An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection

PAN at CLEF-2018

D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev

P. Nakov

Qatar Computing Research Institute, HBKU, Doha, Qatar

Sofia University, Bulgaria

10-14 Sept. 2018, CLEF, Avignon
The Task

Author 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed diam nonummy eimod tempor
invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. Ali et eros et accusam et justo du
doloris et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

expected
answer: no

Author 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed diam nonummy eimod tempor
invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. Ali et eros et accusam et justo du
do bleris et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

yes

Author 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed diam nonummy eimod tempor
invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. Ali et eros et accusam et justo du
do bleris et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

yes

10-14 Sept. 2018, CLEF, Avignon
Related Work

• General approaches for Style Breach Detection:
  ▪ unsupervised methods
  ▪ stylometry and TF-IDF features
• **Wilcoxon Signed Rank test** to check whether two segments are likely to come from the same distribution (Karas et al.)
• Outlier detection using **cosine-based distance** between sentence vectors using pre-trained skip-thought models (Safin and Kuznetsova)
Data Preprocessing

• Special tokens
  ▪ http://www.java2s.com -> _URL_
  ▪ 66657345299563332126532111111 -> _LONG_NUM_
  ▪ /Users/Shared/Client/Blizzard -> _FILE_PATH_
  ▪ =================== -> _CHAR_SEQ_
  ▪ Taumatawhakatangihangakoauauo -> _LONG_WORD_

• Split hyphenated words
  ▪ Pretends-To-Be-Scrum-But-Actually-Is-Not-Even-Agile
Text Segmentation

- Sliding Window
- 1/3 overlap
- Window size: 1/3 of doc length
- Max diff of feature vectors
Lexical Features

Characters:
- spaces
- digits
- commas
- (semi)colons
- apostrophes
- quotes
- parenthesis
- number of paragraphs

Words:
- POS-tags
- short (< 4 chars)
- long (> 6 chars)
- average length
- all-caps
- capitalized

Sentences:
- question
- period
- exclamation
- short (<100 chars)
- long (>200 chars)
More Features

- Stop words: you, the, is, of, ...
- Function words: least, well, etc, whether, ...
- Readability, e.g. Flesch reading ease:
  \[ 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right) \]
- Vocabulary richness
  - Average word frequency class
    - frequency class of 'the' is 1
    - frequency class of 'doppelganger' is 19
  - Proportion of unknown words (not in corpus)
Even More Features

• Repetition
  ▪ average number of occurrences of unigrams, bigrams, ..., 5-grams

• Grammar Contractions
  ▪ *I will* vs. *I'll*
  ▪ *are not* vs. *aren't*

• Quotation variation: ’ vs. “
LightGBM + TF-IDF

- Character [2-6]-grams (up to 300k)
- Word [1-2]-grams (up to 300k)
- Logistic Regression for feature selection
- Parameter tuning to avoid overfitting
- Bagging
- Training TF-IDF on test documents
Stacking
## Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP w/ TF-IDF (Baseline)</td>
<td>validation</td>
<td>70.64</td>
</tr>
<tr>
<td>LightGBM w/ TF-IDF</td>
<td>validation</td>
<td>86.53</td>
</tr>
<tr>
<td>Stacking</td>
<td>validation</td>
<td>80.47</td>
</tr>
<tr>
<td>Stacking w/ LightGBM</td>
<td>validation</td>
<td>87.00</td>
</tr>
<tr>
<td>Stacking w/ LightGBM</td>
<td>test</td>
<td>89.35</td>
</tr>
</tbody>
</table>
## Results

**Table 10.** Evaluation results of the style change detection task.

<table>
<thead>
<tr>
<th>Submission</th>
<th>Accuracy</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zlatkova et al.</td>
<td>0.893</td>
<td>01:35:25</td>
</tr>
<tr>
<td>Hosseinia and Mukherjee</td>
<td>0.825</td>
<td>10:12:28</td>
</tr>
<tr>
<td>Safin and Ogaltsov</td>
<td>0.803</td>
<td>00:05:15</td>
</tr>
<tr>
<td>Khan</td>
<td>0.643</td>
<td>00:01:10</td>
</tr>
<tr>
<td>Schaetti</td>
<td>0.621</td>
<td>00:03:36</td>
</tr>
<tr>
<td>C99-BASELINE</td>
<td>0.589</td>
<td>00:00:16</td>
</tr>
<tr>
<td>rnd2-BASELINE</td>
<td>0.560</td>
<td>–</td>
</tr>
<tr>
<td>rnd1-BASELINE</td>
<td>0.500</td>
<td>–</td>
</tr>
</tbody>
</table>
Style Breach Detection

- **PAN 2017** dataset
  - 134 training examples
  - 0 to 8 breaches
- use the developed *supervised* method
- search for breaches *recursively*
- outperforms *baseline* models
Conclusion

- High accuracy for Style Change Detection is achievable.
- Ensembles perform best.
- Using a supervised method to detect exact breaches is promising, but needs further work.

https://github.com/machinelearning-su/style-change-detection
References