Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus

Habibollah Asghari , Khadijeh Khoshnava , Omid Fatemi , Heshaam Faili

ICT Research Institute

Academic Center for Education, Culture and Reseach (ACECR), Iran

Text Alignment Corpus Construction

Plagiarism Corpus Construction Approaches:

- **Collection:** Find Real-World instances of text reuse or plagiarism, and annotate them.
- **Generation:** Given pairs of documents, generate passages of reused or PLAGIARIZED TEXT between them. Apply a means of obfuscation of your choosing.
  - Simulated
  - Artificial

Our Approach

- **Preprocessing**
  - Unification of letters to Unicode characters designed for Persian and using zero-width non-joiner space
- **Clustering**

  We have proposed our approach for clustered parallel sentences and Wikipedia documents into different topically related groups.
  - **Parallel Sentence Clustering**
    - the clustering procedure of parallel sentences is accomplished to detect the presence of distinct groups and assign parallel sentences to groups.
    - Since the parallel corpus we have used, has been extracted from Wikipedia, so we used the structure of the wiki pages for clustering of sentences.
  - **Document Clustering**
    - For clustering of documents we used the results of parallel sentences clustering stage.
- **Building Plagiarism cases**

  For constructing a plagiarism case, we put together some of the sentences of parallel corpus.
  - The source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

Fragment lengths in sentences.

Type	Length (Sentence)	Ratio (%)
Short	3-5	35
Medium	5-10	38
Long	10-15	27

- **Fragments Obfuscation**
  - Plagiarized fragments have been constructed from Persian sentences and corresponding source fragments have been constructed from English sentences parallel with source sentences.
  - To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with different similarity score were chosen.

Degree of obfuscation in plagiarism cases

Degree	Similarity scores of sentences in fragments		
	0.45-0.65	0.65-0.85	0.85-1
Low	-	-	100 %
Medium	-	25-45 %	55-75 %
High	45-65 %	-	35-55 %

- **Insert Plagiarism Cases in Suspicious Documents**

  In this step, according to suspicious documentâs length, one or more plagiarism cases which are in the same cluster of suspicious documents are selected. For each plagiarized cases, the suspicious and corresponding source fragments inserted at random positions in suspicious and source documents, respectively.

Ratio of Plagiarism fragments in Documents.

Type	Percentage (%)
Little	5-20
Medium	20-40
High	40-60

Data Source Preparation

We have used Wikipedia documents for constructing the main body of source and suspicious documents. Moreover, we exploiteda parallel Persian- English sentence aligned corpus to construct the plagiarized passages.

- **Wikipedia:** We have crawled Persian Wikipedia documents in accordance with corresponding pages in English language. In the process of crawling, we have considered and extracted TITLE, URL , TEXT AND CATEGORIES FIELD OF THE PAGES

WIKIPEDIA
The Free Encyclopedia

- **Persian - English Parallel Corpus:**Using a parallel English-Persian sentence aligned corpus to constructing paired plagiarism passages. A collection of 12 features were used into a Maximum Entropy (MaxEnt) log linear model in order to compute the similarity scores between paired sentences.

  The features are in four categories including features based on sentence length, related to dictionary (IBM model 1), based on alignment and miscellaneous features.

Results

- In this section, the result and statistics of the corpus is presented.

Documents

The number of source documents(English):	19973
The number of suspicious documents(Persian):	
- With plagiarism:	3571
- No plagiarism:	3571

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:	358

Conclusion

- We have discussed our approach to the task of text alignment in the context of PAN 2015 competition.
- This corpus is intended to be used to evaluate the performance of bilingual plagiarism detection systems.
- Our main contribution is to use a novel obfuscation strategy by using the similarity scores between parallel sentences in such a way that the obfuscation degree can be adjusted in plagiarized passages.
- This corpus is the first bilingual plagiarism corpus for Persian language.
- In the future works, we plan to improve our corpus by incorporating other obfuscation strategies such as manual obfuscation and artificial obfuscation in the corpus.

THIS WORK HAS BEEN ACCOMPLISHED UNDER THE SUPPORT OF VICE PRESIDENCY FOR SCIENCE AND TECHNOLOGY OF IRAN GRANT NO. 1164331

{habib.asghari, khadijeh.khoshnava}@ictrc.ir