universität innsbruck

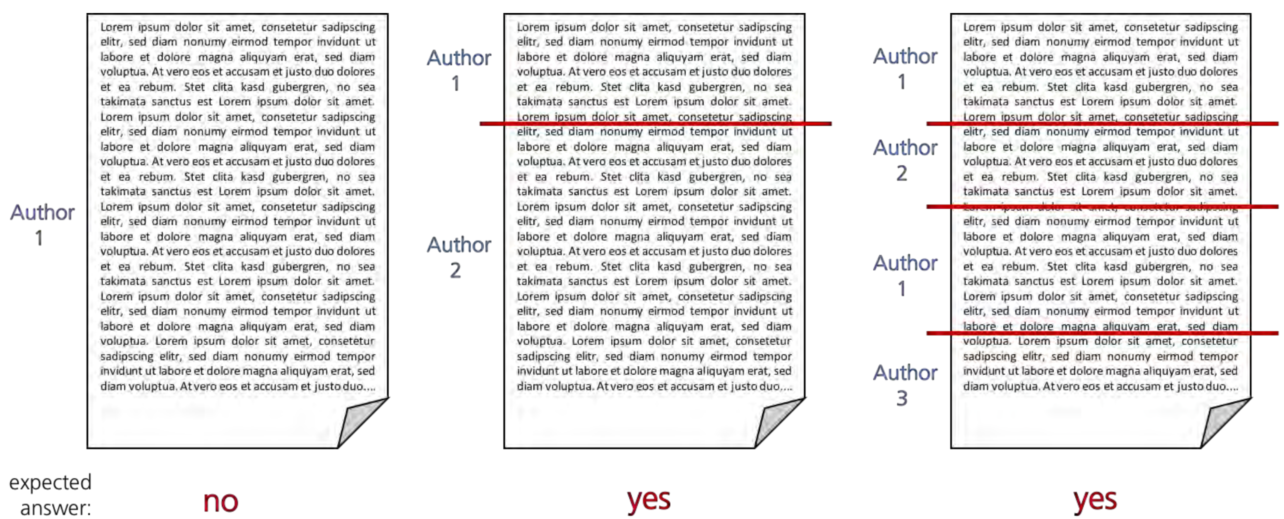
PAN@CLEF 2019
Style Change Detection Task

Eva Zangerle, Michael Tschuggnall, Martin Potthast, Benno Stein

Task Description

Author Identification

- Authorship Attribution
- **Style Change Detection**



Task Description

Given a document, participants should apply intrinsic style analyses to hierarchically answer the following questions:

(a) Is the document written by one or more authors, i.e., do style changes exist or not?

(b) If it is multi-authored, how many authors have collaborated?

Dataset

- Realistic, non-artificial and comprehensive dataset
- Requirements
  - Find multiple authors that write about the same topic
  - Find texts that are freely available and of sufficient length
  - Multi-authored texts need to contain the same topic / subtopic

- Q&A platform **StackExchange** fulfills these requirements

Dataset

StackExchange consists of several sites (170+ sites), data freely available

Each question/answer associated with

- site
- set of tags (subtopics)

Example: photography site – subtopics: lens, canon, nikon, lightroom, photoshop, ...

Dataset

- Cleaning
  - Remove links
  - Remove images
  - Remove code snippets
  - Remove bullet lists
  - Remove block quotes
  - Remove very short questions/answers
  - Remove edited questions/answers
- Using the raw texts, a **training** (50%), **validation** (25%) and **test** (25%) dataset has been created
- Each dataset contains 50% single-author documents and 50% multi-authored documents

Parameters

Parameter	Configuration Options
Number of style changes	0-10
Number of collaborating authors	1-5
Document length	300-1500 tokens
Change positions	End of paragraph, within paragraph, mixed
Segment length distribution	100-1500

Dataset

Dataset	Docs	Authors					Length	
		1	2	3	4	5	Single	Multi
Training	2,546	1,273 (50%)	325 (13%)	313 (12%)	328 (13%)	307 (12%)	977	1,604
Validation	1,272	636 (50%)	179 (14%)	152 (12%)	160 (13%)	145 (11%)	957	1,582
Test	1,210	605 (50%)	147 (12%)	144 (12%)	159 (13%)	155 (13%)	950	1,627

Evaluation

- Two subtasks, scored individually
  - Task a (binary classification): accuracy
  - Task b (classification on number of authors): Ordinal Classification Index (OCI)
- Overall score = $\frac{accuracy+(1-OCI)}{2}$

Approaches

5 registrations, 2 submissions to TIRA:

Treshold-Based and Window-Merge Clustering Methods (Sukanya Nath)

- both tasks tackled at same time
- two clustering approaches for windows
- clustering on pair-wise distance of windows - windows in same cluster are assumed to be written by same author

Feed-forward Neural Networks (Chaoyuan Zuo, Yu Zhao, Ritwik Banerjee)

- subtasks are dealt with individually
- binary classification utilizing multi-layer perceptron (single layer) on tf/idf word vector
- second task: features based on winning solution of 2018 (lexical features (POS, etc.), contracted word form, readability scores, ...), added tf/idf, three different clustering methods applied (k-means on tf-idf, hierarchical clustering on all features, MLP)

Baselines

Baseline-RND

- „advanced“ guessing using text length statistics

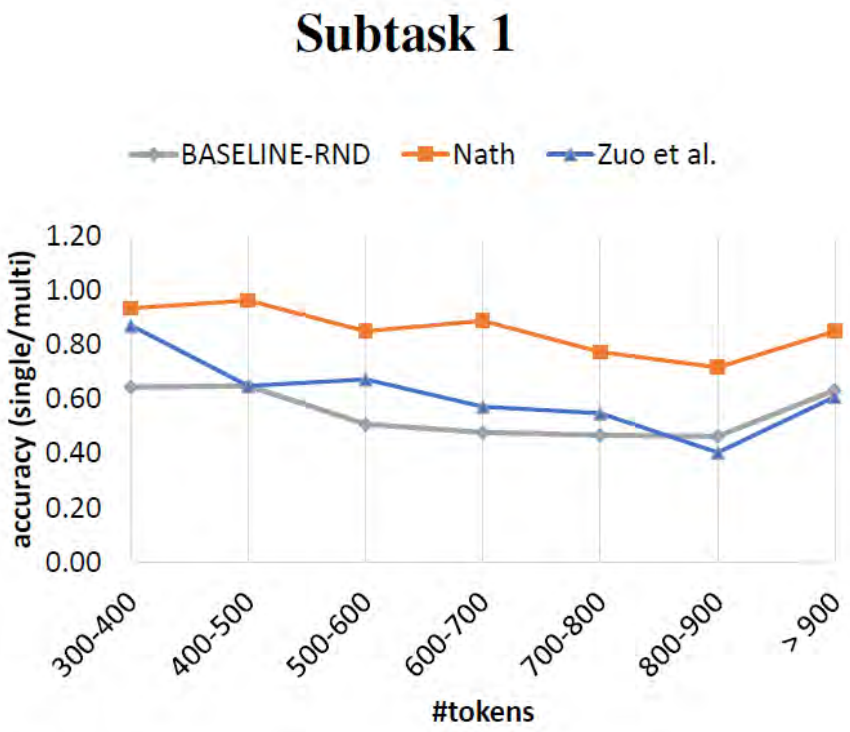
Baseline-C99

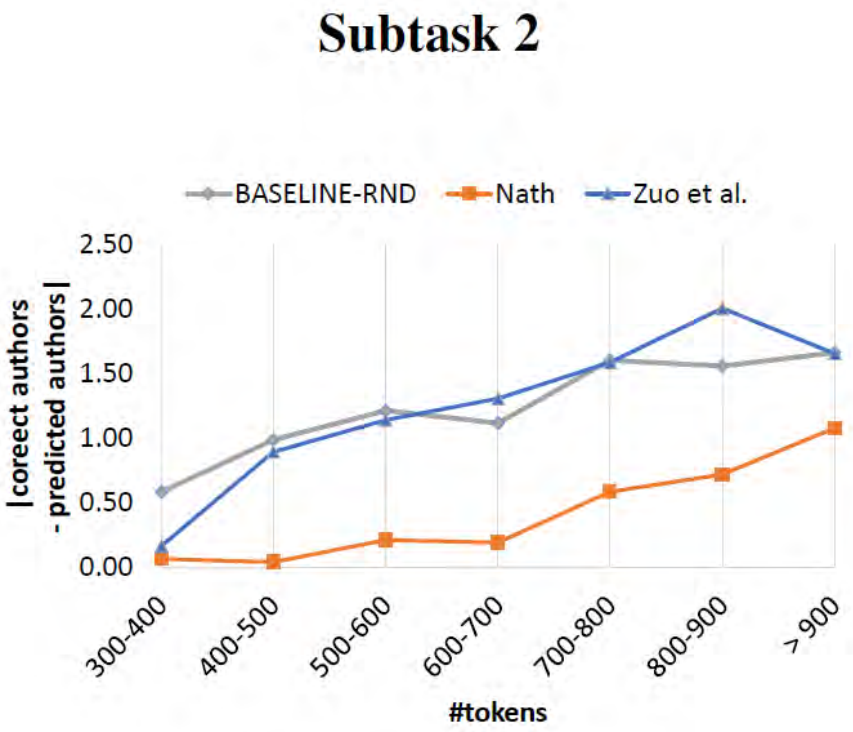
- Utilize C99 text segmentation algorithm (Choi, 2000)
- Let the algorithm determine the number of segments
- If #segments = 1: predict no style changes, otherwise predict changes

Results

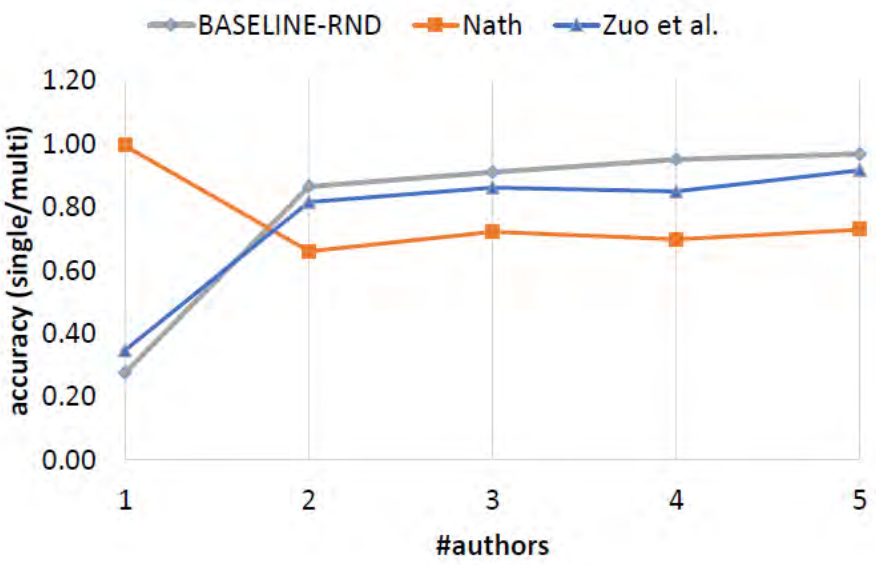
Participant	Accuracy	OCI	Rank	Runtime
Nath	0.848	0.865	0.491	00:02:23
Zuo et al.	0.604	0.809	0.398	00:25:50
Baseline-RND	0.600	0.856	0.372	-
Baseline-C99	0.582	0.882	0.350	00:00:30

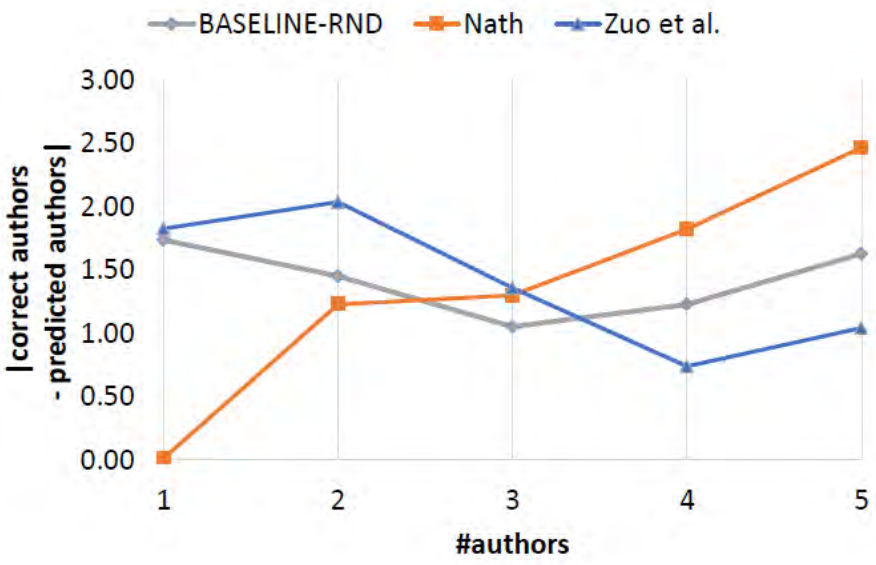
Results (#tokens)





Results (#authors)





Conclusion

- Style change detection task
- Two subtasks were tackled
- Unfortunately only two submissions
- Many exciting plans for next year, looking forward to your submissions next year ☺