A Text Alignment Corpus for Persian Plagiarism Detection

Fatemeh Mashhadirajab, Mehrnoush Shamsfard
Razieh Adelkhah, Fatemeh Shafiee, Chakaveh Saedi

NLP Research Lab, Faculty of Computer Science and Engineering, Shahid Beheshti University
Outline

- Introduction
- Text Alignment Corpus Construction
- Strategies For Plagiarisms Types
- Dataset Statistics
- Conclusions
- References
A taxonomy of plagiarism [2]
Text Alignment Corpus Construction

- Data Source Preparation
- Documents Clustering
- Set of Suspicious and source Documents
- source and suspicious document pairs selection
- Source Documents Segmentation
- Segment Extraction
- Segment Obfuscation
- Obfuscated Segment Insertion
Text Alignment Corpus Construction

- **Data Source Preparation**
  - articles or theses in the fields of computer science and engineering & electrical engineering
    - 4,500 documents from Wikipedia articles
    - 1,500 documents from CSICC\(^1\) articles (2004-2015)
    - 1,500 documents from articles and theses available from online stores
    - 3,589 documents from free Persian resources including mag–iran\(^2\), iran–doc\(^3\), SID\(^4\), prozhe\(^5\), and MatlabSite\(^6\)

---

3. http://www.irandoc.ac.ir
6. www_MATLABSite.com
Text Alignment Corpus Construction

- Data Source Preparation
  - Our corpus contains 11,089 documents
Text Alignment Corpus Construction

- Documents Clustering
Text Alignment Corpus Construction

- Set of Suspicious and source Documents
Text Alignment Corpus Construction

Set of Suspicious and source Documents

They are randomly selected from each cluster

Suspicious Documents

Source Documents
Text Alignment Corpus Construction

- source and suspicious document pairs selection

<table>
<thead>
<tr>
<th>Suspicious Documents</th>
<th>Source Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Documents" /></td>
<td><img src="image2" alt="Documents" /></td>
</tr>
<tr>
<td><img src="image3" alt="Documents" /></td>
<td><img src="image4" alt="Documents" /></td>
</tr>
<tr>
<td><img src="image5" alt="Documents" /></td>
<td><img src="image6" alt="Documents" /></td>
</tr>
<tr>
<td><img src="image7" alt="Documents" /></td>
<td><img src="image8" alt="Documents" /></td>
</tr>
<tr>
<td><img src="image9" alt="Documents" /></td>
<td><img src="image10" alt="Documents" /></td>
</tr>
</tbody>
</table>
Text Alignment Corpus Construction

- source and suspicious document pairs selection

Suspicious Documents

Source Documents

 Similarity Detection system

if the similarity < 50%

NLP Research Lab, Faculty of Computer Science and Engineering, Shahid Beheshti University
Text Alignment Corpus Construction

- source and suspicious document pairs selection

Source Documents

Suspicious Documents

Similarity Detection system

if the similarity < 50%
Text Alignment Corpus Construction

- Source Documents Segmentation

Similarity Detection system
if the similarity < 50%

Source Documents
Suspicious Documents

NLP Research Lab, Faculty of Computer Science and Engineering, Shahid Beheshti University
Text Alignment Corpus Construction

- Segment Extraction

![Diagram showing text alignment corpus construction process](image.png)
Text Alignment Corpus Construction

- Segment Obfuscation

Suspicious Documents | Source Documents

Similarity Detection system

if the similarity < 50%

NLP Research Lab, Faculty of Computer Science and Engineering, Shahid Beheshti University
Text Alignment Corpus Construction

- Obfuscated Segment Insertion

Segment Obfuscation

Suspicious Documents

Source Documents

Similarity Detection system

if the similarity < 50%

NLP Research Lab, Faculty of Computer Science and Engineering, Shahid Beheshti University
Strategies For Plagiarisms Types

- Exact Copy
- Near Copy
- Modified Copy
- Text Manipulation (Paraphrasing)
- Text Manipulation (Summarizing)
- Automatic Translation
- Manual Translation
- Cyclic Translation
- Idea Adoption (semantic-based meaning)
Strategies For Plagiarisms Types

- Exact Copy

Segment Obfuscation

src

susp
Strategies For Plagiarisms Types

- Near Copy

Segment Obfuscation

- Insertion
- Deletion
- Substitution
- Sentence split or join

src  →  Segment Obfuscation  →  susp
Strategies For Plagiarisms Types

- Modified Copy

Segment Obfuscation

- src
- Susp

the Persian sentence understanding and generation system introduced by Adelkhah et al. [7]

semantic representation (sentence understanding) \[\rightarrow\] sentence production based on semantic representation (sentence generation)
Strategies For Plagiarisms Types

- Text Manipulation (Paraphrasing)

Segment Obfuscation

The Persian sentence understanding and generation system introduced by Adelkhah et al. [7]

Each word is replaced with a synonym retrieved from FarsNet or FavaNet
Strategies For Plagiarisms Types

- Text Manipulation (Summarizing)

Segment Obfuscation

Persian summarizer introduced by Shafiee et al. [6]
Strategies For Plagiarisms Types

- Automatic Translation

Segment Obfuscation

src

Google translate
Persian to English

Spell checker
Hunspell

susp

Persian

English
Strategies For Plagiarism Types

- Manual Translation

Segment Obfuscation

Translateion

src

Persian

→

Translateion

English

可疑
Strategies For Plagiarisms Types

- **Cyclic Translation**

Segment Obfuscation

```
src

English

Google translate

Negar spell checker

Hunspell

Google translate

susp

Persian
```
Strategies For Plagiarisms Types

- Idea Adoption (semantic-based meaning)

Segment Obfuscation

src

THE PARAPHRASING STRATEGY

susp
### Dataset Statistics

<table>
<thead>
<tr>
<th>documents</th>
<th>11089</th>
</tr>
</thead>
<tbody>
<tr>
<td>plagiarism cases</td>
<td>11603</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Document purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>languages</td>
</tr>
<tr>
<td>source documents</td>
</tr>
<tr>
<td>suspicious documents</td>
</tr>
<tr>
<td>with plagiarism</td>
</tr>
<tr>
<td>w/o plagiarism</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Document length</th>
</tr>
</thead>
<tbody>
<tr>
<td>short (&lt;10 pages)</td>
</tr>
<tr>
<td>medium (10-100 pages)</td>
</tr>
<tr>
<td>long (&gt;100 pages)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Plagiarism per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>hardly (&lt;20%)</td>
</tr>
<tr>
<td>medium (20%-50%)</td>
</tr>
<tr>
<td>much (50%-80%)</td>
</tr>
<tr>
<td>entirely (&gt;80%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case length</th>
</tr>
</thead>
<tbody>
<tr>
<td>short (&lt;1k characters)</td>
</tr>
<tr>
<td>medium (1k-3k characters)</td>
</tr>
<tr>
<td>long (&gt;3k characters)</td>
</tr>
</tbody>
</table>

### Obfuscation synthesis approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Copy</td>
<td>8%</td>
</tr>
<tr>
<td>Near Copy</td>
<td>12%</td>
</tr>
<tr>
<td>Modified Copy</td>
<td>12%</td>
</tr>
<tr>
<td>Paraphrasing</td>
<td>16%</td>
</tr>
<tr>
<td>Summary</td>
<td>13%</td>
</tr>
<tr>
<td>Manual Translation</td>
<td>18%</td>
</tr>
<tr>
<td>Automatic Translation</td>
<td>8%</td>
</tr>
<tr>
<td>Cyclic Translation</td>
<td>10%</td>
</tr>
<tr>
<td>semantic-based meaning</td>
<td>3%</td>
</tr>
</tbody>
</table>

1 A page is measured as 1500 chars.
Conclusion

✓ This article describes a methodology for building a Persian corpus for evaluating plagiarism detection systems.

✓ This corpus is in PAN format.

✓ This corpus is a variety of plagiarism types in large volume are created.

✓ To produce this corpus, the focus is on the simulation of different types of plagiarism.

✓ Different strategies are employed to create obfuscation in each plagiarism category.
References


NLP Research Lab, Faculty of Computer Science and Engineering, Shahid Beheshti University


References
References


Thanks for your attention.