Cross-Domain Authorship Attribution Based on Compression

Oren Halvani and Lukas Graner

CLEF 2018 **C**onference and **L**abs of the **E**valuation **F**orum - Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 10 - 14 September 2018, Avignon ( )

Fraunhofer Institute for Secure Information Technology (SIT), Darmstadt, Germany

Fraunhofer SIT

Funded by BMBF:

Bundesministerium für Bildung und Forschung

DORIAN

OUTLINE

- Motivation
- Authorship Attribution
- COBAA (proposed method)
- Competition results
- Future work

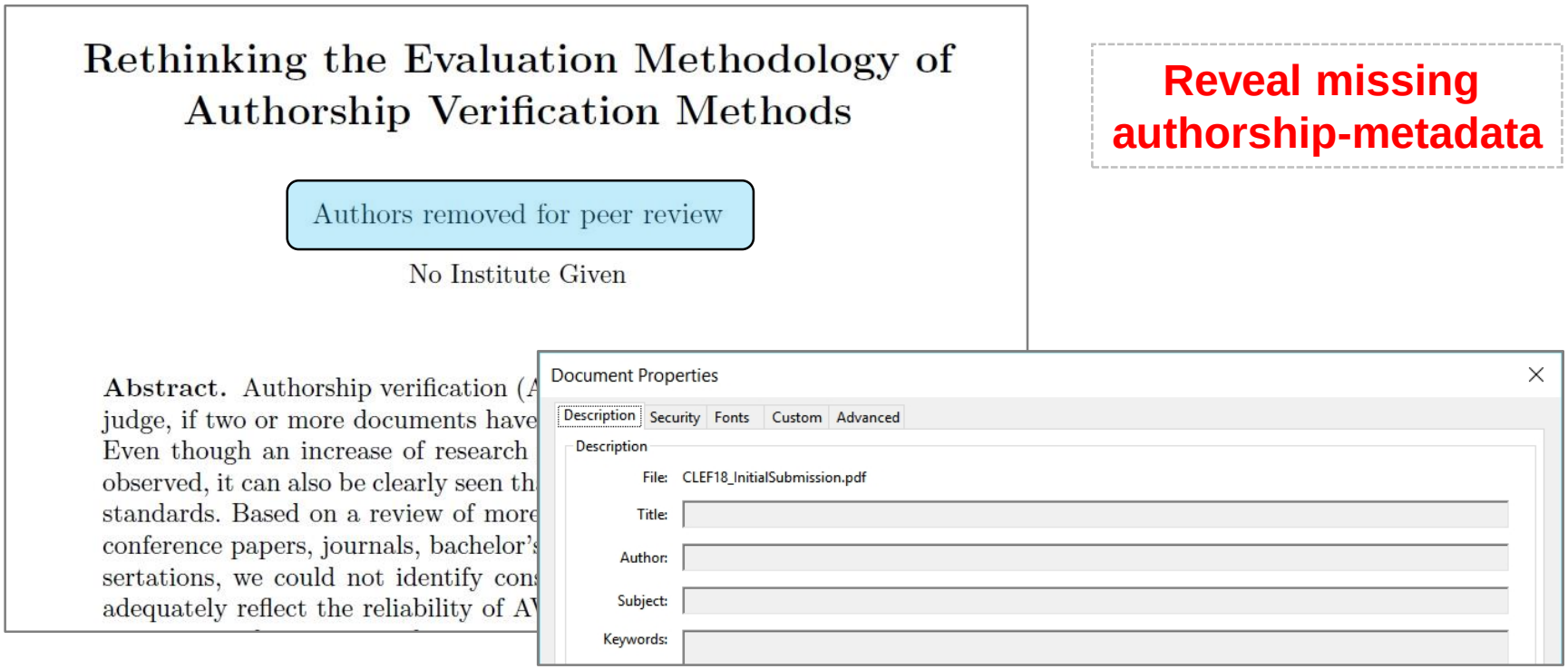
Motivation

MOTIVATION

- Attributing an anonymous text to its most likely author is a very well-studied problem, which dates back to the 19th century.
- Even after more than ten decades, the problem is still far from being solved and has become an important research subject, across many fields and domains.
- The discipline that concerns itself with this problem is known as Authorship Attribution (AA), which is the most prominent subdiscipline of authorship analysis.

MOTIVATION

- What can we do with AA ?

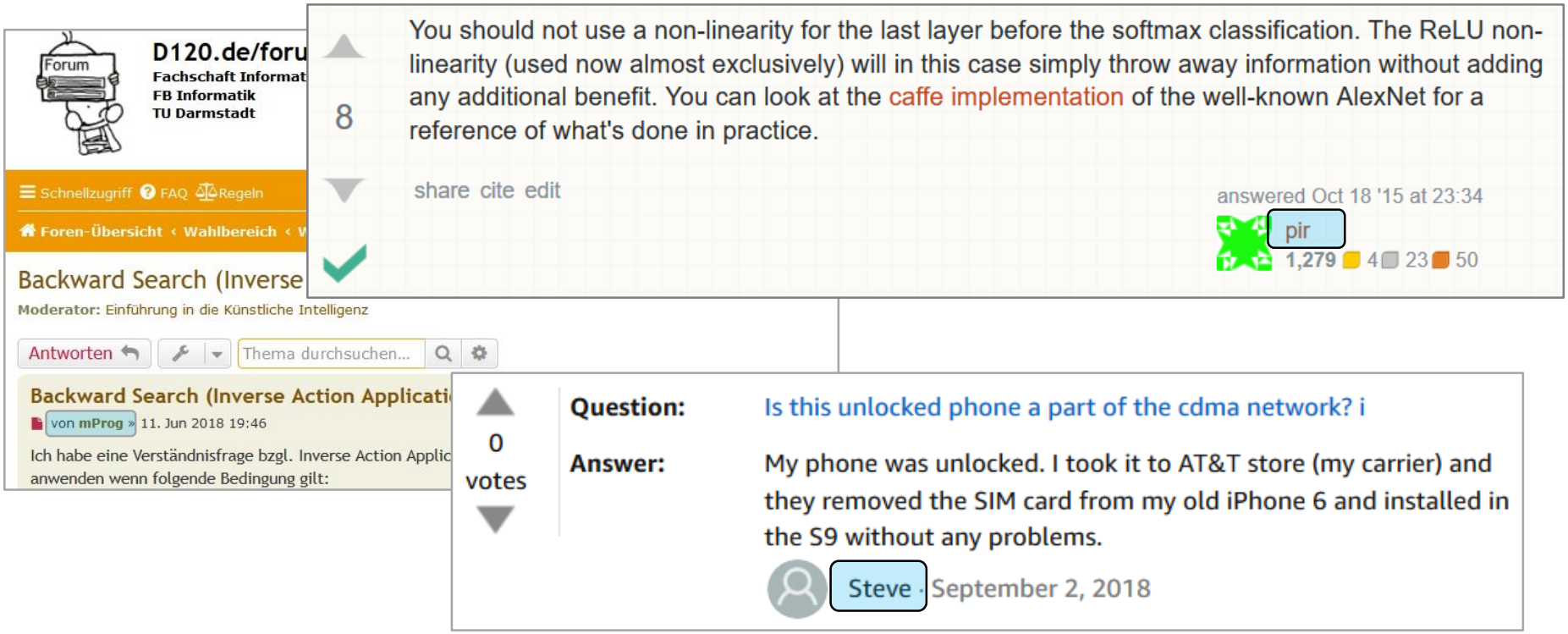


Reveal missing authorship-metadata

MOTIVATION

- What can we do with AA ?

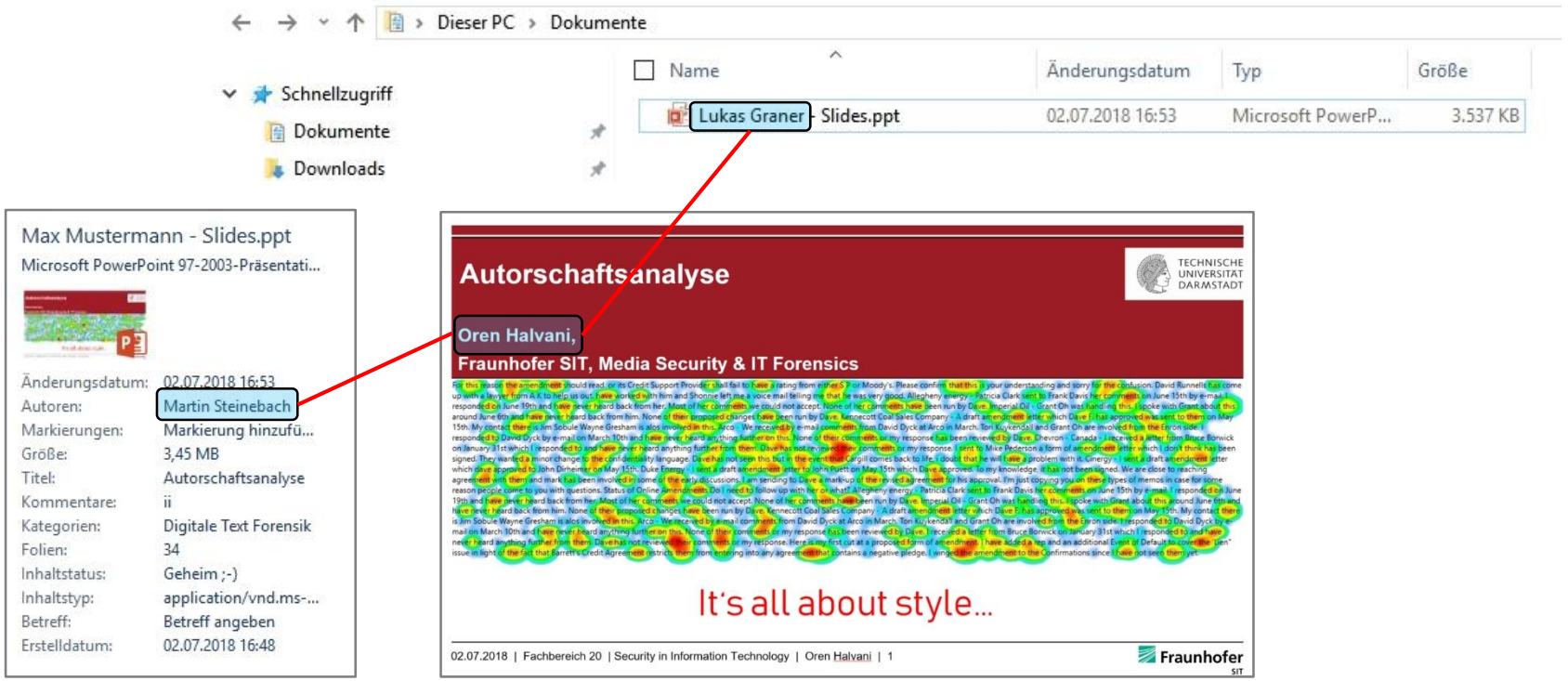
Deanonymize pseudonyms



MOTIVATION

- What can we do with AA ?

Clarify equivocal authorships



MOTIVATION

- So far, numerous machine learning models have been applied to solve AA, including neural networks, SVMs, random forests, logistic regression, *k*-NN, Naïve Bayes and many more.

- The common denominator of these is that they rely on explicitly defined features that serve as an input for the chosen machine learning model.

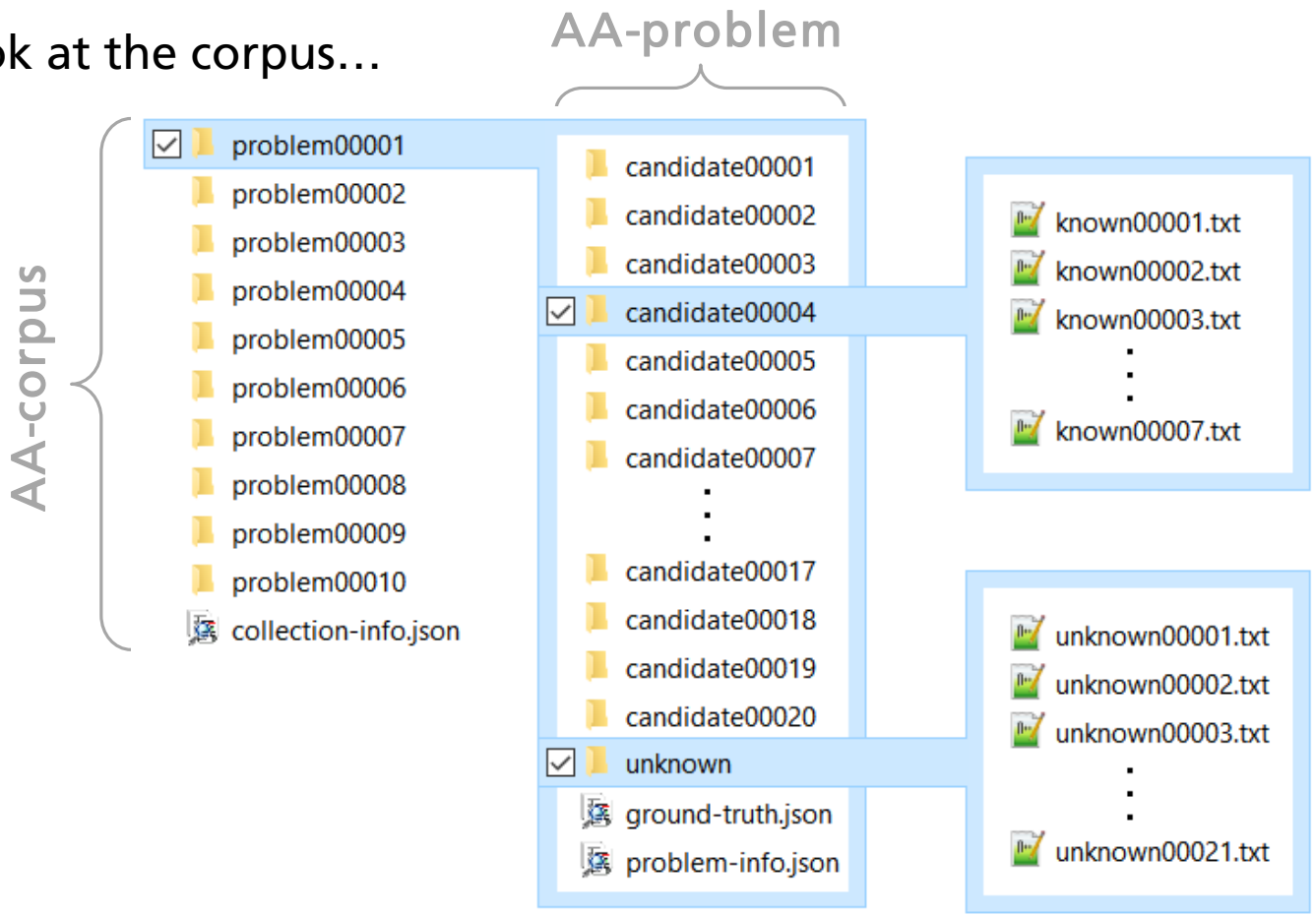
- **Question:** Can we avoid the manual feature engineering process…?

COBAA

(Compression-Based Authorship Attribution)

COBAA

- Lets first look at the corpus…



COBAA

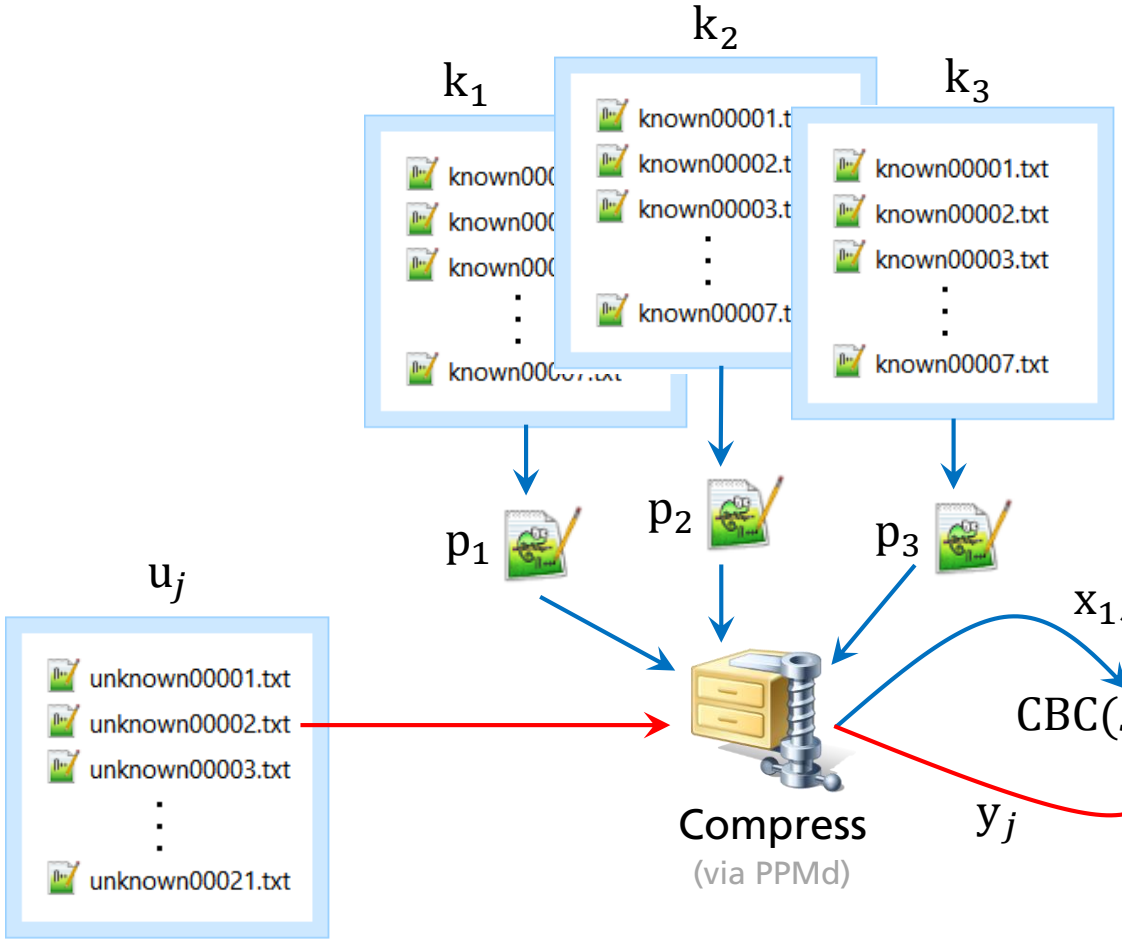
1) For each candidate, COBAA concatenates all of its texts into one profile.

2) Next, each candidate profile as well as each unknown document is compressed using the PPMd compression algorithm.

3) Afterwards, COBAA takes all compressed documents and computes pairwise similarities between a candidate profile and an unknown text via **C**ompression-**B**ased **C**osine $\mathrm{CBC}(\cdot,\cdot)$

4) For each unknown text, the resulting similarity scores are sorted in an descending order such that the candidate with the highest similarity is assumed to be the true author.

COBAA



u_j = unknown text
k_i = known candidate texts
p_i = candidate profile
x_i = compressed profile
y_j = compressed unknown text

$C(\cdot)$ denotes the **length** of the compressed text

$$\mathrm{CBC}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{C(x) + C(y) - C(xy)}{\sqrt{C(x)C(y)}}$$

Competition results

COMPETITION 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

[PAN18]

Performance of all submitted approaches in the cross-domain AA task, using several evaluation measures (ranked by macro F_1).

COMPETITION 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

[PAN18]

Attribution results per language (ranked by macro F_1).

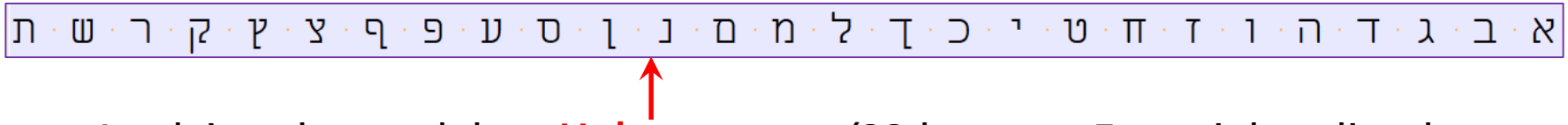
Future work

FUTURE WORK

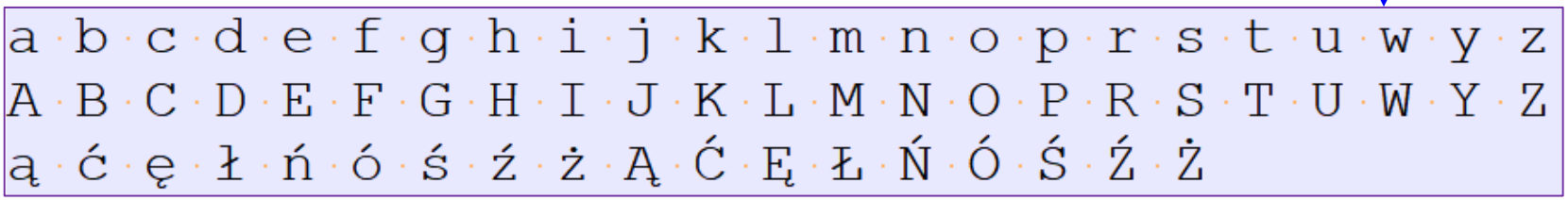
- COBAA is facing a number of issues, which are subject for future work:
  - In its current form, the language model constructed by PPMd remains a **gray box**. Even though we have access to the encoded character sequences (similar to character n-grams), it must be investigated how these contribute to the computed similarity.
  - Unlike other AA approaches that consider a subset of features, COBAA considers the entire text. Therefore, we have to investigate how much the **topic** of the texts influences the classification results. At the moment it is unclear if the AA-task is degenerated to a simple topic classification.

FUTURE WORK

- **Open question:** Why COBAA performs poor regarding specific languages?
- There seems to be no connection between the number of characters in the alphabet and the classification result of the model. For example…

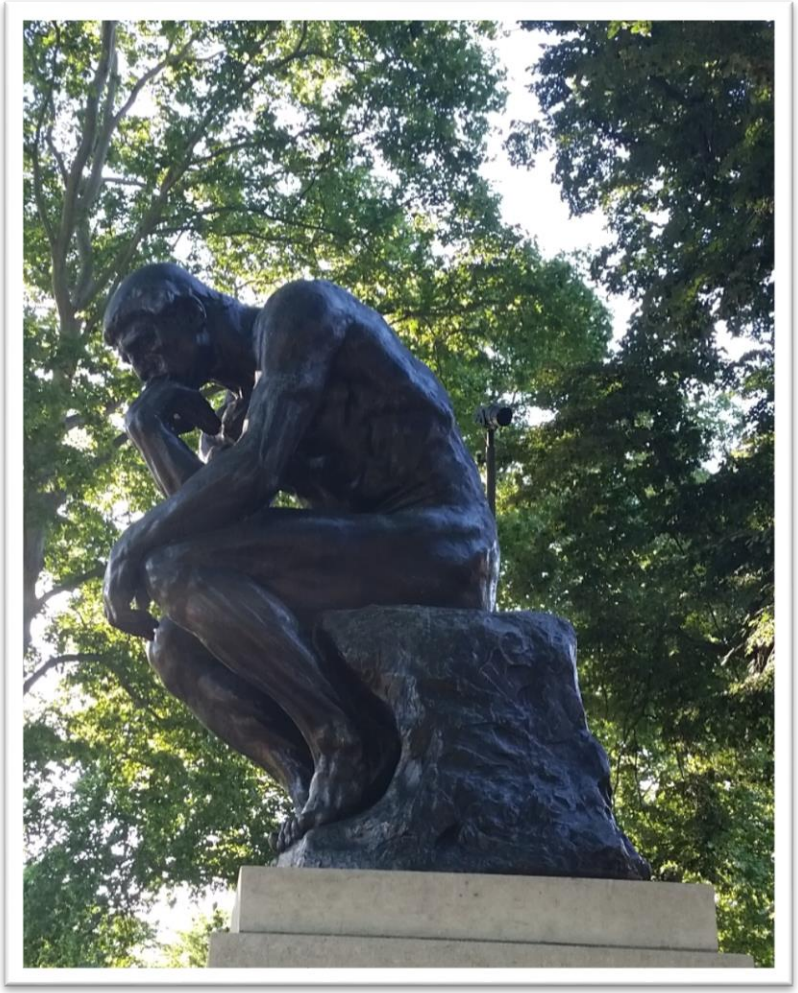


- Applying the model on **Hebrew** texts (22 letters + 5 special ending letters and no casing) leads to similar results as applying it on **Polish** texts (64 characters).



Thank you for listening!

Questions?



REFERENCES

- [PAN18] M. Kestemont, M. Tschugnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast. *Working Notes Papers of the CLEF 2018 Evaluation Labs - Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection*. CEUR Workshop Proceedings, 2018.