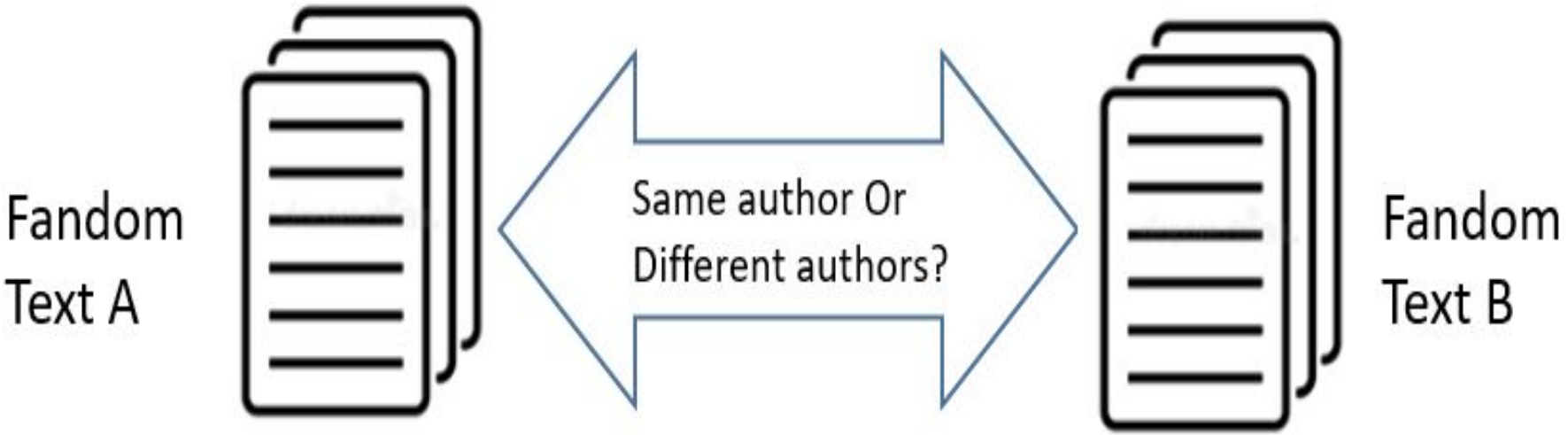
UniNE at PAN-CLEF 2020: Authorship Verification

Catherine Ikae
University of Neuchatel, Switzerland

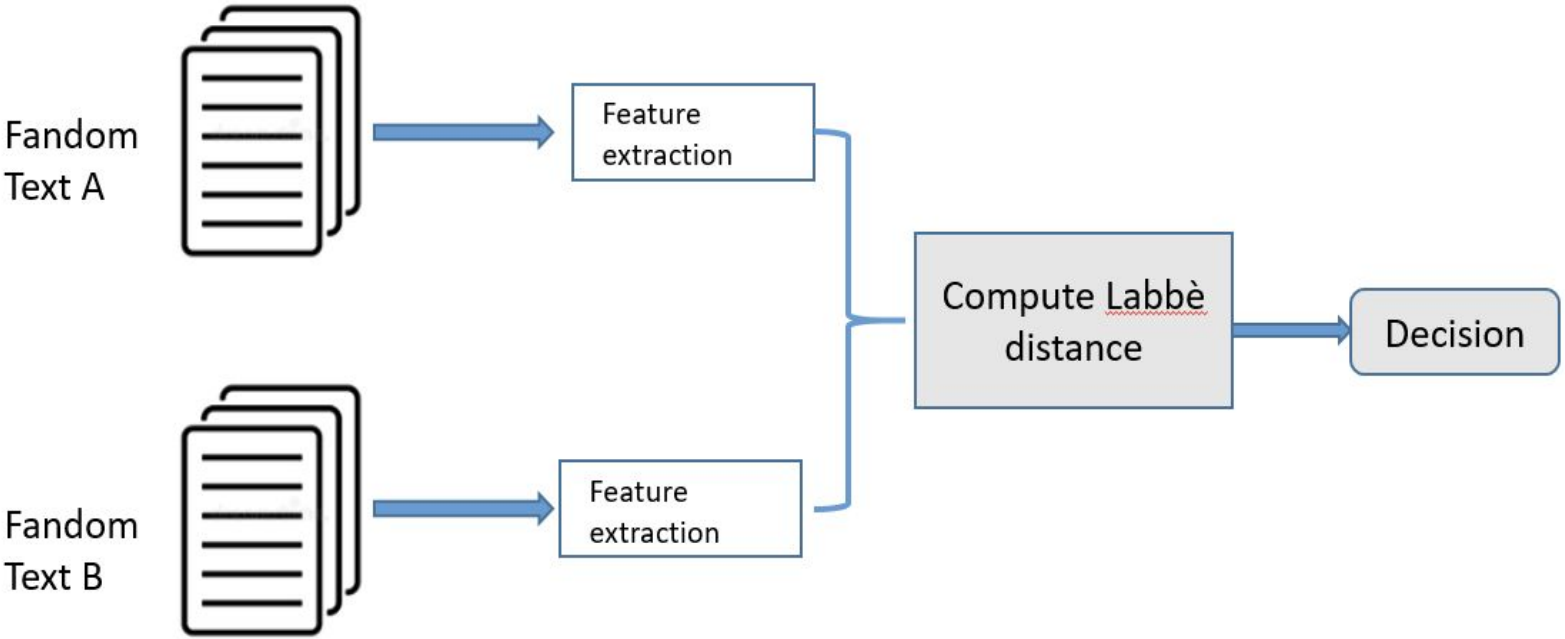
The task


Fandom
Text A
Same author Or
Different authors?
Fandom
Text B

The Data

Guardians of Ga'Hoole	Tokens = 2,235	| Voc |= 1,353
I shift a bit, warily letting my eyes dart from one owl to the other -- but my eyes are trained on the Barn Owl the most. Like Hoole...so like Hoole... He turns a bit, and our eyes meet directly. I can"t describe it...in this next moment, I don"t look away, how awkward it seems. I stare into his eyes. They"re like Hoole"s... They are Barn Owl eyes, but Hoole"s eyes. They"re his eyes...Hoole"s eyes... They hold that light of valor, …		
Hetalia - Axis Powers	Tokens = 2,032	| Voc |= 1,422
"All will become one with Russia," he said, almost simply, his cheer eerie. Fists were already clenched; now they groped about, for a pan, a rifle, a sword-there was nothing. In some way, this brought her but a sigh of relief-Gilbert and Roderich, she was reminded, were not here to suffer as well. If Ivan put his giant hands on Roderich... Click, went an object, and Elizaveta was snapped into the world when her own instincts …		

The Method


Fandom
Text A
Feature
extraction
Fandom
Text B
Feature
extraction
Compute Labbè
distance
Decision

The Method

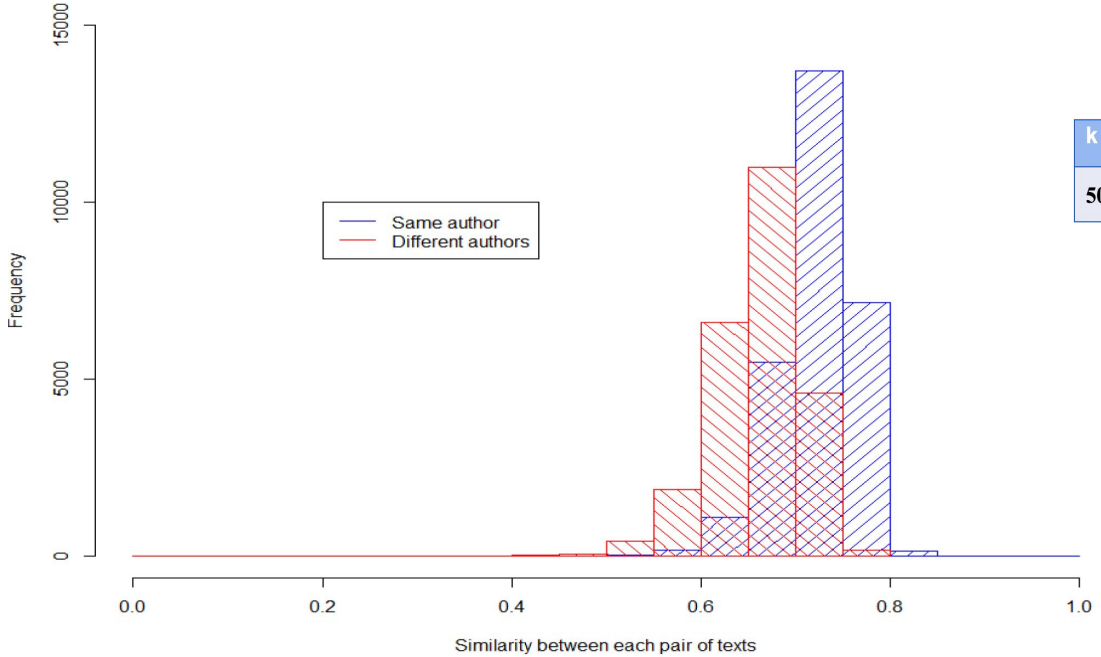
$$\text{Dist Labbé}\,(d_1, d_2) = \frac{\sum_{i=1}^{n} \widehat{rtf}_{i,1} - rtf_{i,2}}{2 * n_{d2}} \quad with \quad \widehat{rtf}_{i,1} = rtf_{i,1} * \frac{n_{d2}}{n_{d1}}$$

$$\text{Decision} \begin{cases} \text{Same author} & \text{if Sim Labbé}\,(d_1, d_2) > 0.5 \\ \text{Different authors} & \text{if Sim Labbé}\,(d_1, d_2) < 0.5 \\ \text{Non decision} & \text{otherwise} \end{cases}$$

Evaluation of the training set

k features	AUC	C@1	F_0.5_u	F1	Overall
100	0.847	0.530	0.585	0.692	0.663
150	0.851	0.535	0.585	0.692	0.665
200	0.854	0.530	0.585	0.692	0.665
250	0.855	0.530	0.585	0.692	0.666
300	0.857	0.530	0.585	0.693	0.666
350	0.858	0.531	0.585	0.693	0.666
400	0.860	0.531	0.585	0.693	0.667
450	0.860	0.531	0.585	0.693	0.667
500	0.860	0.531	0.585	0.693	0.667

Similarity Between Pairs of Text



TIRA results

k features	AUC	C@1	F_0.5_u	F1	Overall
500	0.840	0.545	0.599	0.705	0.672

Conclusion

- We proposed to select features by ranking them according to their frequency of occurrence in each text and taking only the most frequent ones.
- The similarity computation is based on the Labbé distance between two vectors.
- A labbè similarity greater than 0.5 was considered as same authors while those less than 0.5 were different authors.
- There were no cases of indecision.