A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014

Miguel A. Sanchez–Perez, Grigori Sidorov, Alexander Gelbukh
Content

1. Task
2. Methodology
3. Adaptative behavior
4. Results
5. Conclusions
6. Future Work
Task

Text Alignment: Given a **pair of documents**, the task is to identify all contiguous maximal-length passages of **reused text** between them.
Methodology

- Preprocessing
- Seeding
- Extension
- Filtering
Preprocessing

- Sentence splitting (Kiss pretrained punkt model)
- Tokenizing (Treebank word tokenizer)
- Keeping tokens starting with a letter or digit
- Reducing to lowercase
- Stemming (Porter algorithm)
- Joining small sentences (1–2 words) with the next one
Preprocessing

PAN 2014 training corpus
Sentences length histogram (words)
Seeding

Vector representation of sentences:
TF-IDF, where sentences are “documents,” thus called TF-ISF: inverse sentence freq. “Documents”: union of sentences of both docs

Vector similarity:
Cosine similarity $\geq$ threshold $th1$
AND Dice similarity $\geq$ threshold $th2$
Seeding

Seeds: pairs of similar sentences
Extension

Grouping

Group left
Extension  Grouping

Group left

[Diagram showing groupings with arrows and colors]
Extension Grouping

Group right
Extension

Grouping

Group right
Extension

Grouping

Group left
Example: \( \text{maxGap} = 1 \)
Example: \( maxGap = 1 \)
Extension

Grouping

Example: $maxGap = 1$

Group right
Extension

Grouping

Example:
$maxGap = 1$

Group right
Example:
maxGap = 1
Extension Grouping

Example: $maxGap = 1$

Group left
## Extension

### Grouping

<table>
<thead>
<tr>
<th>Iteration</th>
<th>No plagiarism</th>
<th>None</th>
<th>Random</th>
<th>Translation</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>674</td>
<td>6803</td>
<td>6436</td>
<td>7637</td>
<td>3074</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>278</td>
<td>180</td>
<td>246</td>
<td>294</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Example:
maxGap = 2
Extension

Validation

Example: \( maxGap = 2 \)

Cosine similarity

If cosine similarity < \( th3 \)
Regroup with \( maxGap - 1 \)
Extension  Validation
Filtering

1. Resolving overlapping

\[ \text{score} = B + (1 - B) \times A, \]

2. Removing small cases

If \text{n° characters} in left side OR right side < \textit{minPlagLength} then the case is removed
Filtering

Cumulative histogram of plagiarism cases passages

Source documents

Suspicious documents
Adaptative behavior

1. Pre-processing
   minSentLength=3
   Seeding
   th1=0.33
   th2=0.33
   variant A
   Extension
   maxGap=4
   maxGapLeast=2
   minSize=1
   th3=0.4
   1
   variant B
   Extension
   maxGap=24
   maxGapLeast=2
   minSize=1
   th3=0.4
   2
2. Filtering
   minPlagLength=150
   Sum passages in src. doc (src_len) and susp. doc (susp_len)
   no
   src_len>=3susp_len
   yes
   Use variant A
   Use variant B
   End
## Results

**Training:** PAN 2014 = PAN 2013 training corpus. **Evaluation:** PAN 2014, PAN 2013.

<table>
<thead>
<tr>
<th>Obfuscation</th>
<th>2014=2013 training corpus</th>
<th>PAN 2013 test corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Plagdet Recall Prec Granul</td>
<td>Plagdet Recall Prec Granul</td>
</tr>
<tr>
<td>None</td>
<td>0.893 0.978 0.822 1.000</td>
<td>0.900 0.978 0.833 1.000</td>
</tr>
<tr>
<td>Random</td>
<td>0.888 0.858 0.921 1.000</td>
<td>0.884 0.860 0.910 1.000</td>
</tr>
<tr>
<td>Translation</td>
<td>0.883 0.890 0.877 1.000</td>
<td>0.886 0.889 0.884 1.000</td>
</tr>
<tr>
<td>Summary</td>
<td>0.577 0.424 0.994 1.043</td>
<td>0.560 0.412 0.999 1.058</td>
</tr>
<tr>
<td>Entire</td>
<td>0.877 0.879 0.877 1.002</td>
<td>0.878 0.879 0.881 1.003</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Team</th>
<th>Year</th>
<th>None</th>
<th>Random</th>
<th>Translation</th>
<th>Summary</th>
<th>Entire corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanchez-Perez</td>
<td>2013</td>
<td>0.9032</td>
<td>0.88417</td>
<td>0.88659</td>
<td>0.56070</td>
<td>0.87818</td>
</tr>
<tr>
<td>Tonejón</td>
<td>2013</td>
<td>0.92586</td>
<td>0.74711</td>
<td>0.85113</td>
<td>0.34131</td>
<td>0.8222</td>
</tr>
<tr>
<td>Kong</td>
<td>2013</td>
<td>0.8274</td>
<td>0.82281</td>
<td>0.85181</td>
<td>0.43399</td>
<td>0.81896</td>
</tr>
<tr>
<td>Suchomel</td>
<td>2013</td>
<td>0.81761</td>
<td>0.75276</td>
<td>0.67544</td>
<td>0.61011</td>
<td>0.74482</td>
</tr>
<tr>
<td>Samimi</td>
<td>2013</td>
<td>0.84963</td>
<td>0.65668</td>
<td>0.70903</td>
<td>0.11116</td>
<td>0.69913</td>
</tr>
<tr>
<td>Shrestha</td>
<td>2013</td>
<td>0.89369</td>
<td>0.68714</td>
<td>0.62719</td>
<td>0.1186</td>
<td>0.69551</td>
</tr>
<tr>
<td>Falkovskii</td>
<td>2013</td>
<td>0.82431</td>
<td>0.49959</td>
<td>0.60694</td>
<td>0.09943</td>
<td>0.61523</td>
</tr>
<tr>
<td>Nourian</td>
<td>2013</td>
<td>0.90136</td>
<td>0.35076</td>
<td>0.43864</td>
<td>0.11535</td>
<td>0.57716</td>
</tr>
<tr>
<td>Baseline</td>
<td>2013</td>
<td>0.93404</td>
<td>0.07123</td>
<td>0.1063</td>
<td>0.04462</td>
<td>0.42191</td>
</tr>
<tr>
<td>Gillam</td>
<td>2013</td>
<td>0.85884</td>
<td>0.04191</td>
<td>0.01224</td>
<td>0.00218</td>
<td>0.40059</td>
</tr>
<tr>
<td>Jayapal</td>
<td>2013</td>
<td>0.3878</td>
<td>0.18148</td>
<td>0.18181</td>
<td>0.0594</td>
<td>0.27081</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Position</th>
<th>Team</th>
</tr>
</thead>
</table>
| 0.87818  | Miguel A. Sanchez-Perez, Grigori Sidorov, and Alexander Gelbukh  
Instituto Politécnico Nacional, Mexico |
| 0.86933  | Gabriel Oberreuter and Andreas Eisele  
Innovand.io, Chile |
| 0.86806  | Yurii Palkovskii and Alexei Belov  
Zhytomyr Ivan Franko State University, Ukraine |
| 0.85930  | Demetrios Glinos  
University of Central Florida, USA |
| 0.84404  | Prasha Shrestha, Suraj Maharjan, and Thamar Solorio  
University of Alabama at Birmingham, USA |
| 0.82952  | Diego Antonio Rodríguez Torrejón and José Manuel Martín Ramos  
Universidad de Huelva, Spain |
| 0.82642  | Philipp Gross and Pashutan Modaresi  
pressrelations GmbH, Germany |
| 0.82161  | Leilei Kong, Yong Han, Zhongyuan Han, Haihao Yu, Qibo Wang, Tinglei Zhang, Haoliang Qi  
Heilongjiang Institute of Technology, China |
| 0.67220  | Samira Abnar, Mostafa Dehghani, Hamed Zamani, and Azadeh Shakery  
University of Tehran, Iran |
| 0.65954  | Faisal Alvi*, Mark Stevenson*, and Paul Clough*  
*King Fahd University of Petroleum & Minerals, Saudi Arabia, and *University of Sheffield, UK |
| 0.42191  | Baseline |
| 0.28302  | Lee Gillam and Scott Notley  
University of Surrey, UK |
Text alignment task: best result of all 11 participating systems, thanks to:

1. TF–ISF (inverse \textit{sentence} frequency) measure for “soft” removal of stopwords.
2. Recursive extension algorithm: dynamic adjustment of tolerance to gaps
3. Algorithm for resolution of overlapping cases by comparison of competing cases
4. Dynamic adjustment of parameters by type of obfuscation (summary vs. other types)
Future work

- Text reuse focused on paraphrase
- Soft cosine to measure similarity between features
- New strategy to resolve overlapping
Thanks!