Using a Variety of n-Grams for the Detection of Different Kinds of Plagiarism

Prasha Shrestha and Thamar Solorio
Department of Computer and Information Sciences
University of Alabama at Birmingham

Motivation

Text Alignment Task

House appraisal reports for FHA loans are not considered standard and tend to have higher fees due to the extra amount of work and time.

Internal evaluations for FHA loans are not considered standard and tend to have higher fees due to the extraordinary amount of work and time.

Discovery Channel Score worked closely together to identify the scope of the project, conduct a needs assessment, and select vendors. Markmax will work with Deloitte Consulting on software implementation.

Years, a sword called "Seiun" reappears and recalls Takemaru from the grave, and Takemaru decides he wants to play dogcatcher.

Years, sword "Seiun" reappears and grave, and Takemaru decides he wants to play dogcatcher.

Discovery Channel Score worked together closely to the scope across the cliche, conduct a assessment, and select vendors. Markmax will work with Deloitte of software.

Certain Methods are Better Suited to Detect Certain Kinds of Obfuscations.

Methodology

Seeding

Matching

We had a summer season

The summer was hot

Exact Matches (stopword n-grams)

Unordered Matches (named entity and all-words n-grams)

Merging

Merged

Not Merged

Passages in the vicinity plagiarised passages are more likely to be plagiarised.

Postprocessing

Unordered matching of shaded portions of text using all word n-grams with more relaxed parameters

Remove passages that are too short.

Results

Table 1. Evaluation Results for the Training Corpus

<table>
<thead>
<tr>
<th>Plagiarism Type</th>
<th>Precision</th>
<th>Recall</th>
<th>Granularity</th>
<th>Plagdet</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Obfuscation</td>
<td>0.99721</td>
<td>0.88425</td>
<td>1.00000</td>
<td>0.89040</td>
</tr>
<tr>
<td>Random Obfuscation</td>
<td>0.90482</td>
<td>0.71842</td>
<td>1.27195</td>
<td>0.67649</td>
</tr>
<tr>
<td>Translation Obfuscation</td>
<td>0.87069</td>
<td>0.61710</td>
<td>1.23666</td>
<td>0.62194</td>
</tr>
<tr>
<td>Summary Obfuscation</td>
<td>0.91405</td>
<td>0.10747</td>
<td>1.98930</td>
<td>0.12174</td>
</tr>
</tbody>
</table>

Table 2. Evaluation Results for the Test Corpus

<table>
<thead>
<tr>
<th>Plagiarism Type</th>
<th>Precision</th>
<th>Recall</th>
<th>Granularity</th>
<th>Plagdet</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Obfuscation</td>
<td>0.99902</td>
<td>0.89933</td>
<td>1.00083</td>
<td>0.89369</td>
</tr>
<tr>
<td>Random Obfuscation</td>
<td>0.92335</td>
<td>0.71461</td>
<td>1.30962</td>
<td>0.66714</td>
</tr>
<tr>
<td>Translation Obfuscation</td>
<td>0.88008</td>
<td>0.63618</td>
<td>1.26184</td>
<td>0.62719</td>
</tr>
<tr>
<td>Summary Obfuscation</td>
<td>0.90455</td>
<td>0.09897</td>
<td>1.83696</td>
<td>0.11860</td>
</tr>
<tr>
<td>Overall</td>
<td>0.87461</td>
<td>0.73814</td>
<td>1.22084</td>
<td>0.69551</td>
</tr>
<tr>
<td>Best System</td>
<td>0.89484</td>
<td>0.76190</td>
<td>1.00141</td>
<td>0.82220</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.92939</td>
<td>0.34223</td>
<td>1.27473</td>
<td>0.42191</td>
</tr>
</tbody>
</table>

Conclusion

- Three different types of n-grams, each with a different characteristic, collectively can catch passages obfuscated differently. These methods can be combined in such a way that they do not hurt the overall quality of detection of the system.

- Main area that needs improvement is granularity. Named entity n-gram matching inherently produces sparse matches. Although we removed too short passages, removing any more would cost us precision and recall.

- Our postprocessing approach helps to increase detection without compromising the precision. Making our postprocessing approach lenient will help us reduce granularity but will decrease the precision.

- Our approach produces comparatively balanced results across different forms of obfuscations.

Acknowledgement

This research is partially funded by The Office of Naval Research under grant N00014-12-1-0217.