Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus

Salar Mohtaj, Habibollah Asghari, Vahid Zarrabi

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

▶ Documents Clustering

- The documents which are used in the corpus are derived from the Wikipedia Internet encyclopedia project
- Since pages on similar subjects are intended to be grouped together via categories, a bipartite graph of documents-categories has been created to cluster the documents based on their topics
- To detect communities of the graph, the Infomap community detection algorithm has been applied to the graph

▶ Fragments Extraction

We have used two different methods for fragment extraction.

- **Extracting fragments from the source documents:** 50% of the documents are considered as source and 50% are designated as suspicious documents. Note that only 25% of suspicious documents contain plagiarism cases
- **Extracting fragments from the SemEval dataset**

Fragment lengths in words.

Type	Length (Sentence)	Ratio (%)
Short	3-5	50
Medium	6-8	32
Long	9-12	18

▶ Fragments Obfuscation

We have proposed two obfuscation strategies for obfuscation of fragments:

- **Artificial Obfuscation**
  - . None (No Obfuscation), Random Change of Order , POS-preserving Change of Order , Synonym Substitution , Addition / Deletion
- **Simulated Obfuscation**
  - . The pairs of sentences from the dataset of semantic textual similarity task in **SemEval** 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

Obfuscation degree in simulated plagiarism cases

Degree	Similarity		
	3	4	5
Low	-	1 -15 %	85-100 %
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. Then, each of them inserted at random positions in suspicious document. For simulated plagiarism cases, the corresponding source fragments also inserted at random positions in source documents

Ratio of Plagiarism fragments in Documents.

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

SemEval Dataset

- **Semantic Textual Similarity for English in SemEval:** In this task, participants were wanted to develope a system to measure semantic textual similarity between two sentences in English.
- This data covers 5 datasets: paraphrase sentence pairs (MSRpar), sentence pairs from video descriptions (MSRvid), MT evaluation sentence pairs (MTnews and MTeuroparl) and gloss pairs (OnWN).
- The similarity score can range from exact semantic equivalence to complete unrelatedness, corresponding to quantified values between five and zero.
- For Developing the Plagiarism detection corpus the train and test dataset of **STS** task have been used as Simulated plagiarized cases.
- In order to create the cases of plagiarism, we ignore unrelatedness sentences with a similarity degree lower than 3.
- Source fragments constructed by original sentences and corresponding plagiarized fragments are created by corresponding sentences of original ones in the dataset.

Results

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

Documents

The number of source documents:	3309
The number of suspicious documents:	952

Plagiarism cases

The number of plagiarism cases:	
- No obfuscation cases:	10%
- Artificial Obfuscation	
- Low Obfuscation	42%
- High Obfuscation	36%
- Simulated Obfuscation	12%

Plagiarism per Document

The number of Little plagiarized documents:	60%
The number of Medium plagiarized documents:	25%
The number of High plagiarized documents:	15%

Conclusion

- We have discussed our approach to the task of text alignment in the context of PAN 2015 competition.
- Two different approaches have been used for creating the corpus, namely simulated and artificial obfuscation strategies.
- The SemEval dataset is used for generating the simulated plagiarized cases.
- The degree of obfuscation in simulated plagiarism cases is based on similarity scores of pair ed sentences.
- This corpus is intended to be used for testing the performance of plagiarism detection systems for English language.
- Although this corpus is in English text, the obfuscation strategy can also be exploited in other languages.
- In our future work, we plan to improve our corpus by implementing other obfuscation techniques.

This work has been accomplished under the support of Vice Presidency for Science and Technology of Iran grant No. 1164331

{salar.mohtaj, habib.asghari, vahid.zarrabi}@ictrc.ir