Multi-channel Open-set Cross-domain Authorship Attribution

J. E. Custódio, I. Paraboni

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

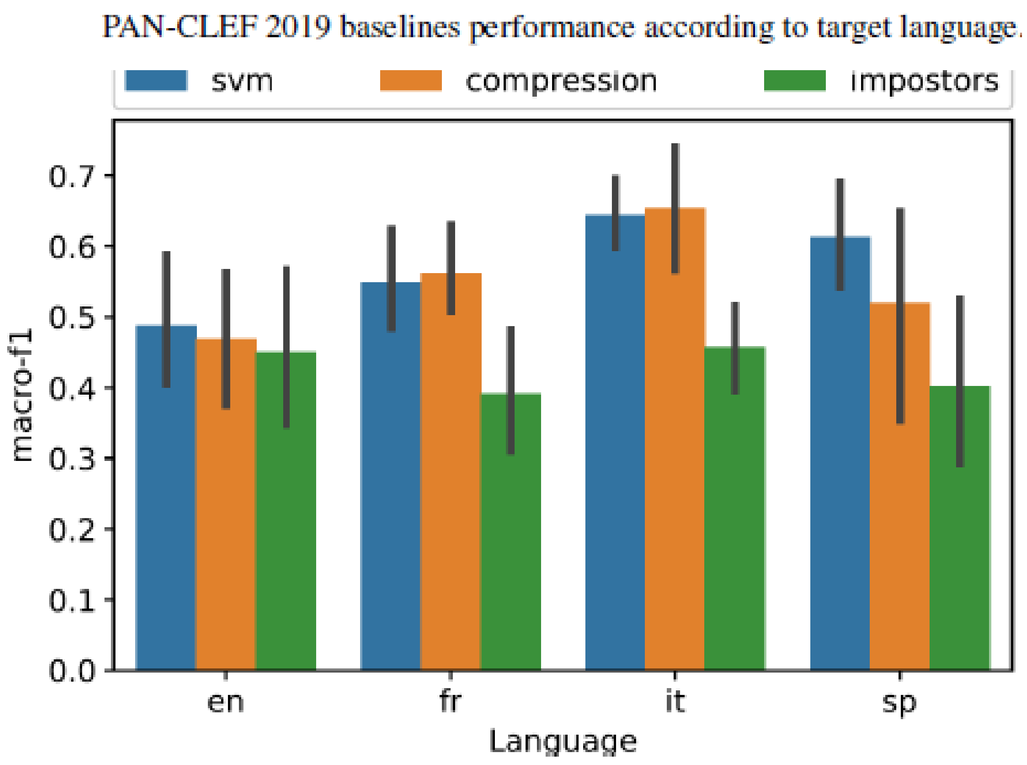
{eleandro,ivandre}@usp.br

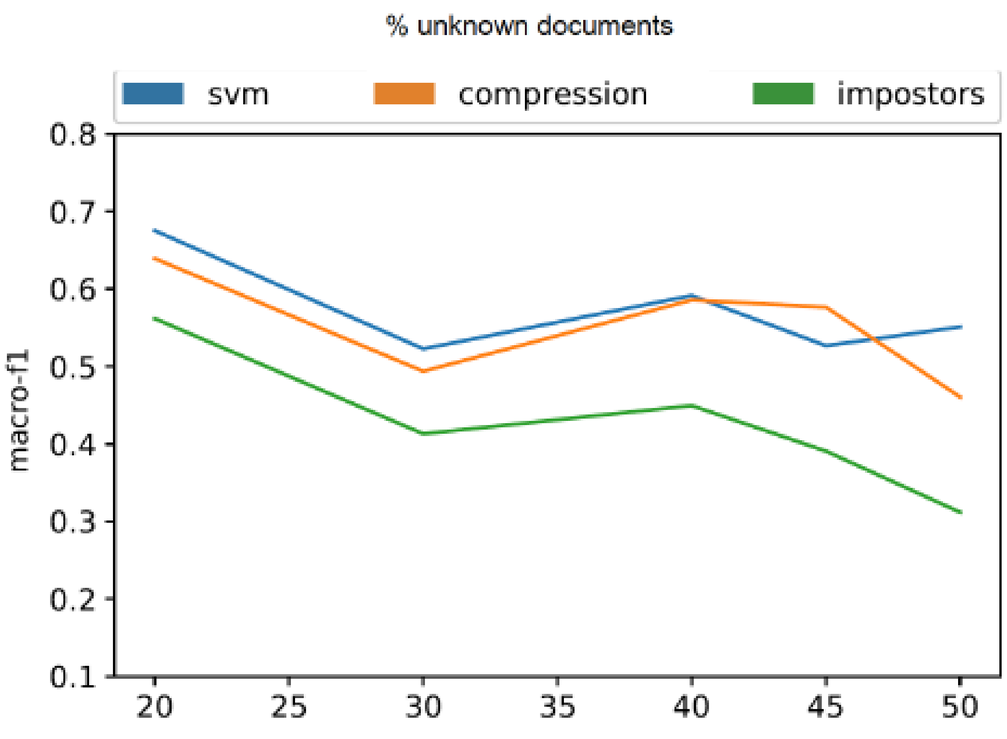
Introduction

- Authorship attribution (AA) is the computational task of identifying the author of a
- given text by examining samples of texts written by a number of candidate authors (Joula, 2012)
- At PAN-CLEF 2019 () an open-set AA task was proposed.
- Current participation is an extension of previous year's EACH-USP ensemble approach to closed-set AA (Custódio & Paraboni, 2018)

Data analysis

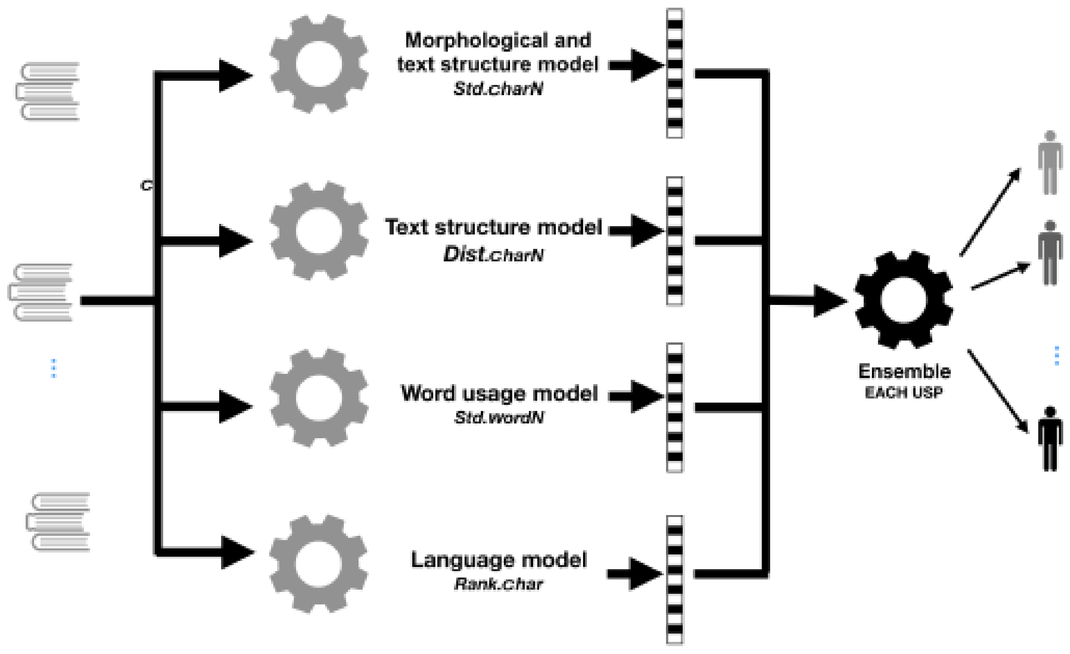
- PAN-CLEF 2019 AA development dataset conveys 20 problems in four languages (English, French, Italian and Spanish), with nine candidate authors per problem, seven documents per candidate and an average of 4500 characters per document
- First step consisted of understanding the AA problem in this setting





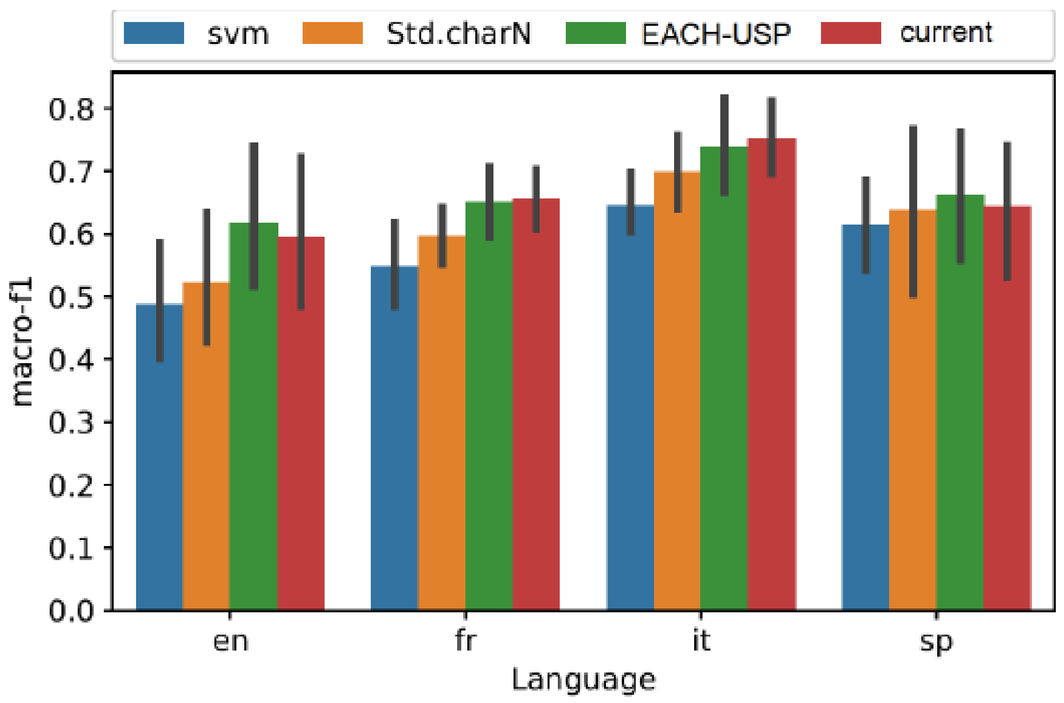
Current work

- EACH-USP ensemble model (Custódio & Paraboni, 2018) combines three models:
- Word n-grams, character n-grams, and a text distortion model
- We presently added a fourth classifier, a character ranking model that computes character adjacency graphs and uses PageRank (Page et. al., 1999) to select the most influential characters of a set of documents of a given author.
- Moreover, the openness aspect of the current AA task is dealt with by assigning the unknown author (<UNK>) label to the input text when the standard deviation of the corresponding row is below a 0:05 threshold.



Results

- Better results for Italian and French
- Still outperformed by the original EACH-USP model for English and Spanish



References

- Juola, P.: An overview of the traditional authorship attribution subtask. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome (2012).
- Custódio, J.E., Paraboni, I.: EACH-USP Ensemble Cross-domain Authorship Attribution: Notebook for PAN at CLEF 2018.
- Page, L., Brin, S., Motwani, R., Winograd, T.: The Pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (1999).

This work has been supported by grant # 2016/14223-0, São Paulo Research Foundation (FAPESP).