A Text Alignment Corpus for Persian Plagiarism Detection

Fatemeh Mashhadirajab
NLP Research Lab,
Faculty of Computer Science and Engineering,
Shahid Beheshti University, Iran
f.mashhadirajab@mail.sbu.ac.ir

Mehrnoush Shamsfard
NLP Research Lab,
Faculty of Computer Science and Engineering,
Shahid Beheshti University, Iran
m-shams@sbu.ac.ir

Razieh Adelkhah
NLP Research Lab,
Faculty of Computer Science and Engineering,
Shahid Beheshti University, Iran
r.adelkhah@yahoo.com

Fatemeh Shafiee
NLP Research Lab,
Faculty of Computer Science and Engineering,
Shahid Beheshti University, Iran
f.shafiee@hotmail.com

Chakaveh Saedi
NLX Lab of university of Lisbon
Department of Informatics
Portugal
Ch_saedi@sbu.ac.ir

ABSTRACT
This paper describes how a Persian text alignment corpus was constructed to evaluate plagiarism detection systems. This corpus is in PAN format and contains 11,089 documents and more than 11,603 plagiarism cases. Efforts were made to simulate various types of plagiarism manually, semi-automatically, or automatically in this large-scale corpus.

CCS Concepts
• Information systems → Near-duplicate and plagiarism detection.
• Information systems → Evaluation of retrieval results.

Keywords
Plagiarism detection; Text alignment corpus; Types of plagiarism; Corpus construction.

1. INTRODUCTION
Plagiarism is using others' phrases, solutions, ideas, or results with no faithful citation. The considerable worldwide growth of plagiarism in recent years emphasizes the importance of dealing with this phenomenon. Plagiarism is an ethical challenge in science to which there are many contributing factors; however, the development in plagiarism detection systems can at least result in a reduction in plagiarism growth. The PAN competition which has been held yearly since 2009 is one famous example in the plagiarism detection area. Such competitions provide a suitable layout for comparing researchers' different approaches and solutions. Having a suitable evaluating corpus is one of the most important requirements in such a competition. This article describes how a corpus for the task of text alignment corpus construction in Persian Plagdet 2016 [1] was constructed. Researchers have produced different taxonomies of plagiarism types [19, 20, 21]. The taxonomy of plagiarism presented by Alzahrani *et al.* [2] is shown in Fig. 1. This taxonomy was used in the current study to construct a data set for evaluating plagiarism detection systems. In the second section, we review available text alignment corpora and in the third section the method for developing a corpus is described. The fourth section explains how to simulate each mentioned type of plagiarism, and finally, dataset statistics for the constructed corpus are given.

2. RELATED WORK
Numerous text alignment datasets, including PAN plagiarism corpora, have been employed to evaluate text alignment algorithms in plagiarism detection competitions since 2009 [3, 8, 9, 16, 17, 18]. The first text alignment data set that was developed by PAN in 2010 [3] includes 27,073 documents in English and 68558 cases of plagiarism. In this massive data set, plagiarism cases are generally provided with two strategies. Simulated Plagiarism is the first strategy in which 907 people were asked to rewrite the given original texts so that the meaning of the original is not changed but the appearance of the text be replaced with different words and phrases. Artificial Plagiarism is the second strategy where automated methods have been used to change the text. Techniques used in this section are divided into three categories. The first category uses techniques to insert, remove and replace words and short phrases, the second category uses techniques to replace words with their synonyms, antonyms, hyponyms, or hypernyms and the third category uses the movement of vocabulary in a sentence with the same POS Tag. Another text alignment corpus which was offered by was used by PAN to evaluate algorithms in 2013 and 2014 [9,18]. This corpus includes the 3653 suspicious document and 4774 source document in English and 8,000 cases of plagiarism. This corpus consists three types of obfuscation strategies: Random obfuscation**,** Cyclic Translation obfuscation and Summary obfuscation. In Random obfuscation they use techniques similar to Artificial Plagiarism strategy. In cyclic translation obfuscation, a text is manually or automatically translated into another language and after edition it is translated into the source language again. To simulate Summary obfuscation which is considered as a plagiarism technique, PAN has used evaluation corpora of summarizer automatic system. Moreover, in year 2015, instead of inviting text alignment algorithms, PAN demanded to have text alignment data set sent, and a total of 8 data sets have been submitted to the PAN 2015. [22-29]. These data sets are in different languages and have used a variety of techniques to obfuscate the text. Alvi' corpus [22] is among such sent corpora, which includes 272 documents in English and 150 plagiarism cases. Alvi uses character-substitution, human-retelling and synonym-replacement techniques to obfuscate text. Asghari [27] has submitted a Persian-English parallel corpus to the PAN 2015. This corpus includes 27115 documents and 11200 plagiarism cases. Cheema'

corpus [23] includes 1000 documents in English and 250 plagiarism cases. In this corpus, in order to obfuscate texts, a number of students of different academic courses were asked to select and rewrite a number of texts related to their fields and put them inside documents with the same subject such as Wikipedia documents. Also A bilingual English-Urdu corpus that includes 1000 documents and 270 plagiarism cases sent to the PAN 2015 competitions by Hanif [24]. In this corpus he used machine translation with and without manual correction of results, with the use of random-obfuscation strategy in some translation results to obfuscate the text. Khoshnavataher [26] has presented a corpus in Persian that includes 2111 documents and 823 plagiarism cases. In order to obfuscate, he used Random obfuscation technique and no-obfuscation technique where a piece of the source document is added to suspicious document without any change. Kong [25] also took part in the PAN 2015competition with 160 documents in Chinese and 152 cases of plagiarism. In order to obfuscate text, Kong asked a number of volunteers to write a paper for topics that have been identified. Mohtaj's corpus [28] also was submitted to PAN 2015 with 4261 documents in English and 2781 plagiarism cases. In this corpus, techniques of no-obfuscation, random-obfuscation and simulated-obfuscation is used to obfuscate text. Palkovskii [29] also makes use of PAN 2013-2014 corpus to prepare a corpus that included 5057 documents in English and 4185 plagiarism cases. Obfuscation was made based on techniques of random-obfuscation, no-obfuscation, cyclic-translation-obfuscation and summary-obfuscation. In the rest of this paper we will describe the construction method we employed to develop a text alignment corpus to evaluate Persian plagiarism detection systems.

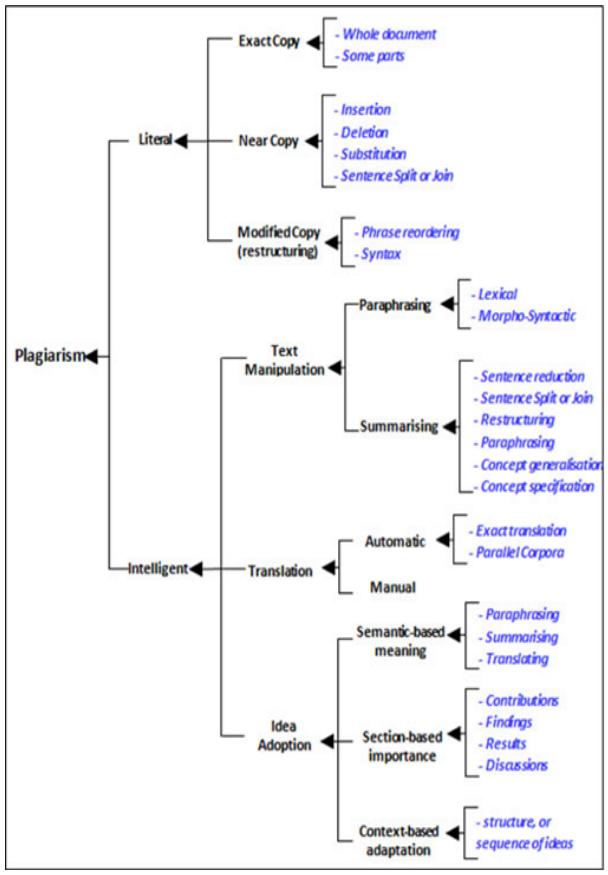

Fig.1. A taxonomy of plagiarism [2]

3. TEXT ALIGNMENT CORPUS CONSTRUCTION

The goal in text alignment is to identify plagiarized segments for each given source and suspicious document pairs [8].

In this study, a text alignment corpus is created to evaluate plagiarism detection systems on Persian scientific documents. The conducted procedure to build this corpus is described herein.

a. Data Source Preparation

We use some documents of source documents collection in Mahtab plagiarism detection system [15] to construct our text alignment corpus. Mahtab plagiarism detector is developed at the Shahid Beheshti University NLP Lab. The goal of Mahtab is detecting plagiarized articles in the fields of computer science and engineering. Our text alignment corpus in this study contains 11,089 documents. They are all articles or theses in the fields of computer science and engineering and also electrical engineering with the following distribution:

4,500 documents from Wikipedia articles;

1,500 documents from CSICC[1] articles (2004-2015);

1,500 documents from articles and theses available from online stores;

3,589 documents from free Persian resources including mag-iran[2], iran-doc[3], SID[4], prozhe[5], and MatlabSite[6].

b. Documents Clustering

Since all documents in the corpus are in the field of computer science, there is a general similarity among them. The method proposed for document clustering is to estimate cluster features first, and then perform clustering based on the introduced features. Finally, an optimization process improves the results. To extract features, all words included in a document are extracted and stemmed using STeP-1 [4]. Each word is then labeled based on Table 1 which is introduced by Makrehchi [5]. For each document, an n-bit histogram vector is produced named V(v_1, v_2, …, v_n) where n is the number of features. If feature_i existed in a document, $v_i = 1$; otherwise, $v_i = 0$. Afterwards, these vectors are classified based on the K-means algorithm and Cosine similarity. To optimize the extracted features in a cluster, the sum of all vectors of a cluster is found and used to produce H (h_1, h_2, …, h_n)), where h1 indicates the number of documents containing the first feature. H is produced for all clusters; Equation (1) can be used to calculate the weight of each feature in the corresponding cluster.

$$w_{h_1} = \frac{h_1}{f_c} \qquad Eq(1)$$

f_c indicates the number of clusters containing this feature. The features are sorted in a descending order based on their weights. Afterwards, the first 100 words of each cluster are considered as the features for that cluster. To improve clustering accuracy, the

[1] Computer Society of Iran Computer Conference, http://csi.org.ir

[2] http://mag-iran.com

[3] http://www.irandoc.ac.ir

[4] http://sid.ir

[5] http://www.prozhe.com

[6] www.MatlabSite.com

membership degree to each cluster must be calculated, and documents must be placed in the most similar cluster. The membership degree for each document is calculated as follows:

$$MembershipDegree = \frac{Sum \times Count}{\sqrt{T_D}} \qquad Eq(2)$$

In which Sum is the number of all seen cluster features (the first 100 words of each cluster based on their weights are considered as cluster features) in the corresponding document, $Count$ is the number of cluster features occurring in the document, and T_D is the document length.

Table 1. Three categories of words in a corpus [5]

frequency of the term in the corpus \ document frequency	Low	Medium	High	
	Key word	Feature	Stop Word	High
	Key word	Feature	Stop Word	Medium
	Stop Word	Stop Word	Stop Word	Low

c. Suspicious Documents Selection

Some documents are randomly selected from each cluster as suspicious documents. Almost half of the documents are employed as source documents and the other half as suspicious documents. Half of the suspicious documents are considered as no-plagiarism documents, and the other half of the documents are used to produce plagiarized documents.

d. Source Set for a plagiarism Document

For each plagiarism document in a cluster, a set of source documents named D_{src} is selected in which there is no repeated document or very similar document to suspicious document, a source document can be used in many suspicious documents so every time a suspicious document can select each source document from the corresponding cluster therefore the selected documents may be selected by this suspicious document before. Moreover if the similarity between source document and suspicious document is more than 50% before adding plagiarism passages to suspicious document, then it is not a good selection because even if a hard strategy is used to obfuscate, plagiarism passages may be discovered by simple similarity detection algorithms. To create D_{src} for each plagiarism document, a document from the corresponding cluster is selected randomly; if the similarity based on the SimHash method [10] between the selected document and each document in D_{src} is more than 50%, the document is considered repeated; otherwise, it is included in D_{src}. This step is continued until there are at least 3 documents in D_{src}. A D_{src} contains a suspicious document and at least 2 source documents. The reason for employing the SimHash method is the noticeable results achieved in [11]. In this phase, the source and suspicious document pairs are specified. In this way, 3,867 paired documents (source- suspicious) are produced to be included in the corpus.

e. Source Set for a No-plagiarism Document

For each no-plagiarism document, a source set is selected as described in step d. However, in this step a similarity detection at the sentence level for each randomly selected source and suspicious document is considered based on the Jaccard similarity measure and a threshold of 0.9; if there are no same sentences between both mentioned files, the source document is added to D_{src}. Using this method, 2,630 pairs of documents are produced in this phase.

f. Source Documents Segmentation

In this step, first a document is divided into its paragraphs. Each subsequence of paragraphs that contain at least 300 words is considered a segment. If a paragraph contain less than 300 words it is combined with the next paragraph. Ultimately, all segments contain at least 300 words.

g. Determine the Length of Plagiarized Segments in each Suspicious Document

The number of plagiarized segments which are employed in a suspicious document depends on the source document length and the length of any plagiarized segments. To decide the number of segments to use from a source document in its paired suspicious one, all paired documents are first labeled. Each randomly selected pair of documents is labeled as "entirely," "much," "medium," or "hardly" as described below.

- **Entirely:** The length of the source document is more than 80% of the length of the suspicious document.
- **Much:** The length of the source document is more than 50%-80% of the length of the suspicious document.
- **Medium:** The length of the source document is about 20%-50% of the length of the suspicious document.
- **Hardly:** The length of the source document is less than 20% of the length of the suspicious document.

If the number of paired documents with the same label is more than one-fourth of the number of paired documents with a label of smaller length that do not have enough paired documents, the label with the lower length is assigned; thus, a uniform distribution is obtained.

h. Segment Extraction

From each source document, some segments are randomly selected. The number of selected segments is based on the classification performed in step g.

i. Segment Obfuscation

This study offers a strategy to manually, semi-automatically, or automatically produce each type of plagiarism mentioned in Alzahrani's taxonomy of plagiarism. In this step, each segment is obfuscated based on one strategy and add to one suspicious document. It is noteworthy that all obfuscated segments included in a document must be obfuscated using the same strategy because according to PAN corpus format, there is no overlap between suspicious documents in different strategies[9] and only one type of plagiarism should be employed in each suspicious document.

j. Obfuscated Segment Insertion

In this step each obfuscated segment is inserted into a suspicious document in a randomly chosen space.

4. STRATEGIES FOR PLAGIARISMS TYPES

- **Exact Copy**

In this strategy, the segments produced in step h were inserted into a suspicious document with no obfuscation. Using this strategy, 324 paired documents were produced.

- **Near Copy**

According to Fig .1 a type of plagiarism is Near Copy [2] that consists insertion, deletion, substitution and sentence split or join methods. To create this kind of plagiarism, the segments produced in step h are obfuscated through deletion, insertion, sentence replacement, and sentence division. With this method, some randomly selected sentences are deleted from the segment and replaced with randomly selected sentences from the suspicious document. Then, some randomly selected sentences are swapped. Finally, complex sentences are identified and broken into main simple sentences. To do this, the complex sentence identifier developed at the Shahid Beheshti University NLP Lab is employed. Each complex sentence in this segment is replaced with its main and subordinate clauses, and 457 paired documents are produced based on this strategy.

- **Modified Copy**

In the taxonomy of plagiarism [2] there is a type of plagiarism called Modified Copy that to obfuscate a text using this strategy, the Persian sentence understanding and generation system introduced by Adelkhah *et al.* [7] is employed. This system performs a bidirectional conversion between Persian sentences and their semantic representation. It changes each sentence to its semantic representation and then generates the Persian sentence using semantic representation. To clarify, this system is composed of 2 sub-systems: 1) semantic representation production for sentences (sentence understanding), and 2) sentence production based on semantic representation (sentence generation). It is noteworthy that in the sentence production phase, in addition to structural changes, there might be samples of chunk relocations in a sentence or samples of word relocations in a chunk. The aim of this system is to produce sentences with the same meaning (deep structure) but different surface structures and words. Using this strategy, 465 paired documents are created.

- **Text Manipulation (Paraphrasing)**

Text Manipulation was performed as described earlier in Modified Copy. The difference here is the word replacement in the sentence generation phase. Each word is replaced with a synonym retrieved from FarsNet (Persian WordNet) [14] or FavaNet (WordNet of Computer domain) [13]. Hence, in addition to structure modification, different words are included in the sentence compared to the main sentence, although the concept remains the same. Chunks may be moved inside a sentence; however, there is no movement for words in a chunk. Using this method, 604 paired documents are produced.

- **Text Manipulation (Summarizing)**

The goal in this step is to obfuscate a text document using summarization methods. To create such queries, the automatic Persian summarizer introduced by Shafiee *et al.* [6] is used, and 506 paired documents are produced.

- **Automatic Translation**

According to types of plagiarism in Fig .1 translation is a type of plagiarism that is divided into automatic and manual translation. Hanif *et al.* [24] use the automatic translation strategy to obfuscate documents in their corpus. Moreover in the PAN 2013-2014 corpus [9, 18] use cyclic translation strategy.

We use all of above three strategies in our corpus (described in Automatic Translation, Manual Translation and Cyclic Translation stages). For the automatic translation strategy, the selected sections are translated from Persian to English by Google translate and the results are checked by Hunspell. Then they are added to the English suspicious documents. 306 paired documents are produced using this strategy.

- **Manual Translation**

The suspicious documents in this step are English articles in the field of computer engineering, and the source documents are Persian articles in the same field. The English articles are clustered as described in step b, and an equivalent English cluster is produced for each Persian one. Then, for each suspicious document, a source document from its equivalent Persian cluster is randomly selected. Based on what was described in steps f, g and h, some sections of the source document are selected. Afterwards, these sections are translated by experts in the fields of computer engineering and are added to the suspicious documents as described in step j. Seven hundred paired documents are produced using this strategy which can be employed to evaluate cross-language similarity detection systems (Persian-English).

- **Cyclic Translation**

With the cyclic translation strategy the selected sections are translated from English to Persian using Google translate, and the results are checked by Negar, a Persian spell checker developed at the NLP Lab of Shahid Beheshti University. The selected sections are then translated again from Persian to English. Finally, the results are checked by Hunspell and add to the English suspicious documents. Using this method, 388 paired documents are created.

- **Idea Adoption (semantic-based meaning)**

The goal in this step is to represent the main idea of a source document using new words/wording. Since most source documents are computer related theses and articles, automatic idea extraction would be a complex task here for which no high accurate system is yet available. Hence, the researchers asked computer experts to rewrite the idea of each document in their own words. To simplify the task, only important sections of source documents, such as the abstract, were considered. Source documents were distributed among three computer PhD candidates and 30 computer MS students, and 109 paired documents were produced.

5. DATASET STATISTICS

Overall, employing all the mentioned strategies, 11,603 plagiarism cases and 6,497 paired documents are produced, from which 2,650 are no-plagiarism, 780 are no obfuscation, and 3,067 are obfuscated ones. The dataset statistics are shown in Table 2.

6. CONCLUSION

This article describes a methodology for building a Persian corpus for evaluating plagiarism detection systems. This large-scale corpus is in PAN format. To produce this corpus, the focus is on the simulation of different types of plagiarism. Different strategies

are employed to create obfuscation in each plagiarism category; hence, a variety of plagiarism types in large volume are created.

Table 2. Dataset statistics for our corpus

documents	11089
plagiarism cases	11603
Document purpose	
languages	fa
source documents	48%
suspicious documents	
with plagiarism	28%
w/o plagiarism	24%
Document length	
short (<10 pages[7])	64%
medium (10-100 pages)	35%
long (>100 pages)	1%
Plagiarism per document	
hardly (<20%)	25%
medium (20%-50%)	20%
much (50%-80%)	26%
entirely (>80%)	29%
Case length	
short (<1k characters)	37%
medium (1k-3k characters)	55%
long (>3k characters)	8%
Obfuscation synthesis approaches	
Exact Copy	8%
Near Copy	12%
Modified Copy	12%
Paraphrasing	16%
Summary	13%
Manual Translation	18%
Automatic Translation	8%
Cyclic Translation	10%
semantic-based meaning	3%

7. REFERENCES

[1] Asghari, H., Mohtaj, S., Fatemi, O., Faili, H., Rosso, P., and Potthast, M., 2016. Algorithms and Corpora for Persian Plagiarism Detection: Overview of PAN at FIRE 2016. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[2] Alzahrani, M., Salim, N. and Abraham, A. 2012. Understanding plagiarism linguistic patterns, Textual features, and detection Methods. *IEEE Trans. SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS*, vol. 42, no. 2.

[3] Potthast, M., Stein, B. and et.al. 2010. An Evaluation Framework for Plagiarism Detection. *Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010 Beijing,_c ACL*.

[4] Shamsfard, M., and Kiani, S., and Shahedi, Y. *STeP-1: Standard Text Preparation for Persian Language*. *CAASL3* Third Workshop on Computational Approaches to Arabic Script- Languages.

[5] Makrehchi, M. and Kamel, M. 2004. A fuzzy set approach to extracting keywords from abstracts. *North American Fuzzy Information Processing Society- NAFIPS 2003*, Banf, Canada.

[6] Shafiee, F. and Shamsfard, M. 2015. The automatic Persian summarizer. *The 20st Computer Society of Iran Computer Conference*.

[7] Adelkhah, R., Sadeghi, R. and Shamsfard, M. 2016. Persian sentence understanding and generation: a mutual conversion. *The 21st Computer Society of Iran Computer Conference*.

[8] Potthast, M., Göring, S. and et.al. 2015. Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. *Working Notes Papers of the CLEF 2015 Evaluation Labs, CEUR Workshop Proceedings*, (September 2015), ISSN 1613-0073.

[9] Potthast, M., Hagen, M., Gollub, T. and et.al. 2013. Overview of the 5th International Competition on Plagiarism Detection", *Working Notes Papers of the CLEF 2013Evaluation Labs and Workshop*, (September 2013), ISBN 978-88-904810-3-1.

[10] Manku, G. S., Jain, A. and Sarma, A. D. 2007. Detecting NearDuplicates for Web Crawling. *Data mining*.

[11] Kamran, K., Ahmadi, A. and Kazemivanhari, F. 2013. Plagiarism detection in Persian text using Fingerprint algorithms. *The 21st Iranian Conference on Electrical Engineering*.

[12] Davarpanah, M. R., sanji, M. and Aramideh, M. 2009. Farsi Lexical Analysis and StopWord List. *Library Hi Tech*, vol. 27, pp 435–449.

[13] Iran Telecommunication Research Center (ITRC), 2013. Buali Sina University. http://217.218.62.234:8080/.

[14] Shamsfard, M., Hesabi, A., Fadaei H. and et.al 2010. Semi Automatic Development of FarsNet; The Persian WordNet. *Proceedings of 5th Global WordNet Conference*.

[15] Mashhadirajab, F. and Shamsfard, M. 2014. *Plagiarism Detection in Persian documents.* Master's thesis. Shahid Beheshti University.

[16] Potthast, M., Eiselt, A and et.al. 2011. Overview of the 3rd International Competition on Plagiarism Detection. *Notebook Papers of CLEF 2011 Labs and Workshops*, (September 2011), ISBN 978-88-904810-1-7.

[17] Potthast, M., Gollub, T. and et.al. 2012. Overview of the 4th International Competition on Plagiarism Detection. *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, (September 2012), ISBN 978-88-904810-3-1.

[18] Potthast, M., Hagen, M. and et.al. 2014. Overview of the 6th International Competition on Plagiarism Detection. *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, (September 2014).

[19] Joy, M. S., Sinclair, J. E. and et.al. 2013. Student perspectives on source-code plagiarism. *International Journal for Educational Integrity*, Vol. 9, No. 1, pp. 3–19.

[7] A page is measured as 1500 chars.

[20] Joy, M. S., Cosma, G. and et.al. 2009. A TAXONOMY OF PLAGIARISM IN COMPUTER SCIENCE. *Proceedings of EDULEARN09 Conference,* (July 2009), ISBN: 978-84-612-9802-0.

[21] Naik, R. R., Landge, M. B., Mahender, C. N. and et.al 2015. A Review on Plagiarism Detection Tools. *International Journal of Computer Applications,* vol. 125 – No.11.

[22] Alvi, F., Stevenson, M., Clough, P. and et.al 2015. The short stories corpus. *Notebook for PAN at CLEF*.

[23] Cheema, W., Najib, F., Ahmed, S. and et.al 2015. A corpus for analyzing text reuse by people of different groups. *Notebook for PAN at CLEF*.

[24] Hanif, I., Nawab, A., Arbab, A. and et.al 2015. Cross-language urdu-english (clue) text alignment corpus. *Notebook for PAN at CLEF*.

[25] Kong, L., Lu, Z., Han, Y. and et.al 2015. Source retrieval and text alignment corpus construction for plagiarism detection. *Notebook for PAN at CLEF*.

[26] Khoshnavataher, K., Zarrabi, V., Mohtaj, S. and Asghari, H. 2015. Developing monolingual Persian corpus for extrinsic plagiarism detection using artificial obfuscation. *Notebook for PAN at CLEF*.

[27] Asghari, H., Khoshnavataher, K., Fatemi, O. and Faili, H. 2015. Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus. *Notebook for PAN at CLEF*.

[28] Mohtaj, S., Asghari, H. and Zarrabi, V. 2015. Developing monolingual english corpus for plagiarism detection using human annotated paraphrase corpus. *Notebook for PAN at CLEF*.

[29] Palkovskii, Y. and Belov, A. 2015. Submission to the 7th international competition on plagiarism detection. http://www.uni-weimar.de/medien/webis/events/pan-15, http://www.clef-initiative.eu/publication/working-notes, From the Zhytomyr State University and SkyLine LLC.