UFRGS@PAN2010: Detecting External Plagiarism
Lab Report for PAN at CLEF 2010

Rafael Corezola Pereira, Viviane P. Moreira, and Renata Galante

Universidade Federal do Rio Grande do Sul
PAN 2010
The Task

• Detect the plagiarized passages in the suspicious documents and their corresponding text fragments in the source documents even if the documents are written in different languages

• Known as **External** plagiarism analysis
The Proposed Approach

1. Language Normalization
2. Retrieval
3. Feature Selection + Classifier Training
4. Plagiarism Analysis
5. Post-Processing

(1) Training Corpus
(2) Norm. Susp. Documents
(3) Suspicious Document
(4) Candidate Documents
(5) Final Result
(1) Language Normalization

• All documents are converted into a common language

• English was chosen
  – More translation resources
  – One of the easiest languages to translate into

• Used a language guesser and an automatic translation tool
The Proposed Approach

1. Language Normalization
2. Retrieval
3. Feature Selection + Classifier Training
4. Plagiarism Analysis
5. Post-Processing

Training Corpus → Feature Selection + Classifier Training → Classification Model

Suspicious Documents → Language Normalization
Original Documents → Language Normalization

Suspicious Document → Retrieval
Candidate Documents → Retrieval
Index

Suspicious Documents
Norm. Susp. Documents
Norm. Orig. Documents

Preliminary Result → Post-Processing → Final Result
(2) Retrieval of Candidate Documents

- Problem: It is not feasible to perform exhaustive comparisons

- Solution: Use the suspicious document as a query to be sent to an IR system

- Documents are divided into paragraphs (subdocuments)

- At the end of this phase, we have a list of at most ten candidate subdocuments for each passage in the suspicious document
The Proposed Approach

1. Language Normalization
2. Retrieval
3. Feature Selection + Classifier Training
4. Plagiarism Analysis
5. Post-Processing

- Training Corpus
- Suspicious Documents
- Original Documents
- Norm. Susp. Documents
- Norm. Orig. Documents
- Index
- Suspicious Document
- Candidate Documents
- Preliminary Result
- Final Result
(3) Feature Selection and Classifier Training

• The goal of the classifier is to decide whether a suspicious passage is plagiarized from a candidate subdocument
• Annotated synthetic examples used for training
• J48 classification algorithm
• Features
  – The cosine similarity between the suspicious passage and the candidate subdocument
  – The similarity score assigned by the IR system
  – The position of the candidate subdocument in the rank generated
  – The length (in characters) of the suspicious and the candidate subdocument
The Proposed Approach

1. Language Normalization
2. Retrieval
3. Feature Selection + Classifier Training
4. Plagiarism Analysis
5. Post-Processing

- Training Corpus
- Original Documents
- Suspicious Documents
- Norm. Susp. Documents
- Norm. Orig. Documents
- Index
- Candidate Documents
- Preliminary Result
- Final Result

For each Index

(1) Language Normalization
(2) Retrieval
(3) Feature Selection + Classifier Training
(4) Plagiarism Analysis
(5) Post-Processing
(4) Plagiarism Analysis

- Submit the test instances to the trained classifier and let it decide whether the suspicious passage is, in fact, plagiarized from one of the candidate subdocuments.
The Proposed Approach

1. Language Normalization
2. Retrieval
3. Feature Selection + Classifier Training
4. Plagiarism Analysis
5. Post-Processing

- Training Corpus
- Suspicious Documents
- Original Documents
- Norm. Susp. Documents
- Norm. Orig. Documents
- Index
- Suspicious Document
- Candidate Documents
- Preliminary Result
- Final Result
(5) Result Post-Processing

• Join the contiguous plagiarized passages detected by the method in order to decrease its final granularity score

<?xml version="1.0" encoding="UTF-8"?>
<document reference="A.txt">
<feature name="detected-plagiarism"
    this_offset="1000"
    this_length="500"
    source_reference="B.txt"
    source_offset="3000"
    source_length="500"
/>
<feature name="detected-plagiarism"
    this_offset="1500"
    this_length="300"
    source_reference="B.txt"
    source_offset="3500"
    source_length="300"
/>
</document>

<?xml version="1.0" encoding="UTF-8"?>
<document reference="A.txt">
<feature name="detected-plagiarism"
    this_offset="1000"
    this_length="800"
    source_reference="B.txt"
    source_offset="3000"
    source_length="800"
/>
</document>
Experiments

- Terrier IR System
  - Porter Stemmer
  - Stop-Word Removal (list of 733 words)

- Weka Data Mining Software
  - J48 classification algorithm

- Google Translator (as language guesser)

- LEC Power Translator
Summary

• Overall results (7th place) / No Obfuscation vs. Translated

<table>
<thead>
<tr>
<th>---</th>
<th>Competition</th>
<th>Only External Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.4036 (7th)</td>
<td>0.4966</td>
</tr>
<tr>
<td>Precision</td>
<td>0.7242 (11th)</td>
<td>0.7242</td>
</tr>
<tr>
<td>Granularity</td>
<td>1.0024 (1th)</td>
<td>1.0017</td>
</tr>
<tr>
<td>Overall Score</td>
<td>0.5175 (7th)</td>
<td>0.5881</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>Translated</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.68</td>
<td>0.60 (4th)</td>
<td>88</td>
</tr>
<tr>
<td>Recall</td>
<td>0.51</td>
<td>0.43 (4th)</td>
<td>84</td>
</tr>
<tr>
<td>Granularity</td>
<td>1.00</td>
<td>1.01 (4th)</td>
<td>99</td>
</tr>
</tbody>
</table>

• The length of the plagiarized passage affects the results considerably
  – The larger the passage the easier the detection

• Low performance while detecting short plagiarized passages
  – Partially explained by our decision of indexing only the subdocuments with length greater than 250 characters
Questions?

Rafael Corezola Pereira, Viviane P. Moreira, and Renata Galante

Universidade Federal do Rio Grande do Sul
PAN 2010
Processing Time

- **Notebook**
  - Intel Core 2 Duo 1.6GHz
  - 2GB RAM
  - HD 5400 RPM

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Analysis Time</strong></td>
<td>~ 230 hours</td>
</tr>
<tr>
<td><strong>Average Time / Suspicious Document</strong></td>
<td>52 seconds</td>
</tr>
<tr>
<td><strong>KB Analyzed / Minute</strong></td>
<td>236KB</td>
</tr>
</tbody>
</table>