RDI System for Extrinsic Plagiarism Detection (RDI_RED)
Working Notes for PAN-AraPlagDet at FIRE 2015

Ahmed Magooda
Computer Engineering Department,
Cairo University
ahmed.ezzat.gawad@gmail.com

Ashraf Y. Mahgoub
Computer Engineering Department,
Cairo University
ashraf.youssef.mahgoub@gmail.com

Mohsen Rashwan
Communication department,
Cairo University
rashwan@rdi-eg.com

Magda B.Fayek
Computer Engineering Department,
Cairo University
magdafayek@ieee.org

Hazem Raafat
Computer Science Department,
Kuwait University
hazem@cs.ku.edu.kw

ABSTRACT
Extrinsic plagiarism detection gathered the attention of many researchers lately. Plagiarism process began to be more and more difficult to be detected due to appearance of other sophisticated plagiarism approaches other than direct copy and paste such as (phrase rephrasing, word shuffling, semantic substitution, etc…). In this paper, we present RDI system for extrinsic plagiarism detection (RDI_RED). RDI_RED system performs remarkably on a wide spectrum of plagiarism techniques starting from simple copy-paste to word shuffling and also complete sentence rephrasing. RDI_RED system achieved the first three positions in Arabic language plagiarism detection competition with a Plagdet (Plagiarism Detection score) of 80% which is 20% higher than the base line and 18% higher than the second best competing system.

Keywords
Extrinsic Plagiarism Detection; Rephrasing; Semantic similarity; Natural Language Processing; Indexing; Chunking; Seeding; Text Alignment.

1. INTRODUCTION
Plagiarism detection is a very interesting task as it is in its core a competition between machines and humans. The essence of plagiarism detection is like reverse engineering human behavior and nullifying all the effort exerted in the process of modifying the plagiarized text. Plagiarism detection can be clustered as two main tracks:(1) Intrinsic plagiarism detection and (2) Extrinsic plagiarism detection.

Intrinsic plagiarism detection is the process of verifying the unity of a document against itself without the need of any external sources. This process is concerned with finding whether the document is written by the same author or there exists some parts that at high probability are not written by the same author.

On the other hand extrinsic plagiarism detection is the process of evaluating a document and verifying if there exists some parts that have been copied from external sources, this process is held with the presence of external source which called (source documents) these documents are treated like probable source of copying.

The system proposed in this paper deals with the later task (extrinsic track). Extrinsic plagiarism detecting is a well-known task that witnessed a lot of work during the last decades. However, the basic drive for developing RDI_RED system is the absence of any reliable system that works on Arabic language.

Lack of Arabic language resources shifted most of the work towards English language based system development only, however we found a good opportunity in this competition to invest our Arabic language knowledge in developing our own system and monitor its performance against other teams with the presence of unified test corpus and benchmark.

2. METHOD
The proposed RDI_RED system consists of three basic modules, (1) Candidate source documents retrieval module, (2) Alignment module and (3) Filtering module. In the following (Param1, Param2, Param3,Param4, Param5 and Param6) are parameters that we vary during the training process.

First we will explain the **retrieval module** in detail. The retrieval module depends on two approaches for candidate source document retrieval:

1.Paragraph based retrieval: In this approach,
(a) The suspicious document is chunked into paragraphs. For each paragraph, named entities are extracted from the document using RDI_NER [5] and Arabic Wikipedia [1] dump module proposed in Mahgoub et al. [2].
(b) Inverse document frequency (IDF) weights are calculated for each term.
(c) Two queries are constructed for each of the resulting paragraphs as follows.
  (i) The first query is constructed by extracting (Param1)words from the current paragraph. These (Param1) words are the (Param1)/2 highest IDF named entities and (Param1)/2 highest IDF words that are not named entities.
  (ii) The second query is constructed by extracting the 10-grams that contains the maximum number of specified

terms, these terms are the terms which were extracted while constructing the first query.

(d) The resulting queries are then issued to a search engine to retrieve set of candidate source documents. Each query is stemmed using RDI_Stemmer [5] and light10 stemmer proposed by Leah S. Larkey et al. [3]. The used search engine uses a paragraph based index that has been constructed using LUCENE search tool [4].

2. Sentence based retrieval: In this approach,
   (a) Source documents are chunked into sentences.
   (b) For each sentence an ID is constructed.
   (c) After that the constructed IDs alongside the chunked sentences are fed into an inverted index using LUCENE.
   (d) For each suspicious document the same procedure is applied and each resulting sentence is treated as a query. Each query is issued to LUCENE resulting in a candidate source document.
   (e) All the retrieved documents for the suspicious document in concern are sorted by the number of queries they were retrieved by (one source document for each query).

The candidate source documents retrieved by the two approaches, are then passed to the **alignment module**. The alignment module presented in this system is based upon three different alignment approaches (1) Skip-gram based approach, (2) Sentence index based approach and (3) Common words based approach.

1. **Skip-gram based approach**: This approach proceeds as follows.
   (a) A suspicious document is scanned by a window of five words with a one word step.
   (b) The five words extracted by the window are stemmed using RDI_Stemmer [5],then all combinations of triple words are extracted out of the five words (Skip-gram).
   (c) The same approach is applied over the retrieved source documents.
   (d) For each of the suspicious document skip-grams, the skip-gram is compared to all of the source documents skip-grams and the matched skip-grams are saved.
   (e) The system then apply an expansion step. In this step for all the matched skip-grams we group consecutive skip-grams that are separated with no more than (Param2) number of characters (either in suspicious document or source document) together.

2. **Sentence index based approach**: This alignment approach depends mainly on the sentence based retrieval approach. Introduced earlier, each sentence of the suspicious document is used as a query where only the first source document match from the index is considered for alignment. For each candidate source document, the following steps are applied:
   (a) The sentences constructed from the suspicious document alongside the matched sentences retrieved added to a list of (suspicious sentence ID – source sentence ID) pairs.
   (b) For each (suspicious sentence ID – source sentence ID) pair
   (i) If there exists pair that resides within a window of length (Param3) from another pair then mark this pair as a valid matching pair.
   (ii) If not, pass the text of both sentences for next module (filtering module)

3. **Common words based approach**: This alignment approach depends on the density of common words between a suspicious document and a candidate source document pair in order to detect the plagiarized parts between them. For each (suspicious-candidate source) pair, the following steps are applied:
   (a) Get list of all common words.
   (b) For each matching word, add list of its indices (positions of the word in both document) into a matching indices list.
   (c) Using a window of words of length (Param4), if the gap between two consecutive matches is wider than (Param4) then we split the list into two separate lists.
   (d) Pass each set of extracted consecutive words from suspicious document with its corresponding set in source list (which has the maximum ratio of common words to their average length) to the next filtering module.

After the retrieval and alignment modules comes the filtering module. **The filtering module** applies set of rules and give a final decision to accept or reject the aligned part. For each of the aligned parts the following rules are applied:

(a) If the source and suspicious chunks are equivalent accept this part, else go to next rule.
(b) If the length of any of the two chunks is shorter than (param5) of characters then this part is rejected, else continue.
(c) If the number of common words is greater than (param6) then accept this part, else reject this part.

3. EVALUATION

In the training phase, the RDI_RED system are trained and tuned with different set of configurations resulting into three different training runs. Each configuration will described alongside its training and testing performance results:

1. The first run uses the following configurations:
   (a) Sentence based retrieval.
   (b) Sentence index based alignment approach, and Common words Alignment approach.
   (c) Parameters tuned: Param3, Param4, Param5 and Param6.

Table 1. First run's performance over training and testing datasets

	PlagDet	Precision	Recall	Granularity
Train	0.74	0.70	0.82	1.02
Test	0.77	0.80	0.79	1.05

2. The second run uses the following configurations:
   (a) Paragraph based indexing and retrieval module, and Sentence based retrieval.
   (b) Skip-gram alignment approach, Sentence index based alignment approach, and Common words Alignment approach.
   (c) Parameters tuned: Param1, Param2,Param3, Param4, Param5 and Param6.

Table 2. Second run's performance over training and testing datasets

	PlagDet	Precision	Recall	Granularity
Train	0.88	0.85	0.94	1.02
Test	0.80	0.85	0.83	1.07

3. The third run uses the following configurations:
   (a) Paragraph based indexing and retrieval module, and Sentence based retrieval.
   (b) Skip-gram alignment approach, and Sentence index based alignment approach.
   (c) Parameters tuned: Param1, Param2,Param3, Param4, Param5 and Param6.

Table 3. Third run's performance over training and testing datasets

	PlagDet	Precision	Recall	Granularity
Train	0.87	0.86	0.91	1.01
Test	0.77	0.85	0.76	1.06

The following table summarizes the final results of Arabic plagiarism competition for the year 2015 [7].

Table 4. Results of Arabic plagiarism competition for year 2015

Rank	1	2	3	4	
Method	Magooda_2	Magooda_3	Magooda_1	Palkovskii	Baseline
Macro precision	0.85	0.85	0.80	0.97	0.99
Macro recall	0.83	0.76	0.79	0.54	0.53
Micro precision	0.92	0.92	0.88	0.99	0.99
Micro recall	0.84	0.78	0.81	0.58	0.59
Granularity	1.07	1.06	1.05	1.16	1.20
Plagdet	0.80	0.77	0.77	0.62	0.60

4. TECHNICAL DETAILS

The systems evaluation carried out over training and test data was performed on a personal machine with plausible specifications, the following specifications are the specifications used during the whole system evaluation process:

- Hardware Specifications:
  - CPU: Intel coreI7 4500U - 2 Cores – 1.8 ~ 3.0 GHz
  - RAM: 16 GB of RAM
- Software Specifications:
  - Operating System: Windows 7 x64
  - Development Environment: Visual Studio 2013
  - Programming Language: .Net C#

The RDI_RED system was trained using the training data provided by the competition to get the best set of parameters for (param1, param2, param3, param4, param5 and param6):
The training time for the previously illustrated approaches are:

Table 5. Training and Testing Time in seconds

	Train	Test
First Run	160500	157666
Second Run	161190	158400
Third Run	150390	147600

Note: The previously reported training running times are per iteration not the whole process of tuning.

5. CONCLUSION

In this paper, the RDI_RED system was introduced for extrinsic plagiarism detection task. The RDI_RED system depends on two different retrieval approaches using LUCENE and three different alignment approaches. Three different configurations are tested and tuned over the provided training dataset. Best results have been achieved by combining more than one alignment approach rather than using each approach as a standalone technique. The combined approach achieved very promising results for Arabic language despite the lack of resources. Despite being a semi-language-independent system RDI_RED achieved comparable results to state of the art English language systems reported in PAN-2014 [6].

6. REFERENCES

[1] https://ar.wikipedia.org/

[2] Mahgoub, Y., A., Rashwan, A. M., Raafat, H., Zahran, A., M. and Fayek, B., M.: *Semantic Query Expansion for Arabic Information Retrieval.* In: EMNLP: The Arabic Natural Language Processing Workshop, Conference on Empirical Methods in Natural Language Processing, Doha, Qatar (2014) 87-92.

[3] Larkey, Leah S., Lisa Ballesteros, and Margaret E. Connell. *Light stemming for Arabic information retrieval*. Arabic computational morphology. Springer Netherlands, 2007. 221-243.

[4] https://lucene.apache.org/

[5] http://www.rdi-eg.com/index.htm

[6] Potthast, Martin, et al. *Overview of the 6th International Competition on Plagiarism Detection*. CLEF Conference on Multilingual and Multimodal Information Access Evaluation. ceur-ws, 2014.

[7] http://misc-umc.org/AraPlagDet