Algorithms and Corpora for Persian Plagiarism Detection

Overview of PAN at FIRE 2016

Habibollah Asghari
School of Electrical and Computer Engineering
College of Engineering
University of Tehran
habib.asghari@ictrc.ac.ir

Salar Mohtaj
ICT Research Institute
Academic Center for Education, Culture and Research (ACECR)
Iran
salar.mohtaj@ictrc.ac.ir

Omid Fatemi
School of Electrical and Computer Engineering
College of Engineering
University of Tehran
omid@fatemi.net

Heshaam Faili
School of Electrical and Computer Engineering
College of Engineering
University of Tehran
hfaili@ut.ac.ir

Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València
Spain
prosso@dsic.upv.es

Martin Potthast
Bauhaus-Universität Weimar
Germany
martin.potthast@uni-weimar.de

ABSTRACT

The task of plagiarism detection is to find passages of text-reuse in a suspicious document. This task is of increasing relevance, since scholars around the world take advantage of the fact that information about nearly any subject can be found on the World Wide Web by reusing existing text instead of writing their own. We organized the Persian PlagDet shared task at PAN 2016 in an effort to promote the comparative assessment of NLP techniques for plagiarism detection with a special focus on plagiarism that appears in a Persian text corpus. The goal of this shared task is to bring together researchers and practitioners around the exciting topic of plagiarism detection and text-reuse detection. We report on the outcome of the shared task, which divides into two subtasks: text alignment and corpus construction. In the first subtask, nine teams participated, whereas the best result achieved was a PlagDet score of 0.922. For the second subtask of corpus construction, five teams submitted a corpus, which were evaluated using the systems submitted for the first subtask. The results show that significant challenges remain in evaluating newly constructed corpora.

CCS Concepts

•General and reference → General conference proceedings.

Keywords

Plagiarism Detection; Evaluation Framework; TIRA Platform; Shared Task; Persian PlagDet.

1. INTRODUCTION

In recent years, a lot of research has been carried out concerning text reuse and plagiarism detection for English. But the detection of plagiarism in languages other than English has received comparably little attention. Although there have been previous developments on tools and algorithms to assist detecting text reuse in Persian, little is known about their detection performance. Therefore, to foster research and development on Persian plagiarism detection, we have organized the first corresponding competition, held in conjunction with the PAN evaluation lab at FIRE 2016.

We overview the detection approaches of nine participating teams and evaluate their respective retrieval performance. Participants were asked to submit their software to the TIRA Evaluation-as-a-Service (EaaS) platform [8] instead of just sending run outputs, rendering the shared task more reproducible. The submitted pieces of software are maintained in executable form so that they can be re-run against new corpora later on. To demonstrate this possibility, we asked participants to also submit evaluation corpora of their own design, which were examined using the detection systems submitted by other participants.

In what follows, Section 2 reviews related work with respect to shared tasks on plagiarism detection. Section 3 describes the main steps of tasks. Section 4 describes the evaluation framework, explaining the TIRA evaluation platform as well as the construction of our training and test datasets alongside the performance measures used. In Section 5, the evaluation results of both the text alignment and the corpus construction subtasks are reported.

2. RELATED WORK

This section reviews recent competitions and shared tasks on plagiarism detection in English, Arabic and Persian.

PAN. Potthast et al. [16] first pointed out the lack of a controlled evaluation environment and corresponding detection quality measures to evaluate plagiarism detection systems as a major obstacle to evaluating plagiarism detection approaches. To overcome these shortcomings, they organized the first international competition on plagiarism detection in 2009 featuring two subtasks: external plagiarism detection and intrinsic plagiarism detection. An important by-product of this competition was the first evaluation framework for plagiarism detection, which consists of a large-scale plagiarism corpus and a detection quality measure called as PlagDet [16, 17].

The PAN competition was continued in the next years, improving the evaluation corpora with each iteration. As of 2012, the competition was revamped in the form of two new subtasks: source retrieval and text alignment. Moreover, at PAN 2015, for the first time, participants were invited to submit their own alignment corpora. Here, participants were asked to compile

corpora comprising artificial, simulated, or even real plagiarism, formatted according to the data format established for the previous shared tasks [20].

AraPlagDet. AraPlagDet is the first international competition on detecting plagiarism in Arabic documents. The competition was held as a PAN shared task at FIRE 2015 and included two sub-tasks corresponding to the first shared tasks at PAN: external plagiarism detection and intrinsic plagiarism detection [1]. The competition followed the formats used at PAN. One of the main motivations of organizers for this shared task was to raise awareness in the Arab world on the seriousness of plagiarism, and, to promote the development of plagiarism detection approaches that deal with the peculiarities of the Arabic language, providing for an evaluation corpus that allows for proper performance comparison between Arabic plagiarism detectors.

PlagDet Task at AAIC. The first competition on Persian plagiarism detection was held as the 3rd AmirKabir Artificial Intelligence Competition (AAIC) in 2015. The competition was the first to plagiarism detection in the Persian language and led to the release of the first plagiarism detection corpus in Persian [10]. Like AraPlagDet, the PAN standard framework on evaluation and corpus annotation has been used in this competition.

3. TASK DESCRIPTION

The shared task of Persian plagiarism detection divides into two subtasks: text alignment and corpus construction.

Text alignment is based on PAN evaluation framework to assess the detection performance plagiarism detectors: given two documents, the task is to determine all contiguous passages of reused texts between them. Nine teams participated in this subtask.

The corpus construction subtask invited participants to submit evaluation corpora of their own design for text alignment, following the standard corpus format. Five corpora were submitted to the competition. Their evaluation consisted of evaluating the validity of annotations via analyzing corpus statistics, such as the length distribution of the documents, the length distribution of the plagiarized passages, and the ratio of plagiarism per document. Moreover, we report on the performance of the aforementioned nine plagiarism detectors in detecting the plagiarism comprised within the submitted corpora.

4. EVALUATION FRAMEWORK

The text alignment subtask consists of identifying the exact positions of reused text passages in a given pair of suspicious document and source document. This section describes the evaluation platform, corpus, and performance measure that were used in this subtask. Moreover, the submitted detection approaches and their respective evaluation results are presented.

4.1 Evaluation Platform

Establishing an evaluation framework for Persian plagiarism detection was one of the primary goals of our competition, consisting of a large-scale plagiarism detection corpus along with performance measures. The framework may serve as a unified test environment for future activities on Persian plagiarism detection research.

Due to the diverse development environments of participants, it is preferable to set up a common platform that satisfies all their requirements. We decided to use the TIRA experimentation platform [8]. TIRA provides for a set of features that facilitate the reproducibility of our shared task while reducing its organizational overhead [6, 7]:

- TIRA provides every participant with a virtual machine that allows for the convenient deployment and execution of submitted software.
- Both Windows and Linux machines are available to participants, whereas deployed software need only be executable from a POSIX command line.
- TIRA offers a convenient web user interface that allows participants to self-evaluate their software by remote-controlling its execution.
- TIRA allows for evaluating submitted software against test datasets hosted at server side. Test datasets are never visible to participants providing for a blind evaluation, and also allowing for sensitive datasets to be used for evaluation that cannot otherwise be shared publicly.
- At the click of a button, the run output of given software is evaluated against the ground truth of a given dataset. Evaluation results are stored and made accessible on TIRA web page as well as for download.

TIRA is widely used as an Evaluation-as-a-Service platform for experimenting information retrieval tasks [9]. In particular, the evaluation platform was used in since the 4th international competition on plagiarism detection at PAN 2012 [18], and now it is a common platform for all of PAN shared tasks [19].

4.2 Evaluation Corpus Construction

In this section we describe the methodology for compiling the Persian Plagdet evaluation corpus used for our shared task. The corpus comprises cases of simulated, artificial, and real plagiarism. In general, there are a number of reasons why collecting only real plagiarism is not sufficient for evaluating plagiarism detectors. First, collections of real plagiarism that have been detected manually are usually skewed towards ease of detection (i.e. the more difficult a plagiarism case is to be detected, the less likely it will be detected after the fact). Second, collecting real plagiarism is expensive and time consuming. Third, a corpus comprising real plagiarism cases cannot be published due to ethical and legal issues [17]. Because of these reasons, methods to artificially create plagiarism, or to simulate plagiarism are often employed to compile plagiarism corpora. These methods aim at emulating humans who try to obfuscate their plagiarism by paraphrasing reused portions of text. An artificial method for compiling plagiarism corpora includes the use of automatic paraphrasing technology to obfuscate plagiarized passages. Simulated passages of plagiarized text are created manually using human resources and crowdsourcing. Simulated methods yield more realistic cases of plagiarism compared to artificial ones, whereas artificial methods are cheaper in terms of both cost and time and hence scalable.

Simulated cases of plagiarism. To create simulated cases of plagiarism, a crowdsourcing approach has been used. For this purpose, a dedicated crowdsourcing platform has been developed, and a paraphrasing task was designed for crowd workers. Paraphrased passages obtained via crowdsourcing were reviewed by experts to ensure quality. All told, about 10% of the crowdsourced paraphrases were rejected because of poor quality. Table 1 gives an overview of the demographics of the crowd workers recruited.

Table 1. Crowd worker demographics.

Worker Demographics		
Age	25 – 30	41%
	30 – 40	38%
	40 – 58	21%
Education	College	5%
	BSc.	25%
	MSc.	58%
	PhD	12%
Tasks per worker	Average	19.0
	Std. deviation	14.5
	Minimum	1
	Maximum	54
Gender	Male	74%
	Female	26%

Artificial cases of plagiarism. In addition to simulated plagiarism based on manual paraphrasing, a large number of artificially created plagiarism has been constructed for the corpus.

As mentioned above, artificial plagiarism is cheaper and faster to compile than simulated plagiarism. To create artificial plagiarism, the previously proposed method of random obfuscation has been used [16]. The method consists of random text operations (i.e. word addition, deletion, shuffling), semantic word variation, and POS-preserving word shuffling. A composition of these operations has been used to create low and high degrees of random obfuscation.

As a result, after the obfuscation of passages extracted from a set of source documents, the simulated and artificial cases of plagiarism were inserted into a selection of suspicious documents. Some key statistics of the plagiarism cases and the final corpus are shown in the Tables 2 and 3.

Table 2. Plagiarism case statistics.

Plagiarism Case Statistics		
Obfuscation	Number of cases	1628
	None (exact copy)	11%
	Artificial	81%
	Simulated	8%
Case length	Short (30 - 50 words)	35%
	Medium (100-200 words)	38%
	Long (200-300 words)	27%

Table 3. Corpus statistics.

Corpus Statistics		
Entire corpus	Number of documents	5830
	Number of plagiarism cases	4118
Document purpose	Source documents	48%
	Suspicious documents	52%
Document length	Short (1-500 words)	35%
	Medium (500-2500 words)	59%
	Long (2500-21000 words)	6%
Plagiarism per Document	Small (5% - 20%)	57%
	Medium (21% - 50%)	15%
	Much (50% - 80%)	18%
	Entirely (>80%)	10%

4.3 Performance Measures

The PlagDet measure was used to evaluate the submitted software. PlagDet is a weighted F-measure that combines character level precision, recall, and granularity into one metric so that plagiarism detection systems can be ranked [17]. The run output of a given detector lists detected passages of allegedly plagiarized text as character offsets and lengths. Detection precision and recall are then computed as shown in Equations 1 and 2 below. In these equations, S is the set of the actual plagiarism cases and R is the set of detected plagiarism cases:

$$prec(S,R) = \frac{1}{|R|}\sum_{r\in R}\frac{|\bigcup_{s\in S}(s \sqcap r)|}{|r|}, \qquad (1)$$

$$rec(S,R) = \frac{1}{|S|}\sum_{s\in S}\frac{|\bigcup_{r\in R}(s \sqcap r)|}{|s|}, \qquad (2)$$

$$where \quad s \sqcap r = \begin{cases} s \cap r & if\ r\ detects\ s \\ \emptyset & otherwise \end{cases}$$

The granularity measure assesses the capability of a detector to detect a plagiarism case as a whole as opposed to in several pieces. The granularity of a detector is defined as follows:

$$gran(S,R) = \frac{1}{|S_R|}\sum_{s\in S_R}|R_S|, \qquad (3)$$

where S denotes the set of plagiarism cases in the corpus, R denotes the set of detections reported by a plagiarism detector, S_R ⊆ S the cases detected by detections in R, and R_S ⊆ R detections that detect cases in S. Finally, the PlagDet measure is a combination of F_1, the equally-weighted harmonic mean of precision and recall, and granularity:

$$plagdet(S,R) = \frac{F_1}{\log_2\big(1 + gran(S,R)\big)}. \qquad (4)$$

5. SUBTASK 1: TEXT ALIGNMENT

This section overviews the submitted software and reports on their evaluation results.

5.1 Survey of Detection Approaches

Nine of 12 registered teams successfully submitted a software to TIRA for the text alignment task. All of the nine participants submitted working notes describing their approaches. In what follows, we survey the approaches.

Talebpour et al. [23] use -trie trees to index the source documents after preprocessing. The preprocessing steps are text tokenization, POS tagging, text cleansing, text normalization to transform text characters into a unique and normal form, removal of stop words and frequent words, and stemming. Moreover, FarsNet (the Persian WordNet) [22] is used to find words' synonyms and synsets. This may allow for detecting cases of paraphrased plagiarism based on replacing words with their synonyms. After preprocessing both documents, all of the words of a source document and their exact positions are inserted into a -trie. After inserting all source documents into a -trie structure, the suspicious document are iteratively analyzed, checking each word one by one against the –trie to find potential sources.

Minaei et al. [14] employ n-grams as seed heuristic to find primary matches between suspicious and source documents. Cases

of plagiarism without obfuscation and similar parts of paraphrased text can be found this way. In order to detect cases of plagiarized passages, matches closer than a specified threshold are merged. Finally, to decrease false positive cases, detected cases shorter than a pre-defined threshold are eliminated.

Momtaz et al. [15] use sentence boundaries to split source and suspicious documents. After text normalization and removal of stop words and punctuations, sentences of both documents are turned into graphs, where words represent nodes and an edge is established between each word and its four surrounding words. Such graphs obtained from suspicious and source documents are compared and their similarity computed, whereas sentences of high similarity are labeled as plagiarism. Finally, to improve granularity, sentences close to each other are merged to create contiguous cases of detected plagiarism.

Gillam et al. [5] use an approach based on their previous PAN efforts. The task of finding textual matching is undertaken without direct using of the textual content. The proposed approach produces a minimal representation of text by distinguishing content and auxiliary words. Moreover it produces matchable binary patterns directly from these dependent words on the number of classes of interest. Although the approach act similar to hashing functions, but no effort is taken to prevent collision. Contrary, hash collision is encouraged over short distances, by preventing reverse-engineering of the patterns, and uses the number of coincident matches to indicate the extent of similarity.

Mansoorizadeh et al. [11] and Ehsan et al. [2] use sentence boundaries to split source and suspicious documents like the approach in [15]. In both approaches, each sentence is represented under the vector space model, using TF-IDF as weighting scheme. Finally, sentences with cosine similarity greater than a pre-defined threshold between corresponding vectors are considered as cases of plagiarism. In [2] a subsequent match merging stage improves performance with respect to granularity. Moreover, overlapping passages and extremely short passages are removed for the same reason. The lack of such a merging stage in Mansoorizadeh et al.'s [11] approach yields high granularity and therefore a poor PlagDet score.

Like most of the submitted software, Esteki et al. [3] split documents into sentences to detect plagiarism cases. After a pre-processing phase, which includes normalization, stemming and stop words removal, a Support Vector Machine (SVM) classifier is used to separate "similar" sentences non-similar ones. The Levenshtein distance, the Jaccard coefficient, and the Longest Common Subsequence (LCS) are used as features extracted from pairs of sentences. Moreover, synonyms are detected to increase the likelihood of detecting paraphrased sentences.

Gharavi et al. [4] use a deep learning approach to represent sentences of suspicious and source documents as vectors. For this purpose, they use Word2Vec to extract words' vectors and to compute sentence vectors as average word vectors. The most similar sentences between pairs of source document and suspicious document are found using the cosine similarity, the Jaccard coefficient, reporting them as plagiarism cases.

Mashhadirajab et al. [12] use the vector space model (VSM) with TF-IDF weighting to create sentence vectors from source and suspicious documents. To gain better results, they use an SVM neural net to predict the obfuscation type in order to adjust the required parameters. Moreover, to calculate the semantic similarity between sentences, FarsNet [22] is used to extract synsets of terms. Finally, within extension and filtering steps similar sentences that are close to each other are merged while passages that either overlap or are too short are removed.

5.2 Evaluation Results

Table 4 shows the overall performance and runtimes of the nine submitted text alignment approaches. As can be seen, the approach of Mashhadirajab [12] has achieved the highest PlagDet score on the complete corpus and is hence ranked highest. Regarding runtime, the submission of Gharavi [4] and Minaei [14] are outstanding: they process the entire corpus in only 1:03 and 1:33 minutes, respectively. Table 5 shows the performance of the submitted software dependent on obfuscation types in the corpus. Although, due to the lack of true positives, no performance values can be computed for the sub-corpus without plagiarism, at least false positive detections for this sub-corpus influence the overall performance of participants on the whole corpus [18]. Gharavi [4] is ranked first in detection performance with highest PlagDet for "No obfuscation," and Mashhadirajab [12] achieves best performance for both "Artificial" and "Simulated" plagiarism. Among all participants, Mashhadirajab achieves best recall across all parts of the corpus, whereas Talebpour [23] and Gharavi [4] outperform it in precision.

6. SUBTASK 2: CORPUS CONSTRUCTION

This section overviews the five submitted text alignment corpora. In the first subsection we will have a survey of submitted corpora and will give a statistical overview of them. In the next subsection the results of validation and evaluation on the submitted corpora will be presented.

6.1 Survey of Submitted Corpora

All of the submitted corpora consist of Persian mono-lingual plagiarism for the task of text alignment, except for Mashhadirajab corpus [13] which also contains a set of cross-lingual English-Persian plagiarism cases. All of the corpora are formatted in accordance with the PAN standard annotation format for text alignment corpora. In particular, this includes two sets of documents, namely source documents and suspicious documents, where the latter are to be analyzed for plagiarism from any of the source documents. The annotations of plagiarism cases are stored separately from the text documents within XML documents for each pair of suspicious and source documents. Therein, each plagiarism case is annotated as follows:

- Start position and length of the source passage in the source document
- Start position and length of the suspicious passage in the suspicious document
- Obfuscation type (e.g., indicating to the way that a source passage has been paraphrased before being added as suspicious passage to the suspicious documents)

6.1.1 Dataset Overview

Table 6 shows an overview of the submitted text alignment corpora in terms of the corpus statistics also reported for our corpus. Mashhadirajab corpus [13] is the biggest one in terms of number of documents, whereas Abnar corpus contains the largest number of plagiarism cases. Samim corpus [21] includes larger documents compared to the other corpora, whereas a large volume of small documents have been used for construction of the ICTRC corpus. Samim corpus and the ICTRC corpus comprise the largest and the smallest plagiarism case, respectively. A variety of different obfuscation strategies have been employed. No

obfuscation (i.e., exact copy) and artificial obfuscation (random text operations) are two common strategies.

The length distributions of documents and plagiarized passages are depicted in Figures 1 and 2. Here, the ICTRC corpus contains stands out, containing the smallest documents and plagiarized passages among all submitted corpora. Figure 3 shows the distribution of the plagiarism ratio per suspicious document. The ratio of plagiarism per suspicious documents in Samim corpus is distributed more uniformly compared to the other submitted corpora. In what follows, the documents used to compile the corpora as well as the construction approaches are discussed in detail.

6.1.2 Document Sources

The first step to compile a plagiarism detection corpus is choosing the documents which will be used as the sets of source documents and suspicious documents. Many plagiarism detection corpora intend to simulate plagiarism in technical texts, so that Wikipedia articles and scientific papers are often employed as source and suspicious documents sources in these corpora. This also pertains to the corpora submitted, which mainly employ journal articles and Wikipedia articles. Wikipedia articles have been used as resource to compiling the ICTRC corpus and Niknam corpus.

Table 4. Overall detection performance for the nine approaches submitted.

Rank / Team	Runtime (h:m:s)	Recall	Precision	Granularity	F-Measure	PlagDet
1 Mashhadirajab	02:22:48	**0.9191**	0.9268	1.0014	**0.9230**	**0.9220**
2 Gharavi	**00:01:03**	0.8582	0.9592	**1**	0.9059	0.9059
3 Momtaz	00:16:08	0.8504	0.8925	**1**	0.8710	0.8710
4 Minaei	00:01:33	0.7960	0.9203	1.0396	0.8536	0.8301
5 Esteki	00:44:03	0.7012	0.9333	**1**	0.8008	0.8008
6 Talebpour	02:24:19	0.8361	**0.9638**	1.2275	0.8954	0.7749
7 Ehsan	00:24:08	0.7049	0.7496	**1**	0.7266	0.7266
8 Gillam	21:08:54	0.4140	0.7548	1.5280	0.5347	0.3996
9 Mansourizadeh	00:02:38	0.8065	0.9000	3.5369	0.8507	0.3899

Table 5. Detection performance of the nine approaches submitted, dependent on obfuscation type.

Team	No obfuscation				Artificial Obfuscation				Simulated Obfuscation			
	Recall	Precision	Granularity	PlagDet	Recall	Precision	Granularity	PlagDet	Recall	Precision	Granularity	PlagDet
Mashhadirajab	**0.9939**	0.9403	**1**	0.9663	**0.9473**	0.9416	1.0006	**0.9440**	**0.8045**	0.9336	1.0047	**0.8613**
Gharavi	0.9825	0.9762	**1**	**0.9793**	0.8979	0.9647	**1**	0.9301	0.6895	**0.9682**	**1**	0.8054
Momtaz	0.9532	0.8965	**1**	0.9240	0.9019	0.8979	**1**	0.8999	0.6534	0.9119	**1**	0.7613
Minaei	0.9659	0.8663	1.0113	0.9060	0.8514	0.9324	1.0240	0.8750	0.5618	0.9110	1.1173	0.6422
Esteki	0.9781	0.9689	**1**	0.9735	0.7758	0.9473	**1**	0.8530	0.3683	0.8982	**1**	0.5224
Talebpour	0.9755	**0.9775**	**1**	0.9765	0.8971	**0.9674**	1.2074	0.8149	0.5961	0.9582	1.4111	0.5788
Ehsan	0.8065	0.7333	**1**	0.7682	0.7542	0.7573	**1**	0.7557	0.5154	0.7858	**1**	0.6225
Gillam	0.7588	0.6257	1.4857	0.5221	0.4236	0.7744	1.5351	0.4080	0.2564	0.7748	1.5308	0.2876
Mansourizadeh	0.9615	0.8821	3.7740	0.4080	0.8891	0.9129	3.6011	0.4091	0.4944	0.8791	3.1494	0.3082

Niknam used 3000 documents larger than 4000 characters, and ICTRC used about 6000 documents larger than 1500 characters. Abnar used texts from a set of novels that were translated to Persian. Despite the genre of books, the documents found in the corpus are not as large as might be expected. Mashhadirajab [13] and Samim [21] used scientific papers to compile their corpora. Mashhadirajab used a combination of Wikipedia articles (40%), articles from the Computer Society of Iran Computer Conference (CSICC) (13%), theses available in online (13%) and Persian open access articles (34%). Samim also collected Persian open access papers from peer reviewed journals to compile their text alignment corpus. The papers used include papers from the humanities (57%), science (25%), veterinary science (10%) and other related subjects (8%).

6.1.3 Obfuscation Synthesis

The second step in compiling a plagiarism detection corpus is to obfuscate passages selected from source documents and then insert them into suspicious documents. Obfuscating text passages aims at emulating plagiarism cases whose authors try to conceal the fact their plagiarized, making it more difficult for human reviewers and plagiarism detection systems alike to identify the plagiarized passages afterwards. As discussed above, creating obfuscated plagiarism manually is laborious and expensive, so that most participants resorted to automatic obfuscation methods. It is remarkable that two of the corpora (the ones of Mashhadirajab and ICTRC) comprise plagiarism that has been manually created. Otherwise, a variety of different approaches have been employed for obfuscation (see Table 6, rows "Obfuscation type"). All of the submitted corpora also contain a portion of plagiarized passages without any obfuscation to simulate verbatim copying.

Niknam employed a set of text operations consisting of addition, deletion and shuffling of words, replacing words with their synonyms and POS-preserving word replacement. Similar obfuscation strategies have been used to compile Samim's corpus. It contains "Random Text Operations" and "Semantic Word Variation" in addition to "No obfuscation." In addition to these obfuscation types, the authors of the ICTRC corpus used a crowdsourcing platform for paraphrasing test passages. About 30 people of various ages, both genders, and different levels of education have participated in the paraphrasing process. Abnar's corpus comprises obfuscation approaches such as replacing words with synonyms, shuffling sentences, circular translation, and a combination of the aforementioned ones. The circular translation approach includes translating the text to an intermediate language and then translating it back to the original one, hoping that the resulting text will significantly differ from the original one while maintaining its meaning. From a diversity point of view, Mashhadirajab's corpus contains the most variety in terms of obfuscation. In addition to artificial and simulated cases, they used summarizing cyclic translation and text manipulation approaches to create cases of plagiarism. Moreover, the corpus comprises also cross-lingual plagiarism where source documents have been translated to Persian using manual and automatic translation.

Table 6. Corpus statistics for the submitted corpora.

		Niknam	Samim	Mashhadirajab	ICTRC	Abnar
Entire corpus	Number of documents	3218	4707	11089	5755	2470
	Number of plagiarism cases	2308	5862	11603	3745	12061
Document purpose	Source documents	52%	50%	48%	49%	20%
	Suspicious documents	48%	50%	52%	51%	80%
Document length	Short (1-10000 words)	35%	2%	53%	91%	51%
	Medium (10000-30000 words)	56%	48%	32%	8%	48%
	Long (> 30000 words)	9%	50%	15%	1%	1%
Plagiarism per document	Hardly (<20%)	71%	29%	39%	57%	29%
	Medium (20%-50%)	28%	25%	14%	37%	60%
	Much (50%-80%)	1%	31%	20%	6%	10%
	Entirely (>80%)	-	15%	27%	-	1%
Case length	Short (1-500 words)	21%	15%	6%	51%	45%
	Medium (500-1500 words)	76%	22%	52%	46%	54%
	Long (>1500 words)	3%	63%	42%	3%	1%
Obfuscation types	No obfuscation (exact copy)	25%	40%	17%	10%	22%
	Artificial (word replacement)	27%	-	-	-	-
	Artificial (synonym replacement)	25%	-	-	-	-
	Artificial (POS-preserving shuffling)	23%	-	-	-	-
	Random	-	40%	-	81%	-
	Semantic	-	20%	-	-	15%
	Near Copy	-	-	28%	-	-
	Summarizing	-	-	33%	-	-
	Paraphrasing	-	-	6%	-	-
	Modified Copy	-	-	4%	-	-
	Circle Translation	-	-	3%	-	21%
	Semantic-based meaning	-	-	1%	-	-
	Auto Translation	-	-	2%	-	-
	Translation	-	-	6%	-	-
	Simulated	-	-	-	9%	-
	Shuffle Sentences	-	-	-	-	21%
	Combination	-	-	-	-	21%

6.2 Corpus Validation

In order to validate the submitted corpora, we analyzed them quantitatively and qualitatively. For the latter, samples have been drawn from each corpus and obfuscation type for manual review. The review involved of validating the plagiarism annotations, such as offsets and lengths of annotated plagiarism in both source and suspicious documents. Moreover, the suspicious passage and its corresponding source have been checked manually to observe the impact of different obfuscation strategies as well as the level of obfuscation. Altogether, no important issues have been found among the studied samples during peer-review.

In addition to manual review, we also analyzed the corpora quantitatively: Figures 1 and 2 depict the length distributions of the documents and the plagiarism cases in the corpora. Both Abnar's corpus and the ICTRC corpus have clear expected values, whereas the other corpora are more evenly distributed. Figure 3 depicts the ratio of plagiarism per document, showing that the ratios are quite unevenly distributed across corpora; Niknam's corpus and the ICTRC corpus comprise mostly suspicious documents with a small ratio of plagiarism. Figures 4 and 5 show the distribution of plagiarized passages in terms of where they start within suspicious documents (i.e., their character offset), and where they start within source documents. The distributions of start offsets within suspicious documents are similar across all corpora with a negative bias against offsets at the beginning of a suspicious document (see Figure 4). The distributions are also similar for the start offsets within source documents with one notable exception: the source passages of Samim's corpus have almost always been chosen from the same offsets of source documents which is a clear bias and may allow for trivial detection.

Finally, we analyzed the plagiarized passages in the submitted corpora with regard to their similarity between source passage and suspicious passage. The experiment consists of comparing source passages with suspicious passages using 10 retrieval models. Each model is an n-gram vector space model (VSM), where n ranges from 1 to 10 words, employing stop word removal, TF-weighting and the cosine similarity [17]. For high-quality corpora, a pattern similar to that of PAN corpora is expected.

Since there are many obfuscate types to choose from, we only compare a selection: the simulated plagiarism cases of Mashhadirajab and ICTRC are compared to the PAN corpora (Figure 6). Moreover, the artificial parts of all corpora are compared to each other (Figure 7). Abnar's corpus is omitted since it lacks artificial obfuscation. Almost all of the corpora show same patterns of similarity for different ranges of n, except the Mashhadirajab's corpus which has a higher range of similarity in comparison others.

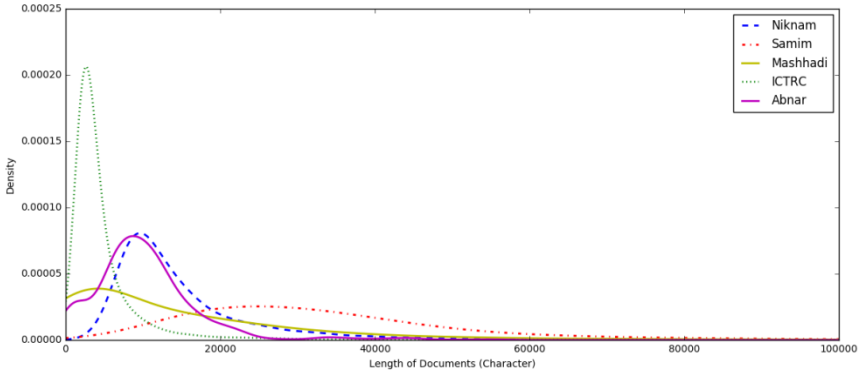

Figure 1. Length distribution of documents.

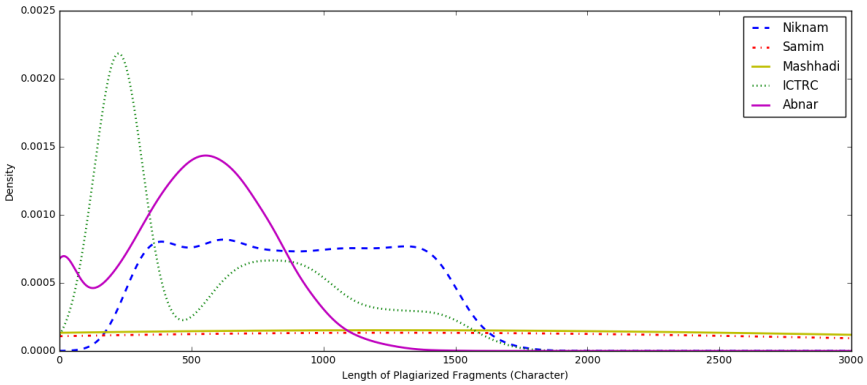

Figure 2. Length distribution of fragments.

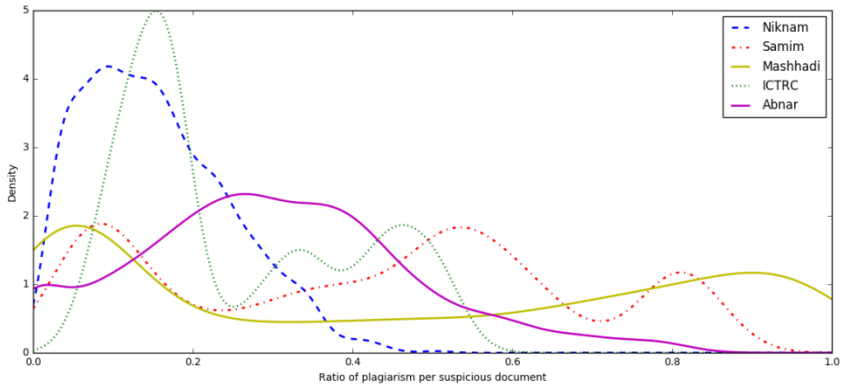

Figure 3. Ratio of plagiarism per document.

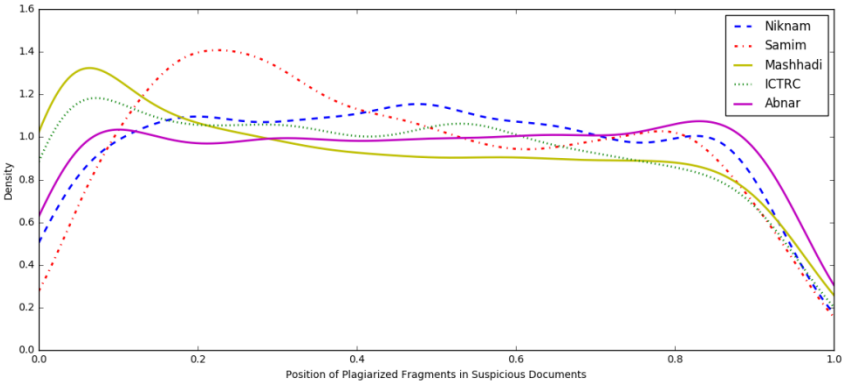

Figure 4. Start position of plagiarized fragments in suspicious documents.

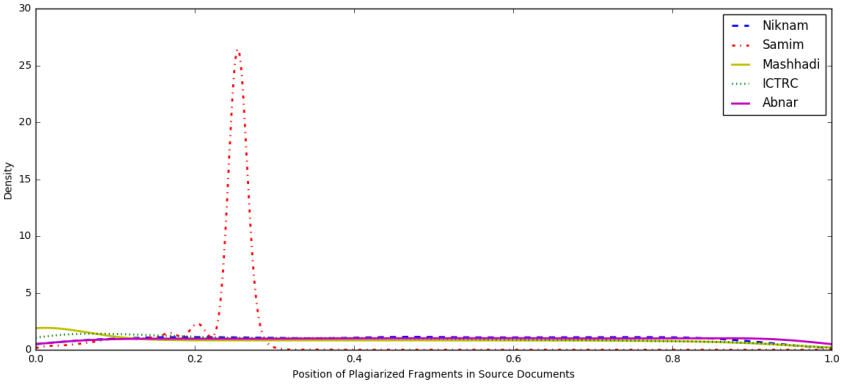

Figure 5. Start position of plagiarized fragments in source documents.

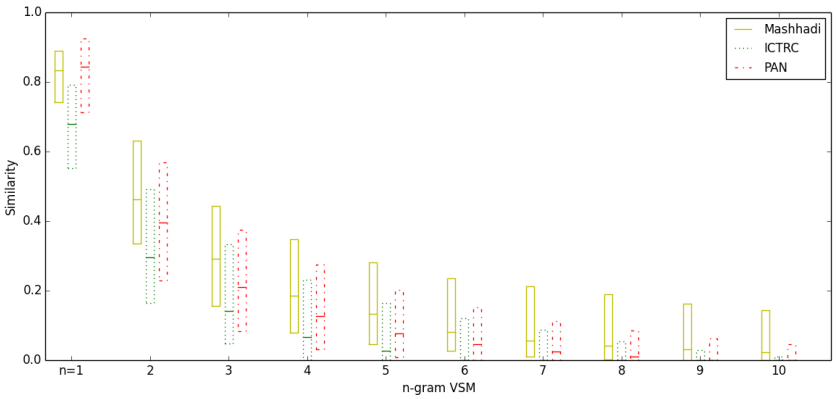

Figure 6. Comparison of Simulated part of Mashhadirajab and ICTRC corpora.

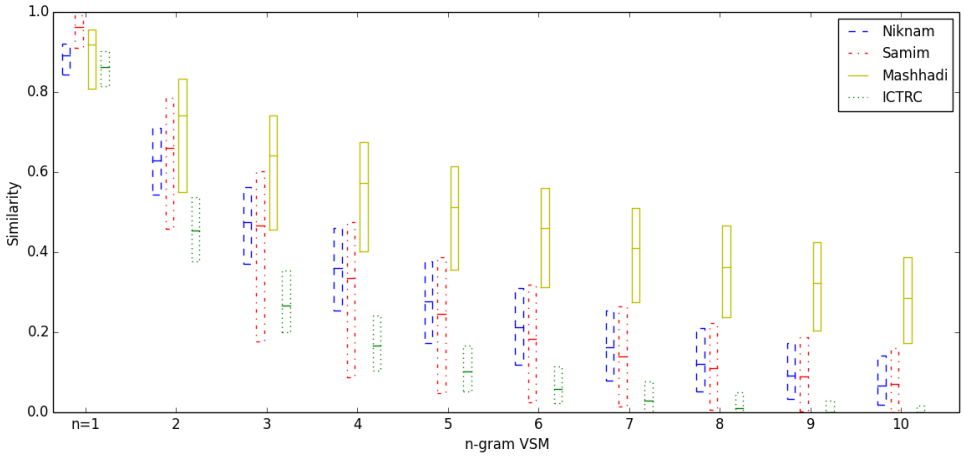

Figure 7. Comparison of Artificial part of Niknam, Samim, Mashhadirajab and ICTRC corpora.

6.3 Corpus Evaluation

Exploiting the virtues of TIRA, our final experiment was to run the nine submitted detection approaches on the five submitted corpora, providing for a first impression on how difficult it is to detect plagiarism within these corpora. Table 7 overviews the results of this experiment. Unfortunately, not all submitted approaches succeeded in processing all corpora. One reason was scalability issues: since some of the submitted corpora are significantly larger than our evaluation corpus, it seems participants did not pay a lot of attention to scalability. The approaches of Talebpour, Mashhadi, and Gillam failed to process the corpora in time. The approaches of Momtaz and Esteki failed to process some of the corpora at first, the results of the former are only partially reliable to date, whereas the latter of which could be fixed in time. This shows that submitting datasets to shared tasks presents its own challenges. Participants will be invited to fix their software to make it work on all corpora, so that further results may become available after publication of this paper, e.g., on TIRA's web page. Considering the detection performance, it can be seen that the PlagDet scores are generally lower compared to our corpus, except for the ICTRC corpus, where the same performance scores have been reached. This shows that the submitted corpora present their own challenges, rendering them more difficult, and presenting future researchers with new opportunities for contributions.

Given the results from all our experiments, the submitted corpora are of reasonable quality. Although some of them are too easy to be solved and comprise a biased sample of plagiarism cases, the diversity of corpora ensures that future evaluations can be done with confidence as long as all available datasets are employed.

Table 7. PlagDet performance of some submitted approaches on the submitted corpora.

Team	Niknam	Samim	Mashhadirajab	ICTRC	Abnar
Gharavi	0.8657	**0.7386**	**0.5784**	**0.9253**	0.3927
Momtaz	0.8161	-	-	0.8924	-
Minaei	**0.9042**	0.6585	0.3877	0.8633	**0.7218**
Esteki	0.5758	-	-	-	0.3830
Ehsan	0.7196	0.5367	0.4014	0.7104	0.5890
Mansourizadeh	0.2984	-	0.1286	-	0.2687

7. CONCLUSION

In conclusion, our shared task has attracted considerable attention from the community of scientists working on plagiarism detection. The shared task has served as a means to establish a new state of the art in performance evaluation for Persian plagiarism detection. Altogether six new evaluation corpora are available now, and nine detection approaches have been evaluated on them. The results show that Persian plagiarism detection is far from being a solved problem. In addition, our contributions broaden the scope of the text alignment task which has been studied mostly for English until now. This may allow future work on plagiarism detection approaches that work on both languages simultaneously.

8. ACKNOWLEDGMENTS

This work has been funded by ICT Research Institute, ACECR, under the partial support of Vice Presidency for Science and Technology of Iran - Grant No. 1164331. The work of Paolo Rosso has been partially funded by the SomEMBED MINECO TIN2015-71147-C2-1-P research project and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030). We would like to thank the participants of the competition for their dedicated work. Our special thanks go to the renowned experts who served on the organizing committee for their contributions and devoted work to make this shared task possible. We would like to thank Javad Rafiei and Khadijeh Khoshnava for their help in construction of evaluation corpus. We

are also immensely grateful to Vahid Zarrabi for his comments and valuable help along the way which greatly assisted this challenging shared task.

9. REFERENCES

[1] Bensalem, I., Boukhalfa, I., Rosso, P., Abouenour, L., Darwish, K., & Chikhi, S. 2015. Overview of the AraPlagDet PAN@ FIRE2015 Shared Task on Arabic Plagiarism Detection, CEUR-WS.org, vol. 1587, pp. 111-122

[2] Ehsan, N, Shakery, A. 2016. A Pairwise Document Analysis Approach for Monolingual Plagiarism Detection, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[3] Esteki, F, Safi Esfahani, F. 2016. A Plagiarism Detection Approach Based on SVM for Persian Texts, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[4] Gharavi, E, Bijari, k, Zahirnia, K, Veisi, H. 2016. A Deep Learning Approach to Persian Plagiarism Detection, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[5] Gillam, L., and Vartapetiance, A., 2016. From English to Persian: Conversion of Text Alignment for Plagiarism Detection, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation*, Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[6] Gollub, T., Burrows, S. and Stein, B., 2012, August. First experiences with TIRA for reproducible evaluation in information retrieval. In *SIGIR* (Vol. 12, pp. 52-55).

[7] Gollub, T., Stein, B., & Burrows, S. 2012, August. Ousting ivory tower research: towards a web framework for providing experiments as a service. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 1125-1126). ACM.

[8] Gollub, T., Stein, B., Burrows, S., and Hoppe, D. 2012. September. TIRA: Configuring, executing, and disseminating information retrieval experiments. In *2012 23rd International Workshop on Database and Expert Systems Applications* (pp. 151-155). IEEE.

[9] Hopfgartner, F., Hanbury, A., Müller, H., Kando, N., Mercer, S., Kalpathy-Cramer, J., Potthast, M., Gollub, T., Krithara, A., Lin, J. and Balog, K., 2015, June. Report on the Evaluation-as-a-Service (EaaS) expert workshop. In *ACM SIGIR Forum* (Vol. 49, No. 1, pp. 57-65). ACM.

[10] Khoshnavataher, K., Zarrabi, V., Mohtaj, S., & Asghari, H. 2015. Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation. Notebook for PAN at CLEF 2015. CLEF (Working Notes).

[11] Mansoorizadeh, M, Rahgooy, T. 2016. Persian Plagiarism Detection Using Sentence Correlations, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[12] Mashhadirajab, F, Shamsfard, M. 2016. A Text Alignment Algorithm Based on Prediction of Obfuscation Types Using SVM Neural Network, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[13] Mashhadirajab, F, Shamsfard, M, Adelkhah, R, Shafiee, F., Saedi, S. 2016. A Text Alignment Corpus for Persian Plagiarism Detection, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[14] Minaei, B, Niknam, M. 2016. An n-gram based Method for Nearly Copy Detection in Plagiarism Systems, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[15] Momtaz, M, Bijari, K, Salehi, M, Veisi, H. 2016. Graph-based Approach to Text Alignment for Plagiarism Detection in Persian Documents, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[16] Potthast, M., Stein, B., Eiselt, A., Barron, Cedeno, A., and Rosso, P., 2009. Overview of the 1st international competition on plagiarism detection. In *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse* (p. 1)

[17] Potthast, M., Stein, B., Barrón-Cedeño, A. and Rosso, P., 2010, August. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 997-1005). Association for Computational Linguistics.

[18] Potthast, M., Gollub, T., Hagen, M., Graßegger,J., Kiesel, J., Michel,M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B., 2012. Overview of the 4th International Competition on Plagiarism Detection. In CLEF (Online Working Notes/Labs/Workshop).

[19] Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E. and Stein, B., 2014, September. Improving the Reproducibility of PAN's Shared Tasks. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 268-299). Springer International Publishing.

[20] Potthast, M., Hagen, M., Göring, S., Rosso, P. and Stein, B., 2015. Towards data submissions for shared tasks: first experiences for the task of text alignment. *Working Notes Papers of the CLEF*, pp.1613-0073.

[21] Rezaei Sharifabadi, M., Eftekhari, S. A. 2016. Mahak Samim: A Corpus of Persian Academic Texts for Evaluating Plagiarism Detection Systems, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[22] Shamsfard, M. 2008. Developing FarsNet: A lexical Ontology for Persian. Proceedings of the 4th global WordNet conference.

[23] Talebpour, A, Shirzadi, M, Aminolroaya, Z. 2016. Plagiarism Detection based on a Novel Trie-based Approach, In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.