EACH-USP Ensemble Cross-Domain Authorship Attribution

Notebook for PAN at CLEF 2018

José Eleandro Custódio and Ivandré Paraboni

School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)
São Paulo, Brazil
{eleandro,ivandre}@usp.br

Abstract. We present an ensemble approach to cross-domain authorship attribution that combines predictions made by three independent classifiers, namely, standard char n-grams, char n-grams with non-diacritic distortion and word n-grams. Our proposal relies on variable-length n-gram models and multinomial logistic regression, and selects the prediction of highest probability among the three models as the output for the task. Results generally outperform the PAN-CLEF 2018 baseline system that makes use of fixed-length char n-grams and linear SVM classification.

1 Introduction

Authorship attribution (AA) is the computational task of determining the author of a given document from a number of possible candidates [1]. Systems of this kind have a wide range of possible applications, from on-line fraud detection to plagiarism and/or copyright protection. AA is presently a well-established research field, and a recurrent topic in the PAN-CLEF shared task series [7,5].

At PAN-CLEF 2018, a cross-domain authorship attribution task applied to fan fiction text has been proposed. In this task, texts written by the same authors in multiple domains were put together, creating a cross-domain setting. The task consists of identifying the author of a given document based on text of a different genre.

The present work describes the results of our own entry in the PAN-CLEF 2018 [2] AA shared task - hereby called the EACH-USP model - using both the baseline system and data provided by the event [1]. This consists of ten individual AA tasks in five languages (English, French, Italian, Polish and Spanish), being two tasks (with 5 or 20 candidate authors) each.

The rest of this paper is structured as follows. Section 3 describes our main AA approach, and Section 4 describes its evaluation over the PAN-CLEF 2018 AA dataset. Section 5 presents our results and those provided by relevant baseline methods. Finally, Section 6 discusses these results and suggests future work.

[1] Available from https://pan.webis.de/clef18/pan18-web/author-identification.html

2 Related Work

The present work shares similarities with a number of AA studies. Some of these are briefly discussed below.

The work in [9] makes use of text distortion methods intended to preserve only the text structure and style in a cross-domain AA setting. The work focused on the use of word-level information, whereas our current proposal will focus on character-level information.

The work in [8] investigates the role of affixes in the AA task by using char n-gram models for the English language. Similarly, the work in [3] addresses the use of char n-grams models for the Portuguese language, and discusses the role of affix information in the AA task. This is in principle relevant to our current work since the Portuguese language shares a great deal of its structure with Spanish and Italian, which are two of the target languages for the PAN-CLEF 2018 AA task.

3 Method

Central to our approach is the idea that the AA task may rely on the combination of different knowledge sources such as lexical preferences, morphological inflection, uppercase usage, and text structure, and that different kinds of knowledge may be obtained either from character-based or word-based text models. These alternatives are discussed as follows.

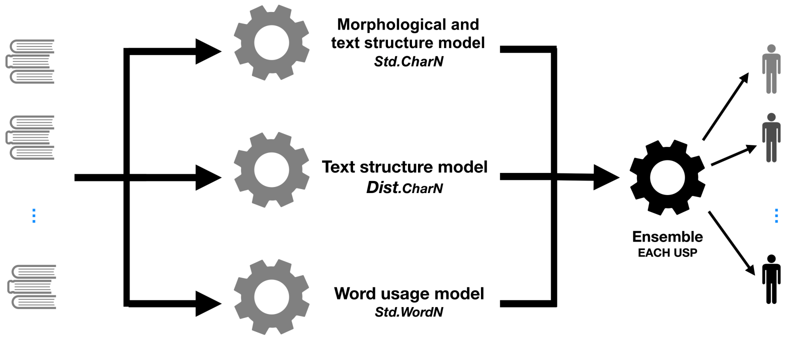
Word or content-based models may indicate word usage preferences, and may help distinguish an author from another. However, we notice that a single author may favour certain words in different domains (e.g., fictional versus dialogue text). Moreover, word-based models will usually discard punctuation and spaces, which may represent a valuable knowledge source for AA. Character-based models, on the other hand, are known for their ability to capture time or gender inflection, among others, as well as punctuation and spacing [6].

Based on these observations, our approach to cross-domain authorship attribution consists of a number of improvements over the standard PAN-CLEF 2018 baseline system organised as an ensemble method. In particular, we replace the original fixed-length n-grams and linear SVM classification for variable-length n-grams and multinomial logistic regression, and we combine predictions made by three independent classifiers to determine the most likely author of a given document as illustrated in Figure 1.

Our proposal - hereby called USP-EACH *Ensemble* model - combines the following three classifiers:

- *Std.charN*: a variable-length char-ngram model
- *Dist.charN*: a variable-length char-ngram model in which non-diacritics were distorted
- *Std.wordN*: a variable-length word-ngram model

Fig. 1. Ensemble cross-domain AA architecture



Both *Std.charN* and *Dist.charN* models are intended to capture syntactic and morphological clues for authorship attribution in a language-independent fashion. In the latter, however, all characters that do not represent diacritics are removed from the text beforehand, therefore focusing on the effects of punctuation, spacing and the use of diacritics, numbers and other non-alphabetical symbols. This form of text distortion [9] is illustrated by the example in Table 1.

Table 1. Example of text distortion using the first document of the 9th training dataset.

Original text	Transformed text
−¿Y cómo sabes que no lo ama? −Inglaterra se preguntó a su vez si habría un muñeco del esposo también.	−¿* *ó** ***** *** ** ** ***? −********** ** *******ó * ** *** ** ****í* ** **ñ*** *** ****** *****é*.

A major motivation for this approach is the observation that, in languages that make use of diacritics, some authors may consistently use the correct spelling (as in 'é', which is Portuguese for 'is') whereas others tend to ignore the need for diacritics by producing the incorrect spelling (e.g., 'e') for the same purpose. In addition to these two character-based models, we also consider a third model that is intended to capture lexical preferences, hereby called *Std.wordN*.

Predictions made by the three classifiers are combined into our *Ensemble* model by selecting the most likely outcome for a given authorship attribution task. To this end, the three individual outputs are concatenated and taken as input features to a fourth soft voting (ensemble) classifier. This in turn performs multinomial logistic regression to select the winning strategy.

4 Experiment

The models introduced in the previous section had their parameters set by using the PAN-CLEF development dataset as follows. Features were scaled using Python's MaxAbsScaler transformer [4], and dimensionality was reduced using a standard PCA implementation. PCA also helps remove correlated features, which is useful in the present case since our models make use of variable length feature concatenation.

The resulting feature sets were submitted to multinomial logistic regression by considering a range of values, as summarised in Table 2.

Table 2. Pipeline - Multinomial logistic regression parameters

Module	Parameters	Possible values
Feature Extraction	N-gram range	Start=(1 to 3) - End=(1 to 5)
	Min document frequency	[0.01, 0.05, 0.1, 0.5]
	Max document frequency	[0.25, 0.50, 0.90, 1.0]
	TF	normal, sublinear
	IDF	normal, smoothed
	Document normalisation	L1, L2
Transformation	Scaling	MaxAbsScaler
	PCA percentage of explained variance	[0.10, 0.25, 0.50, 0.75, 0.90, 0.99]
Classifier	Logistic regression	Multinomial-Softmax

Optimal values for the regression task were determined by making use of grid search and 5-fold cross validation using an ensemble method. The optimal values that were selected for subsequently training our actual models are illustrated in Table 3, in which Start/End values denote the range of subsequences that were concatenated. For instance, Start= 2 and End= 5 represents the concatenation of subsequences $[(2,2),(2,3),\cdots,$ $(4,3),(4,5)]$.

Table 3. Pipeline - Multinomial logistic regression optimal values

Module	Parameters	Optimal values
Feature Extraction	N-gram range	Std.charN - Start=2 End=5
		Dist.charN - Start=2 End=5
		Word.charN - Start=1 End=3
	Min corpus frequency	0.05
	Max corpus frequency	1.0
	TF	sublinear
	IDF	smoothed
	Document normalisation	L2
Transformation	PCA	0.99

Tables 4, 5 and 6 show the ten most relevant features for AA Problem00002, which comprises text written by five authors each in the English language. In this representation, blank spaces were encoded as underscore symbols, and relevance is represented by the weights of multinomial logistic regression. These were estimated by scaling the features to a mean value equal to 0, and to a standard deviation value equal to 1.

Table 4. Most relevant features for *Std.charN*

candidate00001	candidate00002	candidate00003	candidate00004	candidate00005
_as_l	_Sti	_sub	_joi	_day,
_’	_"Can	_suc	_gh	_dev
_prec	_"Ca	_I_fi	_er	_dete
_I’d	_"Be	_succ	_glow	_plac
_"Are	_but	_subs	_Is	_mut
_Re	_but_	_I_f	_sta	_must
_smel	_Ofte	_"T	_gor	_Dro
_leak	_posi	_a_t	_sorr	_day_
_is_s	_For	_"St	_eat_	_she_
_spu	_Ri	_a_sw	_If_t	_chi

Table 5. Most relevant features for *Dist.charN*

candidate00001	candidate00002	candidate00003	candidate00004	candidate00005
*_‘**	_**_-	"*’	*_~_	*_–_*
_**-	_**_(	"*_**	*_~	’*,_*
’	_**_*	!),_*	*_._*	"_~
*).	*!	*!!	’***	*_–
*),_	_**_’	*’*_*	’****	*_–_
_-_*	*!_*	**_*’	"_**’	’*.
_-_	*_“**	**_**	_É***	_“*
_’**	_~_	**_*’	_“*’	_-
!),	_~_*	_**!	_**..	_-_
_’***	_“*’	*!_*_	*_**é	_"***

Being a language-independent approach, information regarding function words was not taken into account, although this might have been helpful since function words usually play a rather prominent role in AA (as opposed to, e.g., content words, which may arguably be more relevant to other text categorisation tasks.) We notice however that function words were made explicit by the *Std.wordN* model. Moreover, we notice that all models also made (to some extent) explicit a number of individual preferences regarding word usage, punctuation and spacing, and that *Std.distN* provides some evidence of the role of punctuation marks, spacing and hyphenation.

Table 6. Most relevant features for *Std.wordN*

candidate00001	candidate00002	candidate00003	candidate00004	candidate00005
about_what	against_his	an_odd	although	and_pulled_him
and_practically	and_it_was	and_then_he	an_eye	and_pulling
any_of	and_so	acknowledged	and_said	across_his
any_more	and_already	and_he_had	and_takes	across_the
and_nearly	and_steve	are_your	and_just	and_all
and_pulled	and_say	again_to	ancient	against_her
agree	accent	and_tell	amount_of	among
all_tony	and_wet	and_forth	always	about_what_to
ah	apparently	are_just	and_grinned	acting
and_wet_and	after	and_grabbing	about_the	about_their

5 Results

Table 7 presents macro F-measure results for the original PAN-CLEF 2018 baseline system, our three individual classifiers and the *Ensemble* model for the ten PAN-CLEF 2018 authorship attribution tasks over the development data. To this end, the baseline was optimised using 4-grams, minimum document frequency of 5 and One-vs-Rest as the SVM multi-class strategy. Our models were optimised individually using the parameters described in Table 2, and output probabilities were combined by using multinomial logistic regression in a soft voting ensemble fashion. Best results are highlighted.

Table 7. Macro-F1 measure results for PAN-CLEF 2018 AA development corpus

Problem	Language	Authors	Baseline	Std.charN	Dist.charN	Std.wordN	Ensemble
001	English	20	0.514	0.609	0.479	0.444	**0.625**
002	English	5	0.626	0.535	0.333	0.577	**0.673**
003	French	20	0.631	0.681	0.568	0.418	**0.776**
004	French	5	0.747	0.719	0.586	0.572	**0.820**
005	Italian	20	0.529	0.597	0.491	0.497	**0.578**
006	Italian	5	0.614	0.623	0.595	0.520	**0.663**
007	Polish	20	0.455	0.470	0.496	0.475	**0.554**
008	Polish	5	0.703	**0.948**	0.570	0.922	0.922
009	Spanish	20	0.709	**0.774**	0.589	0.616	0.701
010	Spanish	5	0.593	0.778	0.802	0.588	**0.830**
Mean			0.612	0.673	0.551	0.563	**0.714**

From these results, a number of observations are warranted. First, we notice that *Std.charN* generally obtained the best results among the three individual classifiers. We also notice that *Dist.charN* performs worse than *Std.charN*. This was to be expected since *Dist.charN* conveys less information, that is, it may be seen as a subset of *Std.charN*.

Our ensemble model consistently outperformed the alternatives by using soft voting. In our experiments, we noticed that combining the three knowledge sources obtained best results. In all cases, the relevant features turned out to be of variable length, ranging from 1 to 5-grams.

6 Final remarks

This paper presented an ensemble approach to cross-domain authorship attribution that combines predictions made by a standard char n-gram model, a char n-gram model with non-diacritic distortion and a word n-gram model using variable-length n-gram models and multinomial logistic regression. Results generally outperform the PAN-CLEF 2018 baseline system that makes use of fixed-length char n-grams and linear SVM classification. As future work, we intend to investigate alternative text models and distortion methods for prefixes, suffixes and other text components.

Acknowledgements. The second author received financial support from FAPESP grant nro. 2016/14223-0.

References

1. Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., benno Stein: Recent trends in digital text forensics and its evaluation: Plagiarism detection, author identification, and author profiling. In: LNCS 8138. pp. 282–302 (2013)
2. 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)
3. Markov, I., Baptista, J., Pichardo-Lagunas, O.: Authorship attribution in portuguese using character N-grams. Acta Polytechnica Hungarica 14(3), 59–78 (2017)
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. Journal of machine learning research 12, 2825–2830 (2011)
5. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN 17: Author identification, author profiling, and author obfuscation. In: LNCS 10456. pp. 275–290 (2017)
6. Rocha, A., Scheirer, W.J., Forstall, C.W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A.R.B., Stamatatos, E.: Authorship Attribution for Social Media Forensics. IEEE Transactions on Information Forensics and Security 12(1), 5–33 (2017)
7. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN 16: New challenges for authorship analysis: Cross-genre profiling, clustering, diarization, and obfuscation. In: LNCS 9822. pp. 332–350 (2016)
8. 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 NAACL HLT 2015. pp. 93–102 (2015)
9. Stamatatos, E.: Authorship attribution using text distortion. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL-2017). Association for Computational Linguistics, Valencia, Spain (2017)