JOINT TALK ON THREE DATA SUBMISSIONS TO TEXT ALIGNMENT AND ONE SOURCE RETRIEVAL ALGORITHM

Mostafa Dehghani
ICT Research Institute, ACECR, Iran
September, 10, 2015
Outline of My Talk
Outline of My Talk

A. Data Submissions to Text Alignment:
   - Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation
   - Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus
   - Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus
Outline of My Talk

A. Data Submissions to Text Alignment:
   - Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation
   - Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus
   - Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus
   - Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection
Outline of My Talk

A. Data Submissions to Text Alignment:
   - Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation
   - Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus
   - Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus
   - Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection

B. Source Retrieval Plagiarism Detection based on Noun Phrase and Keyword Phrase Extraction
Data Submissions to Text Alignment
Corpus Construction Steps

- 
- 
- 
- 
- 
- 
- 
- 
-
Corpus Construction Steps

- Preprocessing
Corpus Construction Steps

- Preprocessing
- Clustering
Corpus Construction Steps

- Preprocessing
- Clustering
- Fragment Extraction
Corpus Construction Steps

- Preprocessing
- Clustering
- Fragment Extraction
- Fragment Obfuscation
Corpus Construction Steps

- Preprocessing
- Clustering
- Fragment Extraction
- Fragment Obfuscation
- Inserting Plagiarism Cases into Documents
Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation

Data resource:

Wikipedia Articles
Mono Lingual Persian Corpus
Mono Lingual Persian Corpus

- Preprocessing
  - Persian is one of the Indo-European languages which have borrowed its script from Arabic, a member of the Semitic language family.
Mono Lingual Persian Corpus

- **Preprocessing**
  - Persian is one of the Indo-European languages which have borrowed its script from Arabic, a member of the Semitic language family

- **Clustering**
  - In this step, collection of Wikipedia documents clustered into different topically related groups
  - A bipartite graph of documents-categories was created to cluster the documents
  - In the next step, the Infomap community detection algorithm was applied to the graph and all communities were detected
  - Finally, Documents within a community are considered as one cluster
Mono Lingual Persian Corpus

- Fragment Extraction
Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents: 25% with Plagiarism – 25% no Plagiarism
Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents: 25% with Plagiarism – 25% no Plagiarism
  - The task of the fragment extraction is to extract fragments from source documents.
Mono Lingual Persian Corpus

Fragment Extraction

- Divided Documents into Two Categories:
  - 50% Source Documents
  - 50% Suspicious Documents: 25% with Plagiarism – 25% no Plagiarism
- The task of the fragment extraction is to extract fragments from source documents.

<table>
<thead>
<tr>
<th>Fragment Length</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>30 – 50 words</td>
</tr>
<tr>
<td>Medium</td>
<td>150 – 250 words</td>
</tr>
<tr>
<td>Long</td>
<td>300 – 500 words</td>
</tr>
</tbody>
</table>
Mono Lingual Persian Corpus

- Fragment Extraction
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents: 25% with Plagiarism – 25% no Plagiarism
  - The task of the fragment extraction is to extract fragments from source documents.
Mono Lingual Persian Corpus

- **Fragment Extraction**
  - Divided Documents into Two Categories:
    - 50% Source Documents
    - 50% Suspicious Documents: 25% with Plagiarism – 25% no Plagiarism
  - The task of the fragment extraction is to extract fragments from source documents.

- **Fragment Obfuscation**
Mono Lingual Persian Corpus

➢ Fragment Extraction
  o Divided Documents into Two Categories:
    • 50% Source Documents
    • 50% Suspicious Documents: 25% with Plagiarism – 25% no Plagiarism
  o The task of the fragment extraction is to extract fragments from source documents.

➢ Fragment Obfuscation
  o Artificial Obfuscation
    • None (No Obfuscation)
    • Random Change of Order
    • POS-preserving Change of Order
    • Synonym Substitution
    • Addition / Deletion
Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document’s length, one or more plagiarism cases are selected.
Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document’s length, one or more plagiarism cases are selected.

<table>
<thead>
<tr>
<th>Plagiarism per Document</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Little</strong></td>
</tr>
<tr>
<td><strong>Medium</strong></td>
</tr>
<tr>
<td><strong>Much</strong></td>
</tr>
<tr>
<td><strong>Very Much</strong></td>
</tr>
</tbody>
</table>
Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document’s length, one or more plagiarism cases are selected.
Inserting Plagiarism Cases into suspicious Documents

- In this step, according to suspicious document’s length, one or more plagiarism cases are selected.
- Each of selected cases inserted at random positions in suspicious document.
Mono Lingual Persian Corpus

- Inserting Plagiarism Cases into suspicious Documents
  - In this step, according to suspicious document’s length, one or more plagiarism cases are selected.
  - Each of selected cases inserted at random positions in suspicious document.
  - Each suspicious document and its corresponding source documents are selected from one cluster.
Inserting Plagiarism Cases into Suspicious Documents

- In this step, according to suspicious document’s length, one or more plagiarism cases are selected.
- Each of selected cases inserted at random positions in suspicious document.
- Each suspicious document and its corresponding source documents are selected from one cluster.

```xml
<document xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"....
  <feature name="project-gutenberg" offset_number="7243" url="http://www.gutenberg.org/files/....
  <feature name="language" value="en"/>
  <feature name="artificial-plagiarism" translation="false" obfuscation="none"
    this_offset="487" this_length="4218" source_reference="source-document03471.txt"
    source_offset="10866" source_length="4226"/>

  <feature name="artificial-plagiarism" translation="false" obfuscation="low"
    this_offset="7507" this_length="1872" source_reference="source-document03471.txt"
    source_offset="4846" source_length="1792"/>

  <feature name="artificial-plagiarism" translation="false" obfuscation="low"
    this_offset="10626" this_length="805" source_reference="source-document03471.txt"
    source_offset="2399" source_length="800"/>
</document>
```
# Mono Lingual Persian Corpus

## Results

<table>
<thead>
<tr>
<th>Documents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of source documents:</td>
<td>1057</td>
</tr>
<tr>
<td>The number of suspicious documents:</td>
<td></td>
</tr>
<tr>
<td>With plagiarism:</td>
<td>529</td>
</tr>
<tr>
<td>No plagiarism:</td>
<td>528</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plagiarism Cases</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of plagiarism cases:</td>
<td></td>
</tr>
<tr>
<td>No obfuscation cases:</td>
<td>259</td>
</tr>
<tr>
<td>With obfuscation cases:</td>
<td>564</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plagiarism per Document</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of Little plagiarized documents:</td>
<td>301</td>
</tr>
<tr>
<td>The number of Medium plagiarized documents:</td>
<td>80</td>
</tr>
<tr>
<td>The number of Much plagiarized documents:</td>
<td>96</td>
</tr>
<tr>
<td>The number of Very much plagiarized documents:</td>
<td>52</td>
</tr>
</tbody>
</table>
Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus

Data resources:

- Wikipedia Articles
- SemEval Dataset
Mono Lingual English Corpus

- Clustering
Mono Lingual English Corpus

- Clustering
- Fragment Extraction
  - Method 1: The fragments are extracted from source documents.
  - Method 2: The fragments are generated based on SemEval dataset sentences.
Mono Lingual English Corpus

- Clustering
- Fragment Extraction
  - Method 1: The fragments are extracted from source documents.
  - Method 2: The fragments are generated based on SemEval dataset sentences.

<table>
<thead>
<tr>
<th>Fragment Length</th>
<th>3 – 5 sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>6 – 8 sentences</td>
</tr>
<tr>
<td>Medium</td>
<td>9 – 12 sentences</td>
</tr>
</tbody>
</table>
Mono Lingual English Corpus

- Clustering
- Fragment Extraction
  - Method 1: The fragments are extracted from source documents.
  - Method 2: The fragments are generated based on SemEval dataset sentences.
Mono Lingual English Corpus

- Fragment Obfuscation
Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
Fragment Obfuscation

- Artificial Obfuscation
- Simulated Obfuscation
  - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
  - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.
Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.

<table>
<thead>
<tr>
<th>Degree</th>
<th>Similarity Scores of Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Low</td>
<td>-</td>
</tr>
<tr>
<td>Medium</td>
<td>25% - 45%</td>
</tr>
<tr>
<td>High</td>
<td>45% - 65%</td>
</tr>
</tbody>
</table>
Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.
Mono Lingual English Corpus

- Fragment Obfuscation
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.

- Inserting Plagiarism Cases into Documents
Mono Lingual English Corpus

- **Fragment Obfuscation**
  - Artificial Obfuscation
  - Simulated Obfuscation
    - The pairs of sentences from the SemEval dataset with their corresponding similarity score are used for constructing the simulated plagiarism cases.
    - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment.

- **Inserting Plagiarism Cases into Documents**

<table>
<thead>
<tr>
<th>Plagiarism per Document</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hardly</strong></td>
<td>5% - 20%</td>
</tr>
<tr>
<td><strong>Medium</strong></td>
<td>20% - 40%</td>
</tr>
<tr>
<td><strong>Much</strong></td>
<td>40% - 60%</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Documents</strong></td>
<td></td>
</tr>
<tr>
<td>The number of source documents:</td>
<td>3309</td>
</tr>
<tr>
<td>The number of suspicious documents:</td>
<td>952</td>
</tr>
<tr>
<td><strong>Plagiarism per Document</strong></td>
<td></td>
</tr>
<tr>
<td>Hardly (5% - 20%)</td>
<td>60%</td>
</tr>
<tr>
<td>Medium (20% - 40%)</td>
<td>25%</td>
</tr>
<tr>
<td>Much (40% - 60%)</td>
<td>15%</td>
</tr>
<tr>
<td><strong>Plagiarism Cases</strong></td>
<td></td>
</tr>
<tr>
<td>The number of plagiarism cases:</td>
<td></td>
</tr>
<tr>
<td>- No obfuscation cases:</td>
<td>10%</td>
</tr>
<tr>
<td>- Random obfuscation:</td>
<td>78%</td>
</tr>
<tr>
<td>- Simulated obfuscation:</td>
<td>12%</td>
</tr>
<tr>
<td><strong>Case Length Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>Short (3 – 5 sentences):</td>
<td>50%</td>
</tr>
<tr>
<td>Medium (6 – 8 sentences):</td>
<td>32%</td>
</tr>
<tr>
<td>Long (9 – 12 sentences):</td>
<td>18%</td>
</tr>
</tbody>
</table>
Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus

Data resources:
- Wikipedia Articles
- Persian-English Parallel Corpus
Bilingual Persian-English Corpus

- Clustering
Bilingual Persian-English Corpus

Clustering

Parallel Sentences Clustering

1. Persian Wikipedia documents were indexed by the Apache Lucene library.
2. We built a query from each Persian sentence.
3. The query was searched in the indexed documents and returns the top document.
4. A bipartite graph of return documents-categories was created. Then, the info-map community detection algorithm was applied to the graph and all communities were detected. Documents within a community are considered as one cluster.
5. Finally, parallel sentences were assigned to the documents in the same cluster.
Clustering

Parallel Sentences Clustering

1. Persian Wikipedia documents were indexed by the Apache Lucene library.
2. We built a query from each Persian sentence.
3. The query was searched in the indexed documents and returns the top document.
4. A bipartite graph of return documents-categories was created. Then, the info- map community detection algorithm was applied to the graph and all communities were detected. Documents within a community are considered as one cluster.
5. Finally, parallel sentences were assigned to the documents in the same cluster.

Documents Clustering

- For each cluster of return documents in the previous stage, the categories of documents have been extracted and considered as label of that cluster.
- The basic documents collected into different topically related clusters based on their categories. The documents are assigned to the cluster with maximum common categories.
Bilingual Persian-English Corpus

- Fragment Extraction
Bilingual Persian-English Corpus

- Fragment Extraction
  - Plagiarism cases are constructed from parallel sentences.
  - Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.
Bilingual Persian-English Corpus

- **Fragment Extraction**
  - Plagiarism cases are constructed from parallel sentences.
  - Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

<table>
<thead>
<tr>
<th>Fragment Length</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short</strong></td>
<td>3 – 5 sentences</td>
</tr>
<tr>
<td><strong>Medium</strong></td>
<td>5 – 10 sentences</td>
</tr>
<tr>
<td><strong>Long</strong></td>
<td>10 – 15 sentences</td>
</tr>
</tbody>
</table>
Bilingual Persian-English Corpus

- Fragment Extraction
  - Plagiarism cases are constructed from parallel sentences.
  - Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

-
Bilingual Persian-English Corpus

➢ Fragment Extraction
  o Plagiarism cases are constructed from parallel sentences.
  o Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

➢ Fragment Obfuscation
  o To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with different similarity score were chosen.
Bilingual Persian-English Corpus

- **Fragment Extraction**
  - Plagiarism cases are constructed from parallel sentences.
  - Source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

- **Fragment Obfuscation**
  - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with different similarity score were chosen.

<table>
<thead>
<tr>
<th>Degree</th>
<th>Similarity scores of sentences in fragments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 - 0.85</td>
</tr>
<tr>
<td>Low</td>
<td>100%</td>
</tr>
<tr>
<td>Medium</td>
<td>55% - 75%</td>
</tr>
<tr>
<td>High</td>
<td>35% - 55%</td>
</tr>
</tbody>
</table>
Bilingual Persian-English Corpus

- Inserting Plagiarism Cases into Documents
Bilingual Persian-English Corpus

- Inserting Plagiarism Cases into Documents
  - In this step, according to suspicious document’s length, one or more plagiarism cases are selected.
  - Persian documents considering as suspicious documents and source documents are English documents.
Bilingual Persian-English Corpus

- Inserting Plagiarism Cases into Documents
  - In this step, according to suspicious document’s length, one or more plagiarism cases are selected.
  - Persian documents considering as suspicious documents and source documents are English documents.

<table>
<thead>
<tr>
<th>Plagiarism per Document</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>5% - 20%</td>
</tr>
<tr>
<td>Medium</td>
<td>20% - 40%</td>
</tr>
<tr>
<td>High</td>
<td>40% - 60%</td>
</tr>
</tbody>
</table>
Bilingual Persian-English Corpus

- Inserting Plagiarism Cases into Documents
  - In this step, according to suspicious document’s length, one or more plagiarism cases are selected.
  - Persian documents considering as suspicious documents and source documents are English documents.
  - English fragment inserted at random positions in source documents and its corresponding Persian fragments has been inserted into suspicious documents.
  - Each suspicious document and its corresponding source documents are selected from one cluster.

<table>
<thead>
<tr>
<th>Plagiarism per Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
</tr>
<tr>
<td>Medium</td>
</tr>
<tr>
<td>High</td>
</tr>
</tbody>
</table>
### Results

<table>
<thead>
<tr>
<th>Documents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of source documents (English):</td>
<td>19973</td>
</tr>
<tr>
<td>The number of suspicious documents (Persian):</td>
<td></td>
</tr>
<tr>
<td>• With plagiarism:</td>
<td>3571</td>
</tr>
<tr>
<td>No plagiarism:</td>
<td>3571</td>
</tr>
</tbody>
</table>

| Plagiarism cases                              |            |
| The number of plagiarism cases:               | 11200      |

| Plagiarism per Document                        |            |
| The number of Little plagiarized documents     | 2035       |
| The number of Medium plagiarized documents     | 536        |
| The number of Much plagiarized documents       | 642        |
| The number of Very much plagiarized documents  | 58         |
Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection

Evaluation of Corpus Submissions to PAN 2015
Corpora Statistical Information
## Corpora Statistical Information

<table>
<thead>
<tr>
<th>Type of Corpus</th>
<th>cheema15</th>
<th>hanif15</th>
<th>Kong15</th>
<th>Alvi15</th>
<th>Palkovskii15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-Lingual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source—Suspicious Language</th>
<th>cheema15</th>
<th>hanif15</th>
<th>Kong15</th>
<th>Alvi15</th>
<th>Palkovskii15</th>
</tr>
</thead>
</table>


Corpora Statistical Information
## Corpora Statistical Information

<table>
<thead>
<tr>
<th></th>
<th>Cheema15</th>
<th>Hanif15</th>
<th>Kong15</th>
<th>Alvi15</th>
<th>Palkovskii15</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Docs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Suspicious Docs</td>
<td>248</td>
<td>250</td>
<td>4</td>
<td>90</td>
<td>1175</td>
</tr>
<tr>
<td>• Source Docs</td>
<td>248</td>
<td>250</td>
<td>78</td>
<td>70</td>
<td>1950</td>
</tr>
<tr>
<td><strong>Length of Docs (in chars)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Min Length</td>
<td>2263</td>
<td>361</td>
<td>394</td>
<td>514</td>
<td>519</td>
</tr>
<tr>
<td>• Max Length</td>
<td>22471</td>
<td>74083</td>
<td>121829</td>
<td>45222</td>
<td>517925</td>
</tr>
<tr>
<td>• Average Length</td>
<td>7239</td>
<td>4382</td>
<td>42839</td>
<td>7718</td>
<td>6512</td>
</tr>
<tr>
<td><strong>Length of Plagiarisms Cases (in chars)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Min Length</td>
<td>134</td>
<td>78</td>
<td>62</td>
<td>259</td>
<td>157</td>
</tr>
<tr>
<td>• Max Length</td>
<td>2439</td>
<td>849</td>
<td>2748</td>
<td>1160</td>
<td>14336</td>
</tr>
<tr>
<td>• Average Length</td>
<td>503</td>
<td>361</td>
<td>423</td>
<td>464</td>
<td>782</td>
</tr>
</tbody>
</table>
Corpora Statistical Information
## Corpora Statistical Information

<table>
<thead>
<tr>
<th>Obfuscation Strategies</th>
<th>Cheema15</th>
<th>Hanif15</th>
<th>Kong15</th>
<th>Alvi15</th>
<th>Palkovskii15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated</td>
<td>123</td>
<td>135</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Real</td>
<td>-</td>
<td>-</td>
<td>109</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Automatic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>-</td>
</tr>
<tr>
<td>Retelling-Human</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>-</td>
</tr>
<tr>
<td>Character-Substitution</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>-</td>
</tr>
<tr>
<td>Translation</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>618</td>
</tr>
<tr>
<td>Summary</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1292</td>
</tr>
<tr>
<td>Random</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>626</td>
</tr>
<tr>
<td>None</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>624</td>
</tr>
<tr>
<td>Sum</td>
<td>123</td>
<td>135</td>
<td>109</td>
<td>75</td>
<td>3160</td>
</tr>
</tbody>
</table>
Manual Evaluation of Corpora
Manual Evaluation of Corpora

- Manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus
Manual Evaluation of Corpora

- Manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus
- Changes in syntactic structure between source and plagiarized passage
Manual Evaluation of Corpora

- Manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus
  - Changes in syntactic structure between source and plagiarized passage
  - Concept preserving from source passage to plagiarized passage
Manual Evaluation of Corpora

- Manually investigate twenty pairs of corresponding source and suspicious fragments in each corpus
  - Changes in syntactic structure between source and plagiarized passage
  - Concept preserving from source passage to plagiarized passage
  - Distribution of obfuscation types in suspicious documents
Automatic Evaluation of Corpora
Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus
Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus
- For Kong15 corpus
Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus

- For Kong15 corpus
  - All source and correspond suspicious fragments are extracted, and the total number of similar “characters n-grams” between source and suspicious plagiarized passages are calculated for n in range of one to four.
Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus
- For Kong15 corpus
  - All source and correspond suspicious fragments are extracted, and the total number of similar “characters n-grams” between source and suspicious plagiarized passages are calculated for n in range of one to four.
- For evaluation of summary obfuscation
Automatic Evaluation of Corpora

- Evaluating two remained obfuscation scenarios:
  - Real obfuscation from Kong15 corpus
  - Summary obfuscation from Palkovskii15 corpus

- For Kong15 corpus
  - All source and correspond suspicious fragments are extracted, and the total number of similar “characters n-grams” between source and suspicious plagiarized passages are calculated for n in range of one to four.

- For evaluation of summary obfuscation
  - From the point of “concept preserving” measure, we have extracted 10% of top words from source fragments based on tf.idf weight.
Source Retrieval based on Noun and Keyword Phrase Extraction

Data resources:

External PD Corpus of PAN 2011
Approach in Use: Five Steps
Approach in Use: Five Steps

- Suspicious Document Chunking
Approach in Use: Five Steps

- Suspicious Document Chunking
- Noun Phrase and Keyword Phrase Extraction
Approach in Use: Five Steps

- Suspicious Document Chunking
- Noun Phrase and Keyword Phrase Extraction
- Query Formulation
Approach in Use: Five Steps

- Suspicious Document Chunking
- Noun Phrase and Keyword Phrase Extraction
- Query Formulation
- Search Control
Approach in Use: Five Steps

- Suspicious Document Chunking
- Noun Phrase and Keyword Phrase Extraction
- Query Formulation
- Search Control
- Document Filtering and Downloading
Suspicious Document Chunking
Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks
Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks
- No fixed pattern to put one plagiarism fragment per chunk
Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks
- No fixed pattern to put one plagiarism fragment per chunk
- Sufficient length of chunks, In order to comprise:
  1. At least one plagiarism fragment per chunk,
  2. And Maximum numbers of extracted queries from the chunks.
Suspicious Document Chunking

- Segmentation of suspicious documents into parts called chunks
- No fixed pattern to put one plagiarism fragment per chunk
- Sufficient length of chunks, In order to comprise:
  1. At least one plagiarism fragment per chunk,
  2. And Maximum numbers of extracted queries from the chunks.
- Individual sentences sets of 500 words Chunks as results.
Noun phrase and keyword phrase Extraction
# Noun phrase and keyword phrase Extraction

<table>
<thead>
<tr>
<th>Operation number</th>
<th>Operation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selection of top 80% long sentences (based on length in chars)</td>
</tr>
<tr>
<td>2</td>
<td>Selection of top 80% sentences (based on number of nouns)</td>
</tr>
<tr>
<td>3</td>
<td>Selection of top three sentences (based on average tf.idf1 values)</td>
</tr>
<tr>
<td>4</td>
<td>Selection of top three sentences (based on number of words with highest values)</td>
</tr>
</tbody>
</table>
Noun phrase and keyword phrase Extraction

<table>
<thead>
<tr>
<th>Operation number</th>
<th>Operation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selection of top 80% long sentences (based on length in chars)</td>
</tr>
<tr>
<td>2</td>
<td>Selection of top 80% sentences (based on number of nouns)</td>
</tr>
<tr>
<td>3</td>
<td>Selection of top three sentences (based on average tf.idf1 values)</td>
</tr>
<tr>
<td>4</td>
<td>Selection of top three sentences (based on number of words with highest values)</td>
</tr>
</tbody>
</table>

- Scenario 1: Operation 1 → Operation 2 → Operation 3 for noun phrase extraction
### Noun phrase and keyword phrase Extraction

<table>
<thead>
<tr>
<th>Operation number</th>
<th>Operation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selection of top 80% long sentences (based on length in chars)</td>
</tr>
<tr>
<td>2</td>
<td>Selection of top 80% sentences (based on number of nouns)</td>
</tr>
<tr>
<td>3</td>
<td>Selection of top three sentences (based on average tf.idf1 values)</td>
</tr>
<tr>
<td>4</td>
<td>Selection of top three sentences (based on number of words with highest values)</td>
</tr>
</tbody>
</table>

- Scenario1: Operation 1 → Operation 2 → Operation 3 for noun phrase extraction
- Scenario2: Operation 1 → Operation 2 → Operation 4 for keyword phrase extraction
Noun phrase and keyword phrase Extraction

<table>
<thead>
<tr>
<th>Operation number</th>
<th>Operation Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Selection of top 80% long sentences (based on length in chars)</td>
</tr>
<tr>
<td>2</td>
<td>Selection of top 80% sentences (based on number of nouns)</td>
</tr>
<tr>
<td>3</td>
<td>Selection of top three sentences (based on average tf.idf1 values)</td>
</tr>
<tr>
<td>4</td>
<td>Selection of top three sentences (based on number of words with highest values)</td>
</tr>
</tbody>
</table>

- Scenario1: Operation 1 → Operation 2 → Operation 3 for noun phrase extraction
- Scenario2: Operation 1 → Operation 2 → Operation 4 for keyword phrase extraction
- Three sentences from each scenario1 and scenario2 selected to query formulation
Query Formulation
Query Formulation

- From each selected sentence, one query is extracted.
Query Formulation

- From each selected sentence, one query is extracted.
- The threshold for the number of words in each query is limited to ten.
Query Formulation

- From each selected sentence, one query is extracted.
- The threshold for the number of words in each query is limited to ten.
- Selection of high weighted terms to reach the ChatNoir limitation.
Query Formulation

- From each selected sentence, one query is extracted.
- The threshold for the number of words in each query is limited to ten.
- Selection of high weighted terms to reach the ChatNoir limitation.
- The terms are placed next to each other based on the order in sentence.
Download Filtering and Search Control
Download Filtering and Search Control

- Download Filtering
  -
  -
  -
  -
  -
  -
Download Filtering and Search Control

- **Download Filtering**
  - 14 top results are selected for each query
Download Filtering and Search Control

- **Download Filtering**
  - 14 top results are selected for each query
  - The query is divided into two sub-queries:
    - Snippet with the length of 500 characters are extracted as a sub-query.
    - Snippets are combined with each other and make a passage.
Download Filtering and Search Control

- **Download Filtering**
  - 14 top results are selected for each query
  - The query is divided into two sub-queries:
    - Snippet with the length of 500 characters are extracted as a sub-query.
    - Snippets are combined with each other and make a passage.
  - If the resulted passage contains at least 50% words of the query
    - The related document is downloaded
    - The document is maintained for search control operation
Download Filtering and Search Control

- **Download Filtering**
  - 14 top results are selected for each query
  - The query is divided into two sub-queries:
    - Snippet with the length of 500 characters are extracted as a sub-query.
    - Snippets are combined with each other and make a passage.
  - If the resulted passage contains at least 50% words of the query
    - The related document is downloaded
    - The document is maintained for search control operation

- **Search Control**
  - Drop a query when at least 60% of its terms are contained in recently downloaded documents set
Search Control

- Drop a query when at least 60% of its terms are contained in recently downloaded documents set.
Evaluation
## Evaluation

<table>
<thead>
<tr>
<th>Downloads</th>
<th>F1</th>
<th>No Detection</th>
<th>Precision</th>
<th>Queries</th>
<th>Recall</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>183.3</td>
<td>0.115</td>
<td>1</td>
<td>0.07539</td>
<td>43.5</td>
<td>0.41381</td>
<td>8:32:37</td>
</tr>
</tbody>
</table>


Evaluation

- Highest rank in “No Detection” measure.
- Highest rank in “Runtime” measure.
Thank you for Your Attention