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

## Notebook for PAN at CLEF 2015

Salar Mohtaj, Habibollah Asghari, Vahid Zarrabi

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

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

**Abstract.** In this paper, we describe an approach to create monolingual English plagiarism detection corpus for the task of text alignment corpus construction in PAN 2015 competition. We propose two different obfuscation methods to fragment obfuscation for creating the cases of plagiarism. The first method is an artificial obfuscation which consists of variety of obfuscation strategies such as synonym substitution, random change of order, POS preserving change of order and addition/deletion. The second obfuscation method is a simulated obfuscation, in which the SemEval dataset is used for creating the cases of plagiarism by using pairs of sentences with their similarity scores.

**Keywords:** Plagiarism Detection, Corpus Construction, Monolingual English Corpus, Human Annotated Paraphrase Corpus

## 1 Introduction

Plagiarism is defined as re-use of another person's ideas, processes, results, or words without explicitly acknowledging the original source [1]. Plagiarism detection algorithms try to search in the large document collections for the retrieval and extraction the patterns of text reuse [2]. Plagiarism detection systems are one of the tools have been using to fight plagiarism and malpractice use of others text [3]. In developing plagiarism detection systems, a plagiarism detection corpus is used for evaluation of the system. It consists of predefined tagged plagiarized materials.

The plagiarism detection task has been running for seven years in PAN competition and each year, it provides a corpus for evaluating of submitted systems. The evaluation corpora in PAN are used for text alignment and source retrieval task for plagiarism detection [4]. Variety of obfuscation strategies have been used to create

text alignment corpora such as artificial obfuscation, simulated obfuscation, translation and summary obfuscation [2, 4, 5, 6, 7].

In this lab report, we have described our approach to generate a monolingual English corpus for the task of text alignment corpus construction. We employ two obfuscation strategies, artificial and simulated ones. Our main contribution is using the SemEval dataset for constructing simulated plagiarism cases. The similarity score of paired sentences in SemEval dataset have been used for establishing the degree of obfuscation for plagiarism cases.

In the following, in section 2 we describe our approach for corpus construction. Then in section 3 we will discuss the statistics of the resulted corpus which is based on Wikipedia articles. Finally, we will conclude and discuss about some future works in section 4.

## 2 Our Approach

In this section, an overview of our approach for constructing a monolingual English plagiarism detection corpus is presented. Our approach includes four main steps: document clustering, fragment extraction, fragment obfuscation and inserting plagiarism cases into the source and suspicious documents. The process of each step is described in the following sections.

### 2.1 Documents Clustering

The documents which are used in the corpus are derived from the Wikipedia Internet encyclopedia project. In this step, the collection of Wikipedia documents is clustered into different topically related categories. 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 [9] has been applied to the graph. Finally, documents within a community are considered as similar documents in one cluster. Each suspicious document and its corresponding source documents are selected from the same cluster.

### 2.2 Fragments Extraction

The documents used in the corpus are divided into two categories: 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.

We have used two different methods for fragment extraction. In the first method, the fragments are extracted from the source documents, while in the second method, the SemEval datasets is used for fragment extraction. The length of fragments is evenly distributed between 3 and 12 sentences. The distribution of fragments' length is shown in Table 1.

**Table 1.** Fragment lengths in sentences

| Fragment Length | |
|---|---|
| Short | 3 – 5 sentences |
| Medium | 6 – 8 sentences |
| Long | 9 – 12 sentences |

### 2.3 Fragments Obfuscation

We have proposed two obfuscation strategies for obfuscation of fragments: Artificial obfuscation and simulated obfuscation. In the following, we described our obfuscation strategies.

**Artificial Obfuscation.** For the purpose of generating artificial plagiarism, obfuscation strategies were applied to fragments extracted from source documents. We have used five obfuscation strategies as follows:

- None (No Obfuscation)

Source fragment without any change considered as the obfuscation fragment. In other words, the obfuscation fragment is an exact copy of source fragment.

- Random Change of Order

Given source fragment, the obfuscation fragment is created by shuffling words at random.

- POS-preserving Change of Order

In order to accomplish this obfuscation strategy, the sequence of parts of speech (POS) tags in source fragment is determined. Then, words are shuffling randomly, while retaining the original POS sequence.

- Synonym Substitution

The plagiarized fragment is created in such a way to replace some words by one of their synonyms.

- Addition / Deletion

The obfuscated fragment is created by inserting or removing words at random.

**Simulated Obfuscation.** The pairs of sentences from the dataset of semantic textual similarity task in SemEval are used for constructing the simulated plagiarism cases. The dataset includes pairs of semantically similar sentences with their corresponding similarity score. The similarity score can range from exact semantic equivalence to complete unrelatedness, corresponding to quantified values between five and zero [8]. In order to create the cases of plagiarism, we ignore unrelatedness sentences with a similarity degree lower than 3.

In this strategy, both source and plagiarized fragments are constructed by SemEval dataset sentences. Source fragments constructed by original sentences and corresponding plagiarized fragments are created by corresponding sentences of original ones in the dataset.

To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with a variety of similarity scores is used in a fragment. The number of sentences and their similarity scores specifies the degree of obfuscation for each plagiarized fragment. More precisely, using sentences with higher degree of similarity (e.g. 5) could lead to plagiarized fragments with lower degree of obfuscation and vice versa. The distribution of different sentences for creating different degrees of obfuscation (namely "Low", "Medium", and "High" obfuscation) is shown in Table 2.

**Table 2.** Obfuscation degree in simulated plagiarism cases

| **Degree** | **Similarity Scores of Sentences** | | |
|---|---|---|---|
| | **3** | **4** | **5** |
| Low | - | 1% -15% | 85% - 100% |
| Medium | 25% - 45% | | 55%- 75% |
| High | 45% - 65% | | 35% - 55% |

### 2.4 Inserting Plagiarism Cases into Suspicious Documents

In this step, one or more plagiarism cases according to the suspicious document's length, within the same cluster have been selected. Then, each of them inserted at random positions in suspicious documents. For simulated plagiarism cases, the corresponding source fragments also inserted at random positions in source documents.

The fraction of plagiarism in each document is not fixed. The percentage of plagiarism in each suspicious document is distributed between 5% and 60% of its length. The ratio of plagiarism per suspicious documents is shown in Table 3.

**Table 3.** Ratio of Plagiarism fragments in Documents

| **Plagiarism per Document** | |
|---|---|
| Hardly | 5% - 20% |
| Medium | 20% - 40% |
| Much | 40% - 60% |

Finally, for each pair of source and suspicious documents, a Metadata file is created which contains meta information about the plagiarism cases. The tags in the file include:

*this_length*: The length of plagiarism case in the suspicious document.
*this_offset*: Start offset of the plagiarism case in the suspicious document.
*source_reference*: Name of source document.
*source_length*: The length of source fragment in the source document.
*source_offset*: Start offset of the source fragment in the source document.

## 3 Results

In this section, the results and statistics of monolingual English corpus for the PAN 2015 competition is presented. This corpus is based on Wikipedia documents. The results of corpus construction are shown in Table 4.

**Table 4.** Statistics of Human Annotated Paraphrase Corpus

| ***Document Statistics*** | |
|---|---|
| ***Document Purpose*** | |
| The number of source documents: | 3309 |
| The number of suspicious documents: | 952 |
| ***Plagiarism per Document*** | |
| Hardly (5% - 20%) | 60% |
| Medium (20% - 40%) | 25% |
| Much (40% - 60%) | 15% |
| ***Plagiarism Case Statistics*** | |
| ***Plagiarism cases*** | |
| The number of plagiarism cases: | |
| - No obfuscation cases: | 10% |
| - With obfuscation cases: | |
| - Random obfuscation: | 78% |
| - Simulated obfuscation: | 12% |
| ***Case Length*** | |
| Short (3 – 5 sentences): | 50% |
| Medium (6 – 8 sentences): | 32% |
| Long (9 – 12 sentences): | 18% |

The established English mono-lingual plagiarism detection corpus is available at the website[1] of "Research Institute for Information and Communication Technology" for research purposes.

## 4 Conclusion and Future Work

In this lab report, we described our approach for constructing a monolingual plagiarism detection corpus. We have used two obfuscation strategies to create our corpus. The first is artificial obfuscation strategy in which the plagiarized fragments are automatically created. In the second strategy, named simulated obfuscation, either source or plagiarized fragments were created by SemEval dataset. The degree of obfuscation in simulated plagiarism cases is based on similarity scores of paired 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.

## Acknowledgement

This work has been accomplished in ICT research Institute, ACECR, under the support of Vice Presidency for Science and Technology of Iran - grant No. 1164331. The authors gratefully acknowledge the support of aforementioned organizations. Special thanks go to the members of ITBM research group for their valuable collaboration.

## References

1. Barrón-Cedeño, Alberto, Marta Vila, M. Antònia Martí, and Paolo Rosso. "Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection." Computational Linguistics 39, no. 4 (2013): 917-947.

2. Potthast, Martin, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein "Overview of the 6th International Competition on Plagiarism Detection." CLEF (Online Working Notes/Labs/Workshop). 2014.

3. Juričić, Vedran, Vanja Štefanec, and Siniša Bosanac. "Multilingual plagiarism detection corpus." In MIPRO, 2012 Proceedings of the 35th International Convention, pp. 1310-1314. IEEE, 2012.

4. Potthast, Martin, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. "Overview of the 5th international compe-

[1] http://www.ictrc.ir/plaglab/corpora/MonoLingual_English_Corpus(mohtaj15).zip

tition on plagiarism detection." In CLEF Conference on Multilingual and Multimodal Information Access Evaluation, pp. 301-331. CELCT, 2013.

5. Potthast, Martin, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. "Overview of the 2nd International Competition on Plagiarism Detection." In CLEF (Notebook Papers/LABs/Workshops). 2010.

6. Potthast, Martin, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. "Overview of the 3rd International Competition on Plagiarism Detection." In *CLEF (Notebook Papers/LABs/Workshops)*. 2011.

7. Potthast, Martin, Tim Gollub, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. Overview of the 4th International Competition on Plagiarism Detection. In Working Notes Papers of the CLEF 2012 Evaluation Labs, September 2012. ISBN 978-88-904810-3-1. ISSN 2038-4963.

8. Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. "sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity." In In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics. 2013.

9. Rosvall, Martin, and Carl T. Bergstrom. "Maps of random walks on complex networks reveal community structure." Proceedings of the National Academy of Sciences 105, no. 4 (2008): 1118-1123.