An Open-Vocabulary Approach to Authorship Attribution

Notebook for PAN at CLEF 2019

Angelo Basile

Independent Researcher
me@angelobasile.it

Abstract The PAN 2019 Authorship Attribution shared task presents the challenge of the open-set condition, i.e. given a text and a set of possible authors, we have to predict who is the true author, but there is no guarantee that she is among the candidates. In this paper we present our participation to this shared task. Our best performing system consists of a linear model using sparse features: in this notebook we present this system in detail. We found that including rare words as features helps our model. Furthermore, we present a series of models which did not outperform the submitted system. On the official test set we achieved 0.613 open-set F1-score.

1 Introduction

Authorship analysis goes back at least to the 15th century, when the Italian humanist Lorenzo Valla showed by means of a linguistic analysis that the *Donation of Constantine* was a forgery. Today it has several applications: history [15], history of philosophy [3], intelligence [1]; [8,21] provide an exhaustive overview on the field and on the methods.

Authorship analysis covers different tasks: *a*) given a text, it is possible to predict the demographics of its authors (*author profiling*), *b*), verify if it was written by a specific author (*author verification*), *c*) compare its style to other texts (*plagiarism detection*) and *d*) if the author of said text is unknown, it can be possible to discover it (*authorship attribution*). We focus on this last task.

In this paper we present our participation to the Cross-domain Authorship Attribution shared task [9], part of the PAN Evaluation Forum [5]. This edition presents several challenges, as it includes texts written in four different languages (English, Italian, French and Spanish) and frames the task as an open-set problem, allowing for the possibility that the true author of a document is not among the candidates. Furthermore, the task organisers designed a cross-domain scenario by sampling the known and unknown texts for each problem from two different sources. All these challenges combined lead us to design a simple, lexical, profile-based model using sparse features.

We experimented with with no success with different stylometric features. In this paper we present our system together with an overview of a series of failed attempts to outperform it.

2 Data

PROBLEM	LANGUAGE	KNOWN TEXTS	VOCABULARY SIZE	UNKNOWN TEXTS
1	en	63	4152	561
2	en	63	4176	137
3	en	63	4049	211
4	en	63	4255	273
5	en	63	4070	264
6	fr	63	3963	121
7	fr	63	4056	92
8	fr	63	4100	430
9	fr	63	4082	239
10	fr	63	4086	38
11	it	63	3085	139
12	it	63	3745	116
13	it	63	3569	196
14	it	63	3676	46
15	it	63	3663	54
16	sp	63	3930	164
17	sp	63	4070	112
18	sp	63	3885	238
19	sp	63	3929	450
20	sp	63	3874	170

Table 1. An overview of the development data.

The data released by the organisers of the tasks consists of fanfiction literature, written in four languages (i.e. English, French, Italian and Spanish). Table 1 shows an overview of the dataset. Fanfiction is a literary genre consisting of writings inspired by certain well-known authors or works, known as fandom. This task challenges participants by asking to predict the author of a text belonging to a given fandom, given only developments texts belonging to different fandoms, providing thus a cross-domain (or cross-fandom) condition. For the development of our system we only used the data released by the organisers.

3 Submitted System

Our best system consists of a linear Support Vector Machine, fed with words and character penta-grams, including all the words, regardless of their frequency. We apply no

pre-processing to the data. Instances assigned a probability lower than 0.1 are classified as written by an unknown author. The features are normalised using TF-IDF. The value of the *C* hyper-parameter of the SVM is 1. For the implementation we used `scikit-learn` [16], starting from the baseline released by the task organisers. Although extremely simple, this system was not outperformed by more complex ones, as described in Section 4. We note that by allowing all the words occurring in the corpus to be part of the feature space, we outperform by 0.2 F1 points an identical system using only words occurring at least five times in the corpus.

PARAMETER	VALUE
n-grams	5
normalisation	tf-idf
lowercasing	false
min. freq	1
C	1.0

Table 2. Details of the submitted system.

4 Additional Experiments

This task presents several challenges, which all combined lead us to the choice of a simple system for the official submission. First, the evaluation platform limits the experimentation with large pre-trained neural models which are dependent on a GPU for running in a reasonable time: considering the reproducibility issues involved with neural models — as described in [19] — this is not necessarily a negative aspect for a shared task. Second, this being a multi-lingual task, we hypothesised that a system relying on linguistic knowledge would have been too dependent on the availability of specific resources (e.g. POS taggers, parsers, etc.). Third, considering that there is no overlap between the set of authors present in the development corpus and those present in the evaluation corpus, we could not rely on traditional techniques for fine-tuning the system.

Given all these constraints, we experimented with language-neutral methods, mostly leveraging frequency- and surface-based features.

Compression A first language-independent method is the compression-based method described in [22]; we used the implementation based on [17]. **Impostor** A second language independent method, also used in an implementation released by the organisers of this shared task, is the Impostors method [11]. **Ensemble** We experimented with a majority-voting ensemble system, built from combining the submitted system with the Compression and Impostors system. **Readability metrics** We tried following [12], by leveraging the readability of a text as a proxy for its true author: we used a battery of readability metrics [7,20,10,4,13,2], using the computed score as a feature in isolation and in combination with word n-grams: both approaches failed. **Bleaching** As shown in

[6], frequency and surface-level features can be useful for cross-lingual author profiling tasks, which are loosely related to authorship attribution. Table 3 shows an illustration of the bleaching feature abstraction method.

TOKEN	FEATURE	EXAMPLE
McDonald's	shape	CcCvcvcc'c
	alphanumeric	True
	length	08
	frequency	17

Table 3. An illustration of the feature bleaching process from [6].

We experimented with the bleaching approach of [6], resulting in a lower performance when compared to the submitted system.

5 Evaluation

The official evaluation is conducted on the TIRA Platform [18], using an F1 metric modified in order to account for the open-set scenario [14]. Here we present *i*) some development results obtained on the development data released by the organisers and *ij*) the official results obtained on the test set on the TIRA platform.

5.1 Development Results

model	open-f1 score
submitted	0.619
compressor	0.554
impostor	0.449
ensemble	0.618
readability	0.078
bleaching	0.133

Table 4. Overview of the cross-validated results from different models.

5.2 Official Results

Our submitted system scored 0.613 open-set macro F1, 0.8 points less than the best-performing system, which scored an average 0.69 open-set F1.

6 Conclusions

In this paper we presented our participation to the PAN Authorship Attribution shared task. We showed that linear models combined with sparse features work well for this task, at least under the constraints of *a)* limited computing power, *b)* language independence and *c)* out-of-domain data. We report that combining different systems into an ensemble model does not help improving performance. We show that word and character n-grams are good features for this task, even though they might allow for interference with topic effects. For reproducibility, we release all the code used in this paper. Our official submission scored 0.613 open-f1 score; we note that our system is the fastest one among the submitted runs on the official test set (00:17:08).

Acknowledgement

The author is thankful to the three anonymous reviewers who helped improving the quality of this paper.

References

1. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems 20(5), 67–75 (2005)
2. Anderson, J.: Lix and rix: Variations on a little-known readability index. Journal of Reading 26(6), 490–496 (1983)
3. Brandwood, L.: The chronology of Plato's dialogues. Cambridge University Press (1990)
4. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology 60(2), 283 (1975)
5. Daelemans, W., Kestemont, M., Manjavancas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., Zangerle, E.: Overview of PAN 2019: Author Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D., Heinatz, G., Cappellato, L., Ferro, N. (eds.) Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Springer (Sep 2019)
6. van der Goot, R., Ljubesic, N., Matroos, I., Nissim, M., Plank, B.: Bleaching text: Abstract features for cross-lingual gender prediction. In: ACL (2018)
7. Gunning, R.: The fog index after twenty years. Journal of Business Communication 6(2), 3–13 (1969)
8. Juola, P., et al.: Authorship attribution. Foundations and Trends® in Information Retrieval 1(3), 233–334 (2008)
9. Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., Stein, B.: Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
10. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)
11. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. JASIST 65, 178–187 (2014)

12. López-Anguita, R., Montejo-Ráez, A., Díaz-Galiano, M.: Complexity Measures and POS N-grams for Author Identification in Several Languages—Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France. CEUR-WS.org (Sep 2018)
13. McLaughlin, G.H.: Clearing the smog. J Reading (1969)
14. Mendes-Junior, P.R., de Souza, R.M., de Oliveira Werneck, R., Stein, B.V., Pazinato, D.V., de Almeida, W.R., Penatti, O.A.B., da Silva Torres, R., Rocha, A.: Nearest neighbors distance ratio open-set classifier. Machine Learning 106, 359–386 (2016)
15. Mosteller, F., Wallace, D.L.: Applied Bayesian and classical inference: the case of the Federalist papers. Springer Science & Business Media (2012)
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830 (2011)
17. Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J.M., Köhler, J., Lötzsch, W., Müller, F., Müller, M.E., et al.: Who wrote the web? revisiting influential author identification research applicable to information retrieval. In: European Conference on Information Retrieval. pp. 393–407. Springer (2016)
18. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)
19. Reimers, N., Gurevych, I.: Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. arXiv preprint arXiv:1803.09578 (2018)
20. Smith, E.A., Senter, R.: Automated readability index. AMRL-TR. Aerospace Medical Research Laboratories (US) pp. 1–14 (1967)
21. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538–556 (2009)
22. Teahan, W.J., Harper, D.J.: Using compression-based language models for text categorization. In: Language modeling for information retrieval, pp. 141–165. Springer (2003)