A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014
Miguel A. Sanchez-Perez, Grigori Sidorov, Alexander Gelbukh

1. Task

Source Retrieval
Collection of documents
Source documents
Text Alignment
Suspicious document
Suspicious passages

2. Methodology

2.1. Pre-processing

Sentence splitting, tokenizing, removal of tokens that do not start from a letter or digit, reducing to lowercase, stemming, joining small sentences (1-2 words) with the next one.

2.1. 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

Seeds: pairs of similar sentences

Vector similarity:
Cosine similarity ≥ threshold th1
AND Dice similarity ≥ threshold th2

2.2. Extension


Resulting groups are considered plagiarism cases

2.3. Filtering

Resolving overlapping

Plagiarism cases shorter than minSentLength characters are removed

3. Adaptive behavior

4. Results

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

For each group like this:
if cosine similarity between left and right sides of the group ≤ threshold th3 then form groups again with maxGap - 1

Example:
After group left and right with maxGap = 2 Grouping with maxGap = 1

5. Conclusions

Text alignment task: best result of all 11 participating systems, thanks to:
1. TF-ISF (inverse sentence frequency) measure for “soft” removal of stopwords
2. Recursive extension algorithm: dynamic adjustment of tolerance to gaps
3. Novel algorithm for resolution of overlapping cases by comparison of competing cases
4. Dynamic adjustment of parameters by type of case (summary vs. other types)

Final result of PAN 2014 Text Alignment

Contact
Miguel A. Sanchez-Perez
masp1988@hotmail.com
Centro de Investigación en Computación,
Instituto Politécnico Nacional, http://www.cic.ipn.mx