EACH-USP Ensemble Cross-domain Authorship Attribution for PAN-CLEF-2018

J. Eleandro Custódio, Ivandré Paraboni
{eleandro,ivandre}@usp.br

Avignon, 11 September 2018

School of Arts, Sciences and Humanities
University of São Paulo
São Paulo Brazil

Overview

- **Context**
- **Motivation**
- **Method**
- **Parameter optimisation**
- **Results**
- **Discussion**

Context

- Early stages of our own Authorship Attribution (AA) research
- Focus on understanding the problem
  (as opposed to Author Profiling)
- Long-term goals:
  - Language- and Content- independent AA
  - Issues for AA in the Brazilian Portuguese language

Motivation

- AA problems come in different flavours

  Contents x Structure

- Bag-of-word methods perform fairly well

  ...but structure also plays a major role in AA

- Ideally we should make use of every possible knowledge source
- Proposal: a simple ensemble method combining well-known AA approaches

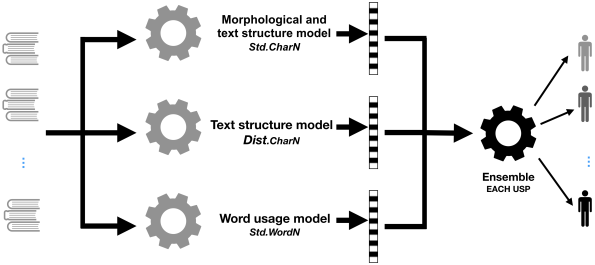
Method

- Possible improvements over the standard PAN-2018 baseline:
  - SVM replaced by multinomial logistic regression
  - Fixed n-gram models replaced by variable length n-grams
  - Ensemble of three classifiers:

    Std.charN a variable-length char-ngram model
    Dist.charN a variable-length char-ngram model in which non-diacritics were distorted
    (Stamatatos, 2017; Granados et. al., 2012)
    Std.wordN a variable-length word-ngram model
- Classifier outputs are combined by soft voting

Method: Architecture

Figure 1: Ensemble architecture



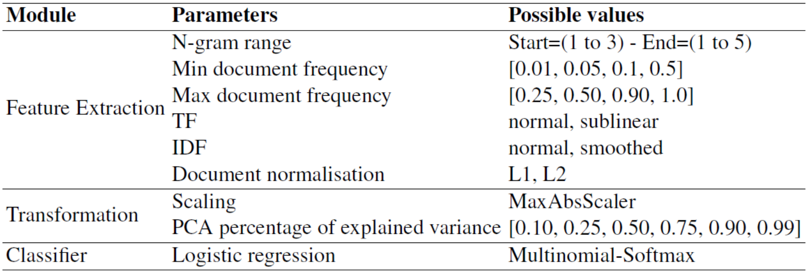
Text Distortion Example

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

First document from Problem 00009 in PAN-CLEF training data.

Parameter optimisation

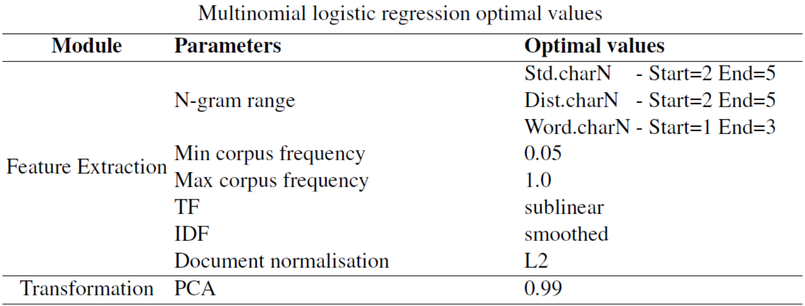
Multinomial logistic regression parameters



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 for each language were determined by making use of grid search and 5-fold cross validation using an ensemble method.
- A single set of values was chosen for all languages.
- Dimensionality was reduced using standard PCA

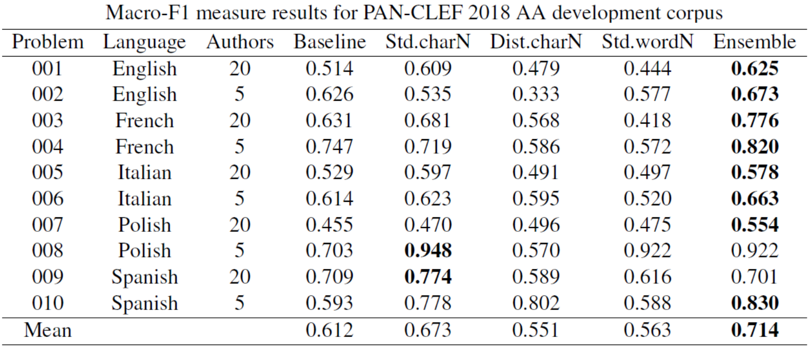
Optimal values



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

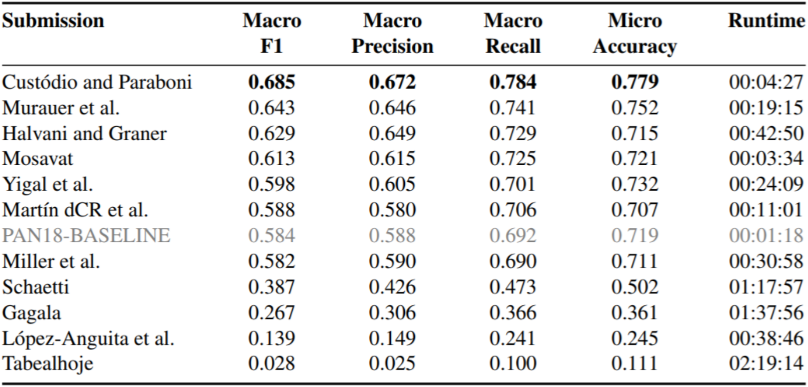
Development results



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

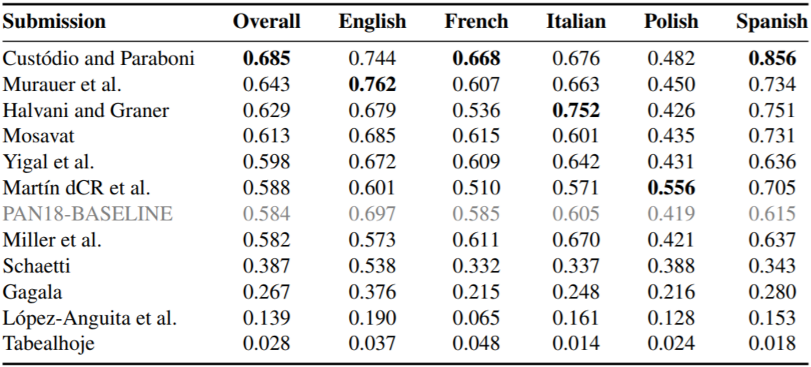
Problem	Language	Authors	Baseline	Std.charN	Dist.charN	Std.wordN	Ensemble
001	English	20	0.514	0.609	0.479	0.444	**0.625**
002	English	5	0.626	0.535	0.333	0.577	**0.673**
003	French	20	0.631	0.681	0.568	0.418	**0.776**
004	French	5	0.747	0.719	0.586	0.572	**0.820**
005	Italian	20	0.529	0.597	0.491	0.497	**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**

PAN-2018 Overall results



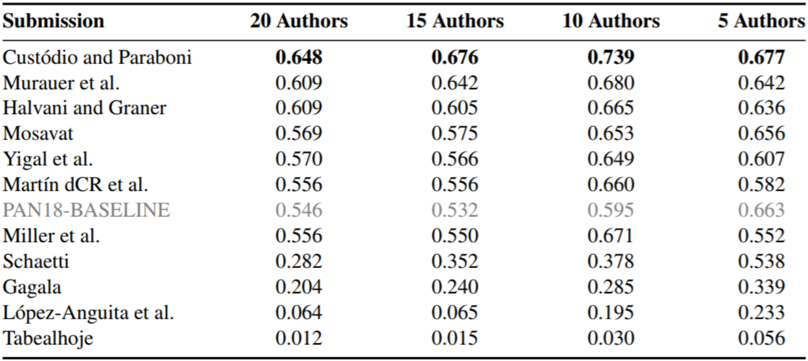
Submission	Macro F1	Macro Precision	Macro Recall	Micro Accuracy	Runtime
Custódio and Paraboni	**0.685**	**0.672**	**0.784**	**0.779**	00:04:27
Murauer et al.	0.643	0.646	0.741	0.752	00:19:15
Halvani and Graner	0.629	0.649	0.729	0.715	00:42:50
Mosavat	0.613	0.615	0.725	0.721	00:03:34
Yigal et al.	0.598	0.605	0.701	0.732	00:24:09
Martín dCR et al.	0.588	0.580	0.706	0.707	00:11:01
PAN18-BASELINE	0.584	0.588	0.692	0.719	00:01:18
Miller et al.	0.582	0.590	0.690	0.711	00:30:58
Schaetti	0.387	0.426	0.473	0.502	01:17:57
Gagala	0.267	0.306	0.366	0.361	01:37:56
López-Anguita et al.	0.139	0.149	0.241	0.245	00:38:46
Tabealhoje	0.028	0.025	0.100	0.111	02:19:14

PAN-2018 Per language results



Submission	Overall	English	French	Italian	Polish	Spanish
Custódio and Paraboni	**0.685**	0.744	**0.668**	0.676	0.482	**0.856**
Murauer et al.	0.643	**0.762**	0.607	0.663	0.450	0.734
Halvani and Graner	0.629	0.679	0.536	**0.752**	0.426	0.751
Mosavat	0.613	0.685	0.615	0.601	0.435	0.731
Yigal et al.	0.598	0.672	0.609	0.642	0.431	0.636
Martín dCR et al.	0.588	0.601	0.510	0.571	**0.556**	0.705
PAN18-BASELINE	0.584	0.697	0.585	0.605	0.419	0.615
Miller et al.	0.582	0.573	0.611	0.670	0.421	0.637
Schaetti	0.387	0.538	0.332	0.337	0.388	0.343
Gagala	0.267	0.376	0.215	0.248	0.216	0.280
López-Anguita et al.	0.139	0.190	0.065	0.161	0.128	0.153
Tabealhoje	0.028	0.037	0.048	0.014	0.024	0.018

PAN-2018 Per dataset size results



Submission	20 Authors	15 Authors	10 Authors	5 Authors
Custódio and Paraboni	**0.648**	**0.676**	**0.739**	**0.677**
Murauer et al.	0.609	0.642	0.680	0.642
Halvani and Graner	0.609	0.605	0.665	0.636
Mosavat	0.569	0.575	0.653	0.656
Yigal et al.	0.570	0.566	0.649	0.607
Martín dCR et al.	0.556	0.556	0.660	0.582
PAN18-BASELINE	0.546	0.532	0.595	0.663
Miller et al.	0.556	0.550	0.671	0.552
Schaetti	0.282	0.352	0.378	0.538
Gagala	0.204	0.240	0.285	0.339
López-Anguita et al.	0.064	0.065	0.195	0.233
Tabealhoje	0.012	0.015	0.030	0.056

Final remarks

- Ensemble generally outperforms individual classifiers
- Best results were obtained for the Spanish language
- Many other opportunities for text distortion
- Future work will combine the use of embedding models for each author

Thank you

This work has been supported by FAPESP grant 2016/14223-0

Special thanks to the PAN-CLEF 2018 organisers!!!

Contact: {eleandro,ivandre}@usp.br