UniNE at CLEF 2019: Style Change Detection

Can we uncover how many authors a document has?

Sukanya Nath
University of Neuchatel

Problem Description

Given a text, find the number of authors based on writing style changes. Using word patterns, discern an exclusive personal style of each author. Presence of multiple styles indicates presence of multiple authors.

Dataset characteristics

Statistics

- *StackExchange* forum posts in English.
- 2546 (training) / 1272 (validation) documents.
- 1-5 authors.
- Mean number of tokens: 1570.

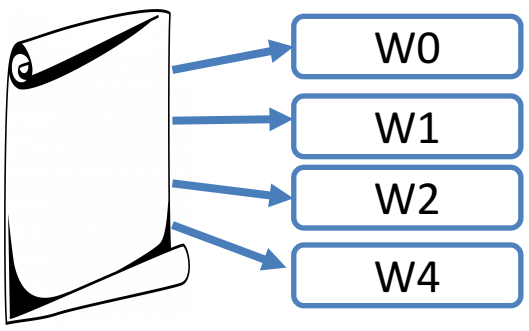

Fig. 1: Split a document into paragraph sized windows

	W0	W1	W2	W3
W0	0	0.8	**0.2**	0.55
W1	0.8	0	0.56	0.42
W2	0.2	0.56	0	0.77
W3	0.55	0.42	0.77	0

Fig. 2: Symmetrical matrix representing distance between windows

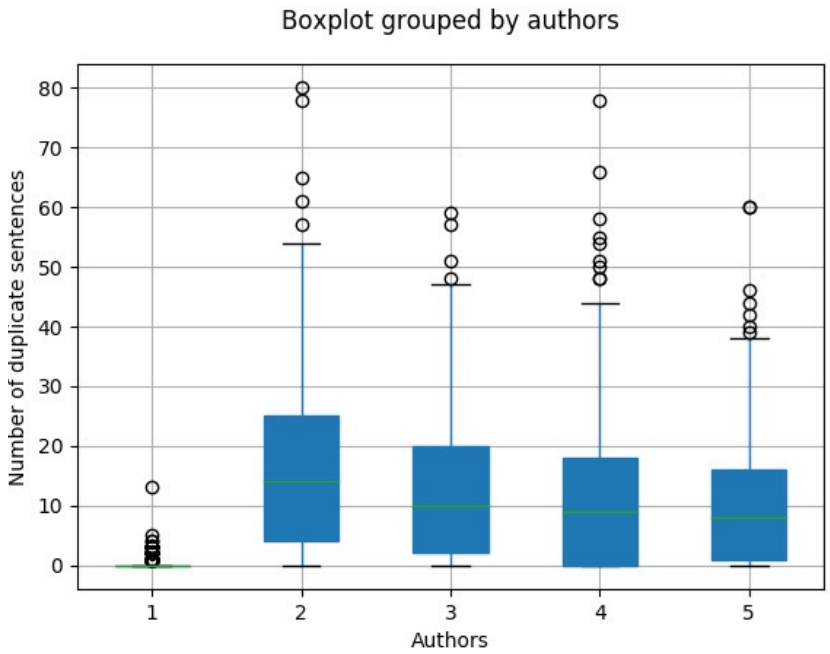

Fig. 3: Chart showing the number of duplicate sentences grouped by authors

Strategy

Preliminary steps

1. Tokenize a document into windows of paragraph length (Fig. 1).
2. Extract 50 Most Frequent Words (MFW) from each window.
3. Compare distance between windows (Matusita) and create distance matrix (see Fig. 2, 4).

Threshold Based Clustering Algorithm (TBC)

- Cluster by selecting windows with smallest distances iteratively (when forming a cluster, the closest members are included first).
- Discriminates later potential members using cluster thresholds (Fig. 5).

Window Merge Clustering Algorithm (WMC)

- Iteratively combine most similar windows to generate a new set of windows.
- Recalculate distance matrix from these new windows, for the next iteration. As a result, the clusters formed are hierarchical (Fig. 6).

Evaluation

Table 1: **Initial Evaluation Results**

	Training set			Validation set		
Algorithm	**Acc.**	**OCI**	**Rank**	**Acc.**	**OCI**	**Rank**
TBC	**0.66**	**0.83**	**0.42**	**0.65**	**0.82**	**0.42**
WMC	0.62	0.91	0.35	0.63	0.88	0.37
Combined Min	0.65	0.92	0.36	0.66	0.9	0.38

Table 2: **Final Evaluation Results** (using duplicates shown in Fig. 3)

	Training			Validation			Official Test		
Algorithm	**Acc.**	**OCI**	**Rank**	**Acc.**	**OCI**	**Rank**	**Rank**	**Acc.**	**OCI**
TBC	**0.83**	**0.87**	**0.48**	**0.83**	**0.85**	**0.49**	**0.85**	**0.87**	**0.49**
WBC	0.72	0.93	0.4	0.74	0.9	0.42	-	-	-
Combined Min	0.70	0.93	0.39	0.72	0.91	0.41	-	-	-

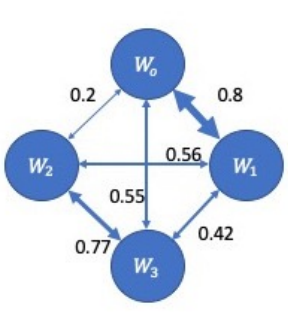

Fig. 4: Distance graph

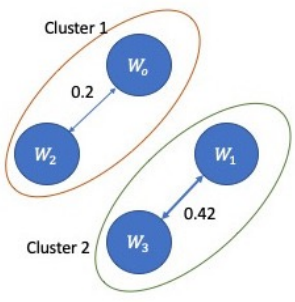

Fig. 5: Clusters representing distance between windows

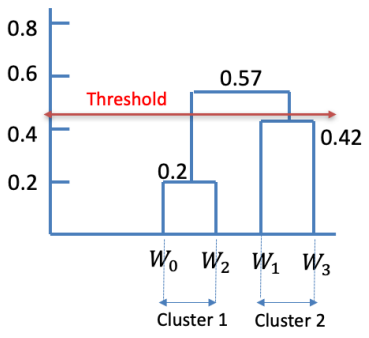

Fig. 6: Hierarchical clusters created by WMC

Conclusion

- TBC performed the best out of three approaches (improvements were observed with duplicated sentences information).
- Winner of the PAN CLEF 2019 SCD challenge.
- Demonstrated that it is possible to identify the number of authors from a relatively short piece of text without any prior training corpus per author.

Contact: Sukanya Nath, Computer Science Department
sukanya.nath@unine.ch
Web: http://www.unine.ch
University of Neuchâtel