CIC-GIL Approach to Cross-domain Authorship Attribution

Notebook for PAN at CLEF 2018

Carolina Martín-del-Campo-Rodríguez[1], Helena Gómez-Adorno[1,2],
Grigori Sidorov[1], and Ildar Batyrshin[1]

[1] Instituto Politécnico Nacional (IPN),
Center for Computing Research (CIC), Mexico City, Mexico
cm.del.cr@gmail.com, sidorov@cic.ipn.mx, batyr1@gmail.com
[2] Universidad Nacional Autónoma de México (UNAM),
Engeneering Institute (II), Mexico City, Mexico
hgomeza@iingen.unam.mx

Abstract We present the CIC-GIL approach to the cross-domain authorship attribution task at PAN 2018. This year's evaluation lab focuses on the closed-set attribution task applied to a Fanfiction corpus in five languages: English, French, Italian, Polish, and Spanish. We followed a traditional machine learning approach and selected different feature sets depending on the language. We evaluated document features such as typed and untyped character n-grams, word n-grams, and function word n-grams. Our final system uses the log-entropy weighting scheme and SVM as classifier.

1 Introduction

The authorship attribution (AA) task consists in identifying the author of a given document among a list of candidates. There are several subtasks within the authorship attribution field such as author identification [4], author obfuscation [11] and author profiling [12]. The AA methods are used for many practical applications like electronic commerce, forensics, and humanities research [2,5]. The Authorship Attribution task is viewed as a multi-class, single-label classification problem, i.e. an automatic method has to assign a single class label (the author) to the unknown authorship documents.

Character n-grams are considered among the best feature representation for authorship attribution problems [16]. In [14], the authors introduced a categorization of character n-grams and showed that some categories have better performance than others in an AA task. Furthermore, several studies indicate that the combination of different types of n-grams introduces useful information to the classification algorithm, providing a robust model [13].

This paper describes our approach to the cross-domain authorship attribution task at PAN 2018 [4,17]. We examined different document features (typed and untyped character n-grams, word n-grams, and function word n-grams), weighting schemes (tf-idf and log-entropy), and machine learning algorithms (support vector machines, multinomial naive Bayes, and multi-layer perceptron).

2 Corpus for Development Phase

The corpus of the authorship attribution shared task at PAN 2018 is focused on cross-domain attribution. It is more challenging than the classical AA setting (the single-topic AA), because the training and testing documents can belong to different domains (eg. thematic area, genre). The documents in the corpus are **fanfics**, i.e., fictional literature based on the theme, atmosphere, style, characters, story world, etc. of a certain known author.

The corpus for development phase corpus (CDP), similarly to the corpus for test phase (CTP), is composed of a training corpus and a test corpus. Although the candidate authors for the CDP and CTP have similar characteristics, the candidate authors do not overlap.

The development phase corpus is composed of 10 problems divided in five languages (two problems each language): English, French, Italian, Polish and Spanish. The specifications of the problems are defined in [4].

3 Methodology

In this section, we first cover the concept of typed character n-grams, then the log-entropy weighting scheme, and finally the experimental settings of the methodology.

3.1 Typed character n-grams

Typed character n-grams, introduced by [14] are subgroups of character n-grams that correspond to three distinct linguistic aspects: morphosyntax (represented by affix n-grams), thematic content (represented by word n-grams) and style (represented by punctuation n-grams). These subgroups are call super categories (SC). Each of these SC are divided in different categories:

- **Affix n-grams:** Capture morphology to some extent (prefix, suffix, space-prefix, space-suffix).
- **Word n-grams:** Capture partial words and other word-relevant tokens (whole-word, mid-word, multi-word).
- **Punctuation n-grams:** Capture patterns of punctuation (beg-punct, mid-punct, end-punct).

Some categories of character n-grams showed higher predictive capabilities in the AA task [14] than using all possible n-grams (categorized and uncategorized). The re-definition stated by [7] of these categories unambiguously assign each 3-gram to exactly one category and do not exclude any n-gram (as in the case of consecutive punctuation marks in the original proposal). Also, the authors showed that some categories have a better performance that others for AA.

3.2 Log-entropy

Global weighting functions measure the importance of a term across the entire collection of documents [3]. Previous research on document similarity judgments [6,9] has shown that entropy-based global weighting is generally better than the TF-IDF model. The log-entropy (*le*) weight is calculated with the following equation (Equation 1):

$$le_{ij} = e_i \times \log(tf_{ij} + 1), \tag{1}$$

$$e_i = 1 + \sum_j \frac{p_{ij} \times \log p_{ij}}{\log n}, \ \ where \ \ p_{ij} = \frac{tf_{ij}}{gf_i}, \tag{2}$$

where n is the number of documents, tf_{ij} is the frequency of the term i in document j, and gf_i is the frequency of term i in the whole collection. A term that appears once in every document will have a weight of zero. A term that appears once in one document will have a weight of one. Any other combination of frequencies will assign a given term a weight between zero and one.

3.3 Experimental Settings

After an evaluation of several classification algorithms, in our final approach we chose Support Vector Machine (SVM) since this algorithm is recommended when the number of dimensions is greater than the number of samples (as in this case) [8]. We used the SVM implementation of sklearn [1], using the strategy one-against-all and the default parameter settings.

We analyzed several text representation schemes: typed character n-grams (with n varying from 2 to 8), untyped character n-grams (with n between 3 and 4), word n-grams (with n varying from 1 to 5) and function word n-grams proposed by Stamatatos [15].

We implemented the character n-gram types introduced by Sapkota *et al.* [14], but with the redefinitions of Markov *et al.* [7], which make them more accurate and complete.

For function word n-grams we used the 50 most frequent stop-words, as described in [15], to form the n-grams (with a value of n equal to 8). For English, the 50 most frequent stop-words mentioned in [15] were used. For the other languages (French, Italian, Polish and Spanish) the 50 most frequent stop-words were extracted from the development corpus (from the training).

We evaluated different combination of features for the different languages in the corpus. We also performed an evaluation study in order to identify the most useful typed character n-gram categories for each language. Table 1 shows the combination of features as well as the types of character n-grams used in our final submission.

Moreover, we experimented with different feature document frequency thresholds. We considered thresholds between 1 and 3, i.e. features that occur in at least 1, 2, or 3 documents in each problem. We found that the features that occur in at least 2 documents achieved the best classification performance in our experiments.

Table 1. Features included for each language in our final submission.

Language	Features	Typed character n-grams categories
English	typed character n-grams (2, 3, 5)	whole-word, mid-word, multi-word, beg-punct, mid-punct, end-punct
French	typed character n-grams (2, 4, 5)	prefix, mid-word, multi-word, beg-punct, end-punct
Italian	word n-grams (1, 2, 3, 5)	
Polish	word n-grams (2, 5)	
Spanish	character n-grams(3), typed character n-grams(4) and word n-grams(1, 2)	beg-punct

Following the experimental settings presented in [3], we examined two feature representations based on a global weighting scheme: log-entropy and tf-idf. Global weighting functions measure the importance of a word across the entire collection of documents. Previous research on document similarity judgments [6,9] and authorship attribution [3] has shown that entropy-based global weighting is generally better than the if-idf model. We use log-entropy as weighting function for out final version.

4 Evaluation Measure and Results

The macro-averaged F1 score is used for evaluating the performance of the systems participating in the authorship attribution shared task at PAN CLEF 2018 [4].

The final configuration of our approach was selected based on the classification performance on the test set of the development phase corpus (DPC). Table 2 shows the results obtained on the DPC with the above-specified configuration evaluated on the TIRA platform [10].

Table 2. Results of the Cross Domain Authorship-Attribution on the Development Phase Corpus

Language	Problem	Macro-Average F1
English	problem 1	0.582
	problem 2	0.783
French	problem 3	0.659
	problem 4	0.938
Italian	problem 5	0.702
	problem 6	0.637
Polish	problem 7	0.589
	problem 8	0.893
Spanish	problem 9	0.804
	problem 10	0.879
Overall score		**0.747**

The results achieved in the test phase corpus (TPC) are shown in Table 3. It can be observed that the performance on the TPC is much lower than in the DPC. This behavior can be explained by our decision of tuning our system based on the classification performance over the test set of the DPC.

Table 3. Results of the Cross Domain Authorship-Attribution Task on the Test Phase Corpus

User	Macro-Average F1	Runtime
custodio18	0.685	00:04:27
murauer18	0.643	00:19:15
halvani18	0.629	00:42:50
mosavat18	0.613	00:03:34
yigal18	0.598	00:24:09
delcamporodriguez18	**0.588**	**00:11:01**
pan18-baseline	0.584	00:01:18
miller18	0.582	00:30:58
schaetti18	0.387	01:17:57
gagala18	0.267	01:37:56
garciacumbreras18	0.139	00:38:46
tabealhoje18	0.028	02:19:14

5 Conclusions

We presented the system that was submitted to the Cross-domain Authorship Attribution task at PAN 2018. Our experiments were performed using different features, finding that a specific set of features per language is the best approach to improve performance.

Our approach had a good performance on the development phase corpus (Macro-Average F1: 0.747), but this performance was severely diminished on the test phase corpus (Macro-Average F1: 0.588). Based on the current technique, there are still opportunities for further enhancements.

In future research, we would like to consider a cross-validation approach for the development phase corpus to make the system more robust.

Acknowledgments

This work was partially supported by the Mexican Government (CONACYT projects 240844, SNI, COFAA-IPN, SIP-IPN 20181849, 20171813) and Honeywell Grant.

References

1. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122 (2013)

2. Coulthard, M.: On admissible linguistic evidence. Journal of Law & Policy 21, 441 (2012)
3. Gómez-Adorno, H., Aleman, Y., Vilariño, D., Sanchez-Perez, M.A., Pinto, D., Sidorov, G.: Author clustering using hierarchical clustering analysis. In: CLEF 2017 Working Notes. CEUR Workshop Proceedings (2017)
4. Kestemont, M., Tschugnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
5. Koppel, M., Seidman, S.: Automatically identifying pseudepigraphic texts. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1449–1454. EMNLP '13 (2013)
6. Lee, M.D., Navarro, D.J., Nikkerud, H.: An empirical evaluation of models of text document similarity. In: Proceedings of the Cognitive Science Society. vol. 27 (2005)
7. Markov, I., Stamatatos, E., Sidorov, G.: Improving cross-topic authorship attribution: The role of pre-processing. In: Proceedings of the 18^{th} International Conference on Computational Linguistics and Intelligent Text Processing. CICLing 2017, Springer (2017)
8. 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)
9. Pincombe, B.: Comparison of human and latent semantic analysis (lsa) judgements of pairwise document similarities for a news corpus. Tech. rep., DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION SALISBURY (AUSTRALIA) INFO SCIENCES LAB (2004)
10. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
11. Potthast, M., Hagen, M., Schremmer, F., Stein, B.: Overview of the Author Obfuscation Task at PAN 2018: A New Approach to Measuring Safety. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
12. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
13. Sanchez-Perez, M.A., Markov, I., Gómez-Adorno, H., Sidorov, G.: Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 145–151. Springer (2017)
14. Sapkota, U., Bethard, S., Montes-y Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies. pp. 93–102. NAACL-HLT'15, Association for Computational Linguistics (2015)
15. Stamatatos, E.: Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology 62(12), 2512–2527 (2011)

16. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
17. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., Potthast, M.: Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (Sep 2018)