem is considered to be relatively easier than open-class as the closed-class problem assumes that the unknown instance is drawn from the classes present in the training set. We proposed a framework for both the closed-class as well as the open class authorship attribution task of the PAN 2012 competition. The PAN 2012 competition dataset considers different types of documents with different lengths, including novel-length ones. The authorship task this year is more challenging than previous years because each sub-task has a very limited number of training

[1] Each linguistic modality refers to a type of feature.

instances per author, and the learning system should be able to effectively model each author's writing style.

In a recent work, Solorio *et al.* (2011) showed that representing documents as a set of separate linguistic modalities in a standard machine learning approach yields good results in AA. On the other hand, profile-based (PB) approaches have consistently shown promising results for the same task [3,8,10,14]. Our PAN system for AA combines these two ideas. We use the notion of linguistic modalities to generate sub-profiles of the authors, one per modality, and we use these sub-profiles to make predictions on authorship based on similarity scores. This independent processing of modalities follows the motivation in [17], where they argue that the independent processing of features by modalities allows to extract more meaningful similarity scores. However, in that previous work each document is represented as a unique instance of the problem, and the similarity scores are used as features to train a machine learning classifier. Here, we concatenate all the documents from a single author to generate the subprofiles. Each sub-profile then represents the author's writing style across a specific dimension, and each dimension will vote on a candidate author.

We participated in both the closed-class and the open-class problem and submitted one run per sub-task. The parameters of the final authorship attribution system, for both the closed-class and the open-class were adjusted using only the training data. Our system yielded very competitive results in the competition, reaching the best accuracies for several tasks. In what follows, we will discuss in more detail the process we followed during the parameter tuning of our system, as well as the evaluation results.

2 Related Work

Authorship attribution problems have been solved mainly using either machine learning or profile-based approaches [19]. Machine learning approaches, where each document in the training set is an instance of the problem, are based on a traditional text classification framework. Here, a machine learning algorithm, such as Support Vector Machines, will be trained using a set of feature vectors, each feature vector representing a sample document [6,2]. Profile-based approaches, on the other hand, solve the authorship attribution task by creating a profile of each author. For each author, all the training instances are concatenated into a single file and appropriate features are extracted to create author profiles. Our system borrows the idea of profiles because it has been successfully used for the authorship attribution task [7,11] and because we consider that having a very small training set, as is the case in this year's competition, will represent a challenge for traditional machine learning approaches. However, as mentioned earlier we generate a set of sub-profiles per author, instead of a single one.

Profile-based approaches have already been successfully used for attributing an unknown text to its real author. Successful examples of this approach include: [7,9,11,15,4]. Frantzeskou *et al.* (2007) showed the use of a profile-based approach on different datasets. They described how an effective and robust classifier can be built with the utilization of a modified similarity measure. They proposed the Source Code Author Profile (SCAP) method using byte-level n-grams to build the profiles, as presented in Keselj et. al.'s (2003) method. The SCAP study carries broad implications for the researchers in authorship attribution as it highlights the use of language-independent features and a simple similarity measure.

A lot of emphasis has also been placed on the use of machine learning for AA. Successful examples include a wide variety of learning algorithms, such as Support Vector Machines [2,17], decision trees [22], and memory based learners [12]. The work more closely related to our system is that of Solorio *et al.* (2011). They proposed a system that uses features from different linguistic dimensions (syntactic, lexical, stylistic, perplexity values) where each dimension is treated independently from each other. These dimensions are used to generate informative meta features, where they assume that meaningful similarity patterns among authors will emerge more clearly if the similarities are computed for each modality. After the similarities for all modalities are extracted, they are merged and used as features in a standard machine learning setting. As mentioned earlier we borrow the idea of modalities in our PAN system. The details are described in the following section.

3 Predicting Authorship Using Sub-profiles

The main difference between our approach and the standard PB approaches is that we use a set of sub-profiles to make the predictions, instead of computing a single one. The different sub-profiles correspond to the different linguistic dimensions in the writeprint of authors. We followed the notion of modalities proposed in previous work [17], although the exact features in each modality is not exactly the same as in previous work. Here we used stylistic, syntactic, semantic, character n-grams and word n-grams, resulting in five different modalities. The stylistic features include number of punctuations, number of sentences, number of tokens, and number of contractions, among other features. The syntactic features are the combination of all the part-of-speech (POS) tag unigrams, and the top n bigrams, trigrams, and grammatical relations from dependency parses. In the semantic modality we use the top n words after removing stop words. With the increased popularity and effectiveness of n-grams for authorship attribution tasks [8,3], we consider character level, as well as word level, n-grams as two different modalities. Character n-grams contain the top n trigrams. The last modality, word n-grams, contains the top 1500 occurrences of each n-gram, where n = 1, 2, 3. Table 1 provides all the features used in the different modalities. After the sub-profiles have been extracted we then compute the similarity scores between the test and candidate authors sub-profiles. Our approach uses the cosine distance [16] as our similarity metric since this has been successfully used in previous work. Note that this decision imposes a restriction on the set of features used in each profile. Typical PB approaches select the set of features in the profile on a per author basis. Our system uses a fixed set of features for all authors, and this set is determined by the training set. For modality m, the cosine similarity between two sub-profiles (the one from author a, and the one for test document t) represented by $\boldsymbol{P_a}$ and $\boldsymbol{P_t}$, each having l_m elements, is computed using Equation 1.

$$sim_m(\boldsymbol{P_a}, \boldsymbol{P_t}) = \frac{\boldsymbol{P_a} \cdot \boldsymbol{P_t}}{|\boldsymbol{P_a}||\boldsymbol{P_t}|} = \frac{\sum_{i=1}^{l_m} \boldsymbol{P_a}_i \times \boldsymbol{P_t}_i}{\sqrt{\sum_{i=1}^{l_m} \boldsymbol{P_a}_i^2} \times \sqrt{\sum_{i=1}^{l_m} \boldsymbol{P_t}_i^2}} \qquad (1)$$

Table 1. Features used in each modality

Modality	**Features Contained**
Stylistic	Total number of sentences
	Number of tokens per sentence.
	Percentage of words with out vowel.
	Number of punctuations per sentence.
	Percentage of contractions.
	Percentage of two consecutive punctuations.
	Number of total alphabetics.
	Percentage of sentence initial words with first letter capitalized.
	Total number of quotations.
Syntactic	Part-of-speech(POS) tag unigrams.
	Part-of-speech(POS) tag bigrams.
	Part-of-speech(POS) tag trigrams.
	Grammatical relations from the dependency parses.
Semantic	Bag of words.
Character n-gram	Character trigrams.
Word n-gram	Word unigrams.
	Word bigrams.
	Word trigrams.

We allow each modality to make authorship predictions based on the cosine similarity of the authors sub-profiles in that modality. Each modality will assign the test document to the author whose sub-profile has the highest cosine similarity in that modality. Final classification decisions are based on the combination of decisions from each modality using an appropriate voting mechanism [1]. Depending upon the type of AA problem, the weight of each modality in the final voting can be adjusted to improve the performance of the AA system.

Due to the limited number of training instances per author, here we use the majority voting to combine the decisions of each modality in the ensemble. Let x be a test instance. The final prediction p_t for test instance x is determined as:

$$p_t = \underset{y}{\operatorname{argmax}} \sum_{m=1}^{k} \mathit{\Gamma}(h_m(x), y), \quad y \in Y, \tag{2}$$

Where $h_m(x)$ is the prediction of modality m, k is the number of modalities, $\mathit{\Gamma}(i, j)$ is an indicator function whose value is 1 iff $i == j$ and 0 otherwise, and $Y = y_1, ..., y_l$ is the set of possible classes (authors). To increase the prediction performance as well as decrease the computation overhead, we decided to perform feature selection based on information gain. To determine the percentage of features to be selected from each modality, we used 50% of the training data as validation set and remaining 50% as the training. We experimented using different percentage of the features for each modality: 20%, 40%, 60%, and 80%. Looking at the performance of the system on the validation set, we determined the number of features to be used for the attribution of test instances. After analyzing the performance on validation set, we decided to use 80% of the stylistic features, 60% of the syntactic features, 20% of the semantic features, 20% of character n-grams, and 40% of word n-grams. To perform the feature selection based on information gain, we converted the feature vector file to arff format and used WEKA [21]. These parameters are fixed for all the datasets so that we can actually analyze how robust and consistent is our AA system on different datasets. This framework for the close-class AA system is evaluated in three datasets released by PAN 2012.

We also participated in the open class authorship attribution task of PAN 2012. The open class task is much harder than the closed class as test documents may belong to an unknown class. We did not do feature selection for this task. However, for the open class problem, we have an extra modality containing perplexity values from 4-gram language models at the character level. The extra modality was added after evaluating the performance of the open-class AA system on the validation set. When the improvement in the performance was observed with the addition of extra modality, we decided to include it for the open-class problem.

To deal with the “out of training set" cases, we decided to use a threshold on the difference between the cosine distance of the 1st and 2nd prediction of each modality. More specifically, to determine if a test document belongs to an unknown author, we compare the highest and the second highest cosine similarity score of the author sub-profiles for each modality. If the difference between them is smaller than the threshold value γ, we decide that the test instance belongs to none of the authors in that modality. We consider the γ threshold a sort of filter for confusing cases, and we assign those confusing cases to the “unknown" category. The idea behind this approach is that if there is indecision, as indicated by a close similarity score between different authors, this indecision is coming from trying to force an assignment of authorship to a document that has no clear match inside the training set, then we should not assign that author to any of our candidate authors, and the system should predict “unknown" in that case.

4 Experimental Results and Evaluation

All the parameters in our system were defined based on different experiments using only the training data. We used 50% of the training instances as validation data and the remaining 50% as the training data to set the parameters. In the cases of n-grams, we selected the top 3000 features, except for word n-grams because the performance on the validation set was better when the top 1500 of each word n-gram were used, where n = 1, 2, 3. We used the Stanford parser [13] to generate the parse trees. For both the closed-class and the open-class problem, we performed experiments separately for each modality on the provided datasets.

There are a total of six datasets, three from the closed-class problem and three from the open-class problem. For each dataset in the closed-class problem, there is a corresponding dataset in the open-class problem. Therefore, the training data is the same for both the closed-class and open-class problems while test data is different. There are three kinds of datasets for each problem: one with three authors, another with eight authors, and the last with 14 authors. As described in Section 3, we first create the author-specific sub-profiles, one per linguistic modality. Then the prediction of the test instances is carried out separately on a modality basis. This means each modality will have its own prediction on the same test instance. The predictions are combined using a majority vote.

Table 2. Accuracies of each modality, as well as the combination for the AA closed-class task of the PAN 2012 test data sets.

Dataset	**Modality Type**					**Majority Voting**
	Stylistic	**Syntactic**	**Semantic**	**Character n-gram**	**Word n-gram**	
3 Authors	83.33	100.00	83.33	100.00	83.33	100.00
8 Authors	75.00	87.50	75.00	62.50	50.00	100.00
14 Authors	28.57	78.57	71.42	64.28	92.85	92.85

Table 2 presents the performance of the individual modalities as well as the combination of the predictions from each modality on three different datasets provided in PAN 2012 authorship attribution competition. Our system achieves an overall accuracy of 100% on two out of the three datasets. The accuracy on the remaining dataset (14 authors) is also very high. As it can be seen, different modalities have different prediction accuracies on different datasets. For 3 Authors, syntactic as well as character n-grams obtain 100% accuracy, while all other modalities obtain 83.33%. Similarly, for 8 Authors, each modality performs differently. As expected, when we combine the individual predictions of the sub-profiles in an ensemble, the accuracy of the final combination is either higher or equal to that of the best individual classifier. The final accuracy for 8 Authors is 100% , which is more than the accuracy of each individual classifier. After looking at the accuracies of other PAN participants on each dataset, we observed that no participants obtained accuracy higher than ours. This proves that our AA system for the closed-class problem worked consistently well across different datasets and was among one of the the top AA closed-class systems in the competition.

In the open-class problem, the challenging part is to determine the threshold value. We performed experiments against different datasets using 0.02, 0.04, 0.06, and 0.08 as the threshold value (γ) and measuring accuracy on the validation set. Based on the performance of the AA system on validation set against the different values of γ we decided to set γ to 0.08. We even tried by increasing the value of γ beyond 0.08, but very poor performance on the validation set was obtained. As a reminder, we compare the highest and the second highest cosine similarity score of the author sub-profiles for each modality. The difference between them is compared against the γ threshold to determine if an instance belongs to an “unknown" category.

Table 3 presents the accuracy from using individual linguistic modalities, as well as the combination of the predictions from each modality for the open class problem of the PAN 2012 competition. Even though our results for the third dataset with 14 authors was not good, for the other two data sets our system yielded promising results. After combining the predictions from individual modalities, an accuracy of 60% was observed for the 3 Authors data set. This accuracy can be considered competitive when compared with that of other competitors on the same dataset. Our open-class AA system performed very well on the 8 Authors data set, reaching an accuracy of 76.47%, which is the highest accuracy obtained on that dataset. After analyzing the performance on

Table 3. Accuracies of each modality as well as the combination for the open-class problem of the PAN 2012 test data sets.

Dataset	**Modality Type**						**Majority**
	Stylistic	**Syntactic**	**Semantic**	**Character n-gram**	**Word n-gram**	**ppl**	**Voting**
3 Authors	60.00	70.00	40.00	50.00	50.00	50.00	60.00
8 Authors	29.41	58.82	52.94	58.82	47.05	52.94	76.47
14 Authors	6.25	31.25	56.25	31.25	68.75	12.50	37.50

the open-class datasets, we see that in only one case (row 2 of Table 3), the combined accuracy is higher than that of the individual classifiers. While for other two datasets, the combination of predictions from the sub-profiles is not the one with the best results. This could be due to having some very weak predictors as part of the ensemble.

For both closed-class and open-class problem, sub-profiling by linguistic modality achieved a very good performance overall in the PAN 2012 authorship attribution challenge. However, further experiments with additional datasets is required to evaluate the performance of the algorithm that we introduced in this paper. For the open-class problem, we proposed a new way of assigning a test instance to an “unknown" category. This new algorithm successfully worked for all the three datasets. However, additional empirical analyses are needed in order to explore different parameter values for this setting.

5 Conclusion

In this paper, we described an approach to the authorship attribution task adopted for the PAN 2012 competition. We introduced the notion of generating sub-profiles that represent the author writeprint along specific linguistic dimensions. Although the idea of linguistic modalities has been explored by previous work, it was used in a machine learning setting, while here we adopt that framework and combine it with a profile-based approach. The evaluation results on different datasets show this is a competitive AA framework for closed-class as well as open-class AA problems. For each dataset in the closed-class AA problem, our system matched the highest accuracy reported in the competition. This clearly illustrates that the proposed algorithm for closed-class problems is very competitive and consistent.

In the open-class AA framework our system did not reach the same high performance, but was also very competitive in the challenge. For one out of the three open-class datasets, we were able to get the best accuracy of the PAN 2012 challenge. The performance of this system for the dataset with 14 authors was not impressive. This leaves us with an impression that we need to keep investigating for a better way to handle open class settings.

Acknowledgements

This work was partially supported by ONR grant N00014-12-1-0217. The authors would like to thank the organizers of the PAN 2012 competition for putting together this event.

References

1. Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. pp. 1–15. Springer-Verlag (2000)
2. Escalante, H.J., Solorio, T., Montes, M.: Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. pp. 288–298. Association for Computational Linguistics (ACL) (2011)
3. Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C.E.: Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method. Journal of Digital Evidence 6(1) (2007)
4. Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C.E., Howald, B.S.: Identifying authorship by byte-level n-grams: The source code author profile (scap) method. IJDE (2007)
5. 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)
6. Houvardas, J., Stamatatos, E.: N-gram feature selection for author identification. In: Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications. LNCS, vol. 4183, pp. 77–86. Springer, Varna, Bulgaria (2006)
7. Joula, P.: Authorship attribution. Foundations and Trends in Information Retrieval 1(3), 233–334 (2006)
8. 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)
9. 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)
10. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language Resources and Evaluation 45, 83–94 (2011)
11. Lambers, M., Veenman, C.J.: Forensic authorship attribution using compression distances to prototypes. In: Geradts, Z.J.M.H., Franke, K.Y., Veenman, C.J. (eds.) IWCF 2009. vol. LNCS 5718, pp. 13–24 (2009)
12. Luyckx, K., Daelemans, W.: Shallow text analysis and machine learning for authorship attribution. Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands (CLIN) pp. 149–160 (2005)
13. Marneffe, M.D., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: LREC 2006 (2006)
14. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship: The Federalist. Addison-Wesley (1964), http://hdl.handle.net/2027/mdp.39015004063254
15. Peng, F., Shuurmans, D., Keselj, V., Wang, S.: Language independent authorship attribution using character level language models. In: Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics. vol. 1, pp. 267–274. Budapest, Hungary (2003)

16. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing Management 24(5), 513 – 523 (1988)
17. Solorio, T., Pillay, S., Raghavan, S., y Gómez, M.M.: Generating metafeatures for authorship attribution on web forum posts. In: 5th International Joint Conference on Natural Language Processing, IJCNLP 2011. pp. 156–164 (2011)
18. Stamatatos, E.: Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology (2011), http://dx.doi.org/10.1002/asi.21630
19. Stamatatos, E.: A survey on modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
20. 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)
21. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kauffmann, 2nd edn. (2005)
22. Zhao, Y., Zobel, J.: Effective and scalable authorship attribution using function words. In: Proceedings of 2nd Asian Information Retrieval Symposium. LNCS, vol. 3689, pp. 174–189. Jeju Island, Korea (2005)