RDI System for Intrinsic Plagiarism Detection (RDI_RID)
Working Notes for PAN-AraPlagDet at FIRE 2015

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

Ahmed Magooda
Computer Engineering Department,
Cairo University
ahmed.ezzat.gawad@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
Many researchers have been investigating the task of plagiarism detection lately. In this paper we present RDI system for intrinsic plagiarism detection (RDI_RID). RDI_RID system was the only system that participated in intrinsic track of the Arabic language plagiarism detection competition. RDI_RID system achieved a PlagDet (Plagiarism Detection score) of 19% compared to 38% achieved by the base line system. The proposed system is based on vector representation of stylometric features extracted from document's text.

Keywords
Intrinsic Plagiarism Detection; Stylometry; Natural Language Processing; Part of Speech.

1. INTRODUCTION
Due to major advances in plagiarism techniques, plagiarized documents have become too difficult and sophisticated to be detected by traditional plagiarism detection methodologies. Therefore, efficient plagiarism detection techniques are needed to detect intelligently manipulated text.

Plagiarism detection systems are special types of Information Retrieval (IR) systems as their task is not limited to recognize relevant documents only, they also analyze suspicious documents and detect chunks of text which are plagiarized from another source. Therefore, plagiarism detection can be viewed as a process of reverse engineering applied on suspicious document to reformulate it to its origins.

There exists two types of plagiarism detection systems. The first type uses stylometric features extracted from the suspicions document itself in order to highlight chunks of text that do not align with the rest of the document and hence are more probable to be brought from another document. This type of systems is called intrinsic plagiarism detection systems.

The second type is called extrinsic plagiarism detection systems. In these systems, the suspicious document is tested against a set of external sources in order to detect which parts have been plagiarized from which source document.

The system proposed in this paper deals with the first track (intrinsic track). Intrinsic plagiarism detection is similar to the problem of authorship detection. To our knowledge there is no reliable system that detects the plagiarism intrinsically on Arabic documents.

While a considerable portion of research in this field was targeting the English Language, the well-constructed data sets in addition to previous Arabic processing knowledge gained in related tasks were the main motivations for participating in this competition.

2. METHOD
The proposed RDI_RID system consists of three basic modules, (1) Document chunking module (2) Vector representation module(3) Filtering module.

1. **Chunking module**. This module divides the document text using a sliding window with the following configuration:
   (a) The Window should contain (param1) alphanumeric characters (no punctuation or diacritics).To do so, the window can be allowed to expand up to 1.5*(param1) characters to guarantee the existence of param1 alphanumeric characters within the window or the window reaches the maximum size of 1.5*(param1).
   (b) The window slides by (param1)/2 characters at a time.
2. After the chunking process, the chunks **conversion to vector** is carried out. The following features are extracted to represent the chunk as a vector according to the method proposed by Zechner et al.[1]. For each chunk the following is estimated:
   (a) Stop words frequency: These include all Arabic stop words such as (... إلى, على, من) each stop word has a specific dimension in the final vector; each dimension is set with the corresponding frequency.
   (b) Punctuation frequency: These include all Arabic punctuations such as (“" : . ,”). Each punctuation type has a specific dimension in the final vector; each dimension is set with the corresponding frequency.
   (c) Part of speech frequency: for each part of speech category the module counts the frequency of this POS category in the considered chunk. Each POS category has a specific dimension in the final vector; each dimension is set with

the corresponding frequency. Documents are processed using RDI_POS tagger [2].

(d) Word type frequency: a corpus that combines (Gigaword [3] and classical Arabic) was used to calculate word frequency for Arabic language, we have categorized the Arabic words into a fixed number of classes (26 classes in our case) by calculating $\log_2(Word\ Frequeny)$. For example class 1, represents words that occur between 2 and 4 times, class 2 represents words that occur between 4 and 8 times, etc... . For each class the module counts the frequency of this class in the considered chunk. Each class category has a specific dimension in the final vector; each dimension is set with the corresponding frequency.

After representing all chunks as vectors, comes the role of the filtering module. The filtering module constructs a mean vector for all chunks' vectors. After calculating the mean vector, the cosine distance is calculated between each chunk vector and mean vector. The module then calculates a mean cosine distance for the previously calculated distances. Then using all chunk vectors the module calculates standard deviation. Using the mean value and standard deviation for each chunk vector if cosine distance with the mean vector is less than (mean_value – (param2) * standard_deviation) this chunk is classified as plagiarism. The consecutive chunks are then combined and reported as one plagiarism part.

3. EVALUATION

In the training phase, the RDI_RID system is trained in order to tune the two parameters controlling the RDI_RID system (param1) & (param2). The training process was held using the following configuration: param1 = 500, param2 = 0.5, the best performance achieved is presented in Table1.

Table 1. Best results RDI_RID achieved on training and testing data

	Recall	Precision	Granularity	PlagDet
Training	0.14	0.18	1.0	0.16
Testing	0.18	0.19	1.0	0.19

Table 2 summarizes final results of the intrinsic plagiarism detection sub-task at AraPlagDet-2015 competition.

Table 2. Results of Arabic intrinsic plagiarism competition for year 2015

Method	Magooda	Baseline
Macro precision	0.19	0.27
Macro recall	0.20	0.78
Micro precision	0.15	0.29
Micro recall	0.20	0.49
Granularity	1.00	1.09
Plagdet (macro)	0.19	0.38

4. TECHNICAL DETAILS

The systems evaluation carried out over training and test data was performed on a personal machine with plausible specs, 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#

RDI_RID was trained using the supported training data to get the best set of parameters for (param1 and param2)
The training time for the system is:

Table 3. Training Timeand testing time.

Training	11527 seconds
Testing	10440 seconds

It should be noted that the previously reported training time is per iteration not the whole process of tuning.

5. CONCLUSION

In this paper, the RDI_RID system was introduced for intrinsic plagiarism detection task. The RDI_RID system depends on vector representation of chunks using stylometric features. Despite being behind the baseline system, the lack of Arabic language resources made the process of developing such a system a hard task. RDI_RID system was the only system that participated in the competition; the unavailability of comparison to other systems(other than the baseline) limited the evaluation scope of RDI_RID system. We intend to introduce other features for RDI_RID system to boost RDI_RID performance by a huge factor.

6. REFERENCES

[1] Zechner, Mario, et al. *External and intrinsic plagiarism detection using vector space models*.Proc. SEPLN. 2009.

[2] www.rdi-eg.com/index.htm

[3] catalog.ldc.upenn.edu/LDC2011T11