From English to Persian: Conversion of Text Alignment for Plagiarism Detection

Lee Gillam
Department of Computer Science
University of Surrey
UK
l.gillam@surrey.ac.uk

Anna Vartapetiance
Department of Computer Science
University of Surrey
UK
a.vartapetiance@surrey.ac.uk

ABSTRACT
This paper briefly describes the approach taken to Persian Plagiarism Detection based on modification to the approach used for PAN between 2011 and 2014 in order to adapt to Persian. This effort has offered us the opportunity to evaluate detection performance for the same approach with another language. A key part of the motivation remains that of undertaking plagiarism detection in such a way as to make it highly unlikely that the content being matched against could be determined based on the matches made, and hence to allow for privacy.

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

Keywords
Plagiarism detection; text alignment; Persian; PAN

1. INTRODUCTION
Detection of plagiarism has been shown to be beneficial to both education and research, ensuring that students and researchers alike are demonstrating their own understanding and findings. The same techniques can be used against the document archives of an organization to improve how such information is managed. And we have successfully demonstrated how related techniques can be used to protect by preventing the accidental propagation of corporate information deemed to be of high value in an Innovate UK project on Intellectual Property Protecting Cloud Services in Supply Chains (IPCRESS) collaboratively with Jaguar Land Rover and GeoLang Ltd [1]. This work was grounded in our previous PAN efforts, e.g. [2], undertaken with respect to finding matching text without, at the time of the match, directly using the textual content itself (e.g. n-grams) or using patterns as could be uniquely resolved to the textual content. The approach [3] produces a minimal representation of an input text by distinguishing content words from auxiliary words, and producing matchable binary patterns directly from these dependent on the number of classes of interest in each. This acts like hashing, but no effort is taken to ensure collision-avoidance; indeed, the approach actively encourages hash collision over short distances, as acts to prevent reverse-engineering of the patterns, and uses the number of coincident matches to indicate the extent of similarity. Such an approach is, therefore, more suited for longer initial pattern matching. Further, with the intention to undertake match without access to the textual content, there is a need for subsequent verification of potential matches based on access to the text, which can be undertaken automatically here, but would be anticipated to involve delegation of permissions within the kinds of system envisaged.

In PAN 11, this approach gained 4th place, with PlagDet=0.2467329, Recall=0.1500480, Precision=0.7106536, Granularity=1.0058894. In 2012, we showed good granularity, with high recall and precision for non-obfuscated text, but not such great recall when faced with higher orders of obfuscation, and subsequent results are consistent or slightly better.

In this paper, we assess our efforts against texts in Persian, with some commentary on the effects of the make-up of the dataset. Section 2 presents findings in respect to the training data of Persian Plagdet, and Section 3 addresses the test data. Section 4 concludes the paper and considers future work.

2. TRAINING DATA

The Persian Plagdet training data, dated 2016-07-17 comprises 1563 source documents and 1525 suspicious documents. There are some 2749 associated annotations, the breakdown of which is shown below in Table 1.

Table 1. Breakdown of 2749 potential cases of plagiarism within training data.

01-no-plagiarism	783	28%
02-no-obfuscation	208	8%
03-random-obfuscation	1611	59%
04-simulated-obfuscation	147	5%
Total files	2749	

It is striking that there is such a significant minority of unobfuscated cases, in contrast to 59% 'random'. A high-performing system must therefore be able to address the nature of this randomness, and this was uncovered during training as it seemed difficult to push performance beyond a certain point.

Following re-implementation of our approach in Python for ease of inspection (and necessary upgrading to Python 3.4 so that wide characters were handled properly), and a switch to a Persian stoplist[1], initial tests were conducted to ascertain performance against the training data.

Key parameters for the system are:

1. length used for match (windowSize, measured in words)
2. distance within which two non-overlapping matches can be merged (merge_dist, measured in characters)

[1] Obtained from: https://github.com/kharazi/persian-stopwords/blob/master/persian, comprising some 778 words. An alternative of just 330 words was also considered but not used: http://www.ranks.nl/stopwords/persian

3. minimum proportion of shared words in matched segments (match_threshold), required due to the methodology used as verification of match
4. minimum length of a final match (min_length), used later to filter out short matches.

Brute force efforts were used to obtain approximate values for parameters 1-3, with:

1. windowSize from 15 to 35, in increments of 5
2. merge_dist from 100 to 225 in increments of 25, and then honed in increments of 5
3. match_threshold from 0.5 to 0.95 in increments of 5

The best initial combination was windowSize=30, merge_dist=210, match_threshold=0.75, producing the following:

Table 2. Best scores on training data following initial brute force determination.

Plagdet Score	0.1894
Recall	0.1055
Precision	0.9542
Granularity	1.0043

Following the investigation into the make-up of the training data, explained above, the performance against each subset was evaluated. Initially, we determine the detection capability with match_threshold=0 (second column), i.e. without content checking, then look at each subset with checking (subsequent columns), as shown in Table 3.

Such scores demonstrate acceptable performance for the approach used, which is geared towards copying with low levels of obfuscation, as even without subsequent checking of matches recall is very high for non-obfuscated text. The low recall for obfuscation therefore leads us to explore the nature of the kind of obfuscation in use, and here the use of so-called random obfuscation within the training data certainly merits discussion.

To begin with, if passages are constructed which would not be meaningful within the language, there would seem to be limited gain from its treatment here, as there is a question over how this reflects the reality of the problem being addressed. In addition, if the extent of change is high as would require significant human effort to reproduce, the likelihood of such highly edited passages in the wild would seem to be lessened unless approaches are partially automated and do not undergo post-editing. It also becomes difficult to address the difference between necessary inclusion within a focused discussion, and an act of deliberate copying.

Table 3. Breakdown of scoring against specific subsets of the training data.

	02 before content check	02	03	04
Plagdet Score	0.5740	0.9050	0.0031	0.0426
Recall	0.9817	0.9729	0.0016	0.0357
Precision	0.4055	0.8461	0.0476	0.0605
Granularity	1.0000	1.0000	1.0000	1.0769
% cases	8%	8%	59%	5%

Consider the example passages below with matches between source and suspicious, with colour used to identify those passages shared. These passages are not fully matchable: lowermost in the table are two matched sub-passages where the maximal fragment length is 5, and the sub-matches are mostly smaller. Clearly, when the initial extents are short this will favour those approaches that address short n-grams. In addition, these passages include extents of text in that are quite different between the two, and this brings implications to the treatment of the gap between passages as well as to any verification step as addresses word overlap.

Table 4. Matching segments in random obfuscation, within annotations provided for the training data. Highlighted parts lowermost show differences between matched segments, as assists in identifying favourable n-gram sizes for matching.

Suspicious

ه عهدآبخو ، گنج ، کدخدا و همه خروس ، کاه است ، ساختمان اصلی نادرستیست کتاب شیعی تشکیل قصیده دادهاند .
ابراهیم گلستان در اعتراض کردن اسم غیرقانونی تیغ قندک و جراحی شده کتابش به جمهوری اسلامی ایران دو کار مینویسد . اولین بیشیلهپیله نامه طفره اندر
تپه تاریخ ۳۱ تیر قمر ۱۳۷۴ خطاب به مدیر است یا کتاب در آن یغما شده نوشته_شده_است ، شرمآور انتشاراتیای ، اگر خواروبارفروشی برای شرم کهنسال .
این اقدام شما . چاپ نامه این . من وقتم را از تلف نمیکنم معصیتبار زاویه که از منافی علاقه مندی کسانی چنین کردهاند شکایتی اشغال به را ماخذ به اقامتگاه
ببرم کاه چنین مقام و مرجع را اگر هم باشد ، نمیشناسم ... بهتر است این باشد کش رفتنها و مثله کردنها بماند برقرار شخ برای عملهٔ شک

Source

.جزیره، گنج، کدخدا و از همه مهمتر خروس، که مظهر بیداری است، ساختارهای اصلی این کتاب را تشکیل دادهاند
ابراهیم گلستان در اعتراض به چاپ غیرقانونی و جراحی شده کتابش در ایران دو نامه مینویسد. اولین نامه در تاریخ ۳۱ تیر ماه ۱۳۷۴ خطاب به مدیر
:انتشاراتیای که کتاب در آن چاپ شده نوشته شده است
شرمآور است، اگر برای شرم معنایی باقیمانده باشد. این اقدام شما نادرست بوده است. اسم این کار نادرستیست. من وقتم را بیش از تلف نمیکنم که از کسانی که
چنین کردهاند شکایتی به مرجعی یا به مقامی ببرم، که چنین مقام و مرجع را اگر هم باشد، نمیشناسم ... بهتر است این کش رفتنها و مُثله کردنها بماند برای
.عملهٔ شکنجه

گنج	کدخدا	و	***	همه	***	خروس																
گنج	کدخدا	و	از	همه	مهمتر	خروس																
تاریخ	۳۱	تیر	قمر	۱۳۷۴	خطاب	به	مدیر	است	یا	کتاب	در	آن	یغما	شده	نوشته	شده	است	شرمآور				
تاریخ	۳۱	تیر	ماه	۱۳۷۴	خطاب	به	مدیر	انتشاراتی	که	کتاب	در	آن	چاپ	شده	نوشته	شده	است	شرمآور				

Assuming that the test data would be formulated similarly to the training data, peak performance would be somewhat constrained, and with time available, further tests were performed using rather shorter initial word numbers (windowSize values) and lower values for other parameters.

Table 5. Example scores from further brute force determination, showing increases in recall for smaller windowSize, at the cost of drops in precision and granularity.

merge_dist	50	50	50
windowSize	20	10	8
match_threshold	0.35	0.35	0.35
min_length	200	200	200
Plagdet Score	0.1478	0.3669	0.4231
Recall	0.0859	0.3055	0.4568
Precision	0.9514	0.8493	0.7586
Granularity	1.0935	1.3370	1.5453

3. TEST DATA

From Table 4, above, values used for the submission based on test data are in the final column. Results from all participants are shown below in Table 5. Although only achieving 8th place of 9, with all other participants are from Iranian institutions, we are satisfied that we have managed to maintain the core of our approach, which we were already aware was only robust to a certain extent of obfuscation and is not readily tuned to random effects as may not necessarily be readable unless via a codebook.

Table 6. Results of all participants.

Rank	Plagdet	Granularity	Precision	Recall
1	0.92204	1.00146	0.92688	0.91919
2	0.90593	1	0.95927	0.85820
3	0.87103	1	0.89258	0.85049
4	0.83015	1.03968	0.92034	0.79602
5	0.80083	1.0	0.93337	0.70124
6	0.77496	1.22759	0.96383	0.83615
7	0.72662	1	0.74962	0.70499
8	**0.39968**	**1.52803**	**0.75484**	**0.41407**
9	0.38994	3.53698	0.90002	0.80659

4. CONCLUSIONS

In this paper, we briefly described the approach taken to Persian Plagiarism Detection, based on modification to the approach used for PAN between 2011 and 2014. Detection performance for Persian is appropriate with respect to the nature of application we have in mind, and the large proportion of randomly obfuscated data, allied to the manner in which obfuscation is conducted, limits what our approach would achieve. Runtime performance is also inadequate, however this is believed due in significant part to using a Python implementation instead of our main C++ codebase – and we are aware that some suggest Python 3 is significantly slower than C++ for the majority of tasks [4]. Evaluating a standard n-gram approach, via C++, would be expected to improve detection performance against these data.

It is worth noting that the first author has no familiarity with Persian languages, and only sought the co-author's advice on reasons for the detection performance in respect to random obfuscation after it was apparent that an improved plagiarism detection score could not be achieved.

5. ACKNOWLEDGMENTS

The authors gratefully recognize prior contributions of Neil Cooke, Scott Notley, Peter Wrobel, Neil Newbold and Henry Cooke in respect to the approach in previous competitions, and by Neil Cooke and Peter Wrobel to the patents generated from these. We recognize, also, prior support from the EPSRC and JISC (EP/I034408/1) and by Innovate UK (TSB, 169201). The authors are also grateful for the efforts of the Persian Plagdet organizers in formulating proceedings [5] and for system and data provision [6], [7], [8].

6. REFERENCES

[1] Gillam, L., Notley, S., Broome, S. and Garside, D. 2015 IPCRESS: Tracking Intellectual Property through Supply Chains in Clouds. In Raghavendra Rao, N. (ed.) *Enterprise Management Strategies in the Era of Cloud Computing*. IGI-Global.

[2] Cooke, N., Gillam, L., Wrobel, P. Cooke, H. and Al-Obaidli, F. 2011 "A high performance plagiarism detection system". *Proceedings of the 3rd PAN workshop*.

[3] Cooke, N and Gillam, L. 2012. System, process and method for the detection of common content in multiple documents in an electronic system. U.S. Patent filing US13/307,428, filed 30th November 2011.

[4] The Computer Language Benchmarks Game: http://benchmarksgame.alioth.debian.org/u64q/compare.php?lang=python3&lang2=gpp

[5] 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.

[6] Potthast, M., Stein, B., Barrón-Cedeño, A. and Rosso, P. 2010. An Evaluation Framework for Plagiarism Detection. In Huang, C-R and Jurafsky, D. (eds.), 23rd International Conference on Computational Linguistics (COLING 10), pages 997-1005, Stroudsburg, Pennsylvania, August 2010.Association for Computational Linguistics.

[7] Gollub, T., Stein, B. and Burrows, S.. 2012. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In Hersh, B., Callan, J., Maarek, Y. and Sanderson. M. (eds.), 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12), pages 1125-1126, August 2012. ACM. ISBN 978-1-4503-1472-5.

[8] Potthast, M., Gollub, T., Rangel,, F., Rosso, P., Stamatatos, E. and Stein, B.. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Kanoulas, E. et al, editors, Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14), pages 268-299, Berlin Heidelberg New York, September 2014. Springer. ISBN 978-3-319-11381-4.