PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition

Alberto Barrón-Cedeño[1], Paolo Rosso[1], Sobha Lalitha Devi[2], Paul Clough[3], and Mark Stevenson[3]

[1] NLE Lab - ELiRF, Universidad Politécnica de Valencia, Spain
[2] AU-KBC Research Centre, Chennai, India
[3] University of Sheffield, UK
{lbarron, prosso}@dsic.upv.es sobha@au-kbc.org
p.d.clough@sheffield.ac.uk M.Stevenson@dcs.shef.ac.uk

Abstract. The development of models for automatic detection of text re-use and plagiarism across languages has received increasing attention in the last years. However, the lack of an evaluation framework composed of annotated datasets has caused these efforts to be isolated. In this paper we present the CL!TR 2011 corpus, the first manually created corpus for the analysis of cross-language text re-use between English and Hindi. The corpus was used during the Cross-Language !ndian Text Re-Use Detection Competition. Here we overview the approaches applied the contestants and evaluate their quality when detecting a re-used text together with its source.

1 Introduction

Text re-use occurs when pre-existing written material is consciously used again during the creation of a new text or version [11, 6]. This might include the re-use of an entire text (e.g. duplicate web pages), or smaller segments (e.g. chunks, paragraphs and sentences) from one or more existing texts. Plagiarism, perhaps the most widely known example of text re-use, can be defined as "the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source" [17]. The problem has received attention from various research areas and even generated new terms such as *copy-paste syndrome* [29, 18] and *cyberplagiarism* [12]. The increased availability and accessibility of content online (e.g. texts, images, videos and sounds) is making text re-use easier than ever before and subsequently the automatic detection of text re-use, and in particular plagiarism detection, is of considerable importance.[4]

Recent efforts have focused on developing datasets with which to evaluate text re-use and plagiarism detection. The PAN International Competition on Plagiarism Detection (PAN@CLEF) [27, 24][5], held in conjunction with the

[4] See [9, 10, 21, 24] for an overview of the state of the art in automatic plagiarism detection.
[5] http://pan.webis.de

Cross-Language Evaluation Forum (CLEF), is perhaps the most widely-known example. Benchmarks that have been developed for the PAN competitions include corpora containing examples of automatically generated and simulated plagiarism[6] and evaluation metrics [26]. As a result, for the first time it has been possible to objectively evaluate and compare diverse methods for plagiarism detection.

The CL!TR@FIRE track focuses on *cross-language text re-use*, a more specific form of text re-use.[7] In the cross-language text re-use scenario the re-used text fragment and its source(s) are written in different languages, making the detection of re-use harder than when both texts are in the same language. Cross-language text re-use is an emerging research area that has begun to receive attention in recent years [5, 7, 19, 25]. There are various motivations for this interest: (*i*) speakers of under-resourced languages [4] are often forced to consult documentation in a foreign language; (*ii*) people immersed in a foreign country can still consult material written in their native language and (*iii*) cross-language text re-use, and in particular plagiarism, is becoming a problem. However, benchmarks are needed to assist in the development and evaluation of methods for detecting cross-language text re-use. The Cross-Language !ndian Text Re-Use detection task (CL!TR)[8] at FIRE addresses this issue.

The CL!TR@FIRE track has focussed on the re-use of Wikipedia articles as they are often a preferred source for plagiarised examples [16, 20]. Therefore, the collection of potential sources for a given case of re-use in CL!TR is composed of Wikipedia articles on several topics, including computer science and tourism (the latter from Incredible !ndia).

2 Corpus

A set of potentially re-used documents written in Hindi, D_{hi}, and a set of potential source documents written in English, D_{en}, were provided to participants. The documents in D_{hi} are those probable cases where text re-use has occurred. D_{en} included a total of 5,032 Wikipedia articles written in English; D_{hi} a total of 388 documents, written in Hindi (shown in Table 1).

The documents in D_{hi} are potentially re-used from the documents in D_{en}, however the languages are different and therefore makes the detection of text re-use a more difficult task. All of them were manually created. The topics included are computer science and tourism.

[6] Automatically generated plagiarism is created without any human involvement by altering a source text automatically, for example by deleting words or replacing them with equivalent terms (e.g. synonyms). Simulated plagiarism is generated manually by asking people to re-use text. PAN used automatically generated and simulated examples of plagiarism since cases of true plagiarism, where the writer has re-used text with the intent of claiming authorship, are difficult to identify and distribute.

[7] http://www.dsic.upv.es/grupos/nle/fire-workshop-clitr.html

[8] The name of our initiative is partially inspired by the *Incredible !ndia* campaign name (http://www.incredibleindia.org/).

Table 1. CL!TR 2011 corpus statistics. The figures are shown for the two sets D_{en} and D_{hi}. The column headers stand for: $|D|$ number of documents in the corpus (partition), $|D_{tokens}|$ total number of tokens, $|D_{types}|$ total number of types. k= thousand, M = million.

Partition	$\|D\|$	$\|D_{tokens}\|$	$\|D_{types}\|$
D_{hi}	388	216 k	5 k
D_{en}	5,032	9.3 M	644 k

2.1 Generating Cases of Plagiarism

A set of simulated plagiarised documents was created using an approach based on the one described by [8]. Participants were provided with a set of questions and asked to write a short answer, either by re-using text from a source provided (Wikipedia) or by looking at learning material (e.g. textbook, lecture notes, or websites). To simulate different degrees of obfuscation participants were asked to use one of four methods to write the answer:

Near copy Participants were asked to answer the question by simply copying text from the relevant Wikipedia article (i.e. performing cut–and–paste actions). No instructions were given about which parts of the article to copy (selection had to be performed to produce a short answer of the required length, 200-300 words). In this case using automatic translation was mandatory.

Light revision Participants were asked to base their answer on text found in the Wikipedia article and were, once again, given no instructions about which parts of the article to copy. They were instructed that they could alter the text in some basic ways including substituting words and phrases with synonyms and altering the grammatical structure (i.e. paraphrasing). Participants were also instructed not to radically alter the order of information found in sentences. Participants were allowed to use automatic translators.

Heavy revision Participants were once again asked to base their answer on the relevant Wikipedia article but were instructed to rephrase the text to generate an answer with the same meaning as the source text, but expressed using different words and structure. This could include splitting source sentences into one or more individual sentences, or combining more than one source sentence into a single sentence. No constraints were placed on how the text could be altered. Participants were not allowed to use automatic translation.

Non-plagiarism Participants were provided with learning materials in the form of either lecture notes, sections from textbooks, or web pages from Incredible !ndia that could be used to answer the relevant question. Participants were asked to read these materials and then attempt to answer the question using their own knowledge (including what they had learned from the materials provided). They were also told that they could look at other materials to answer the question but explicitly instructed not to look at Wikipedia.

Table 2. CL!TR 2011 potentially re-used documents distribution.

Training partition		Test Partition	
Re-used	130	Re-used	146
– Light revision	30	– Light revision	69
– Heavy revision	55	– Heavy revision	43
– Exact copy	45	– Exact copy	34
Original	68	Original	44
Total	198	Total	190

The first three methods are designed to generate examples of simulated-plagiarism in which the source text has been obfuscated to different levels. The final method, Non-plagiarism, generates answers which are not plagiarised to be used for comparison. This approach was originally developed for the creation of a monolingual corpus of simulated plagiarism [8]: the sources (i.e. Wikipedia articles and learning materials) were in English and participants were asked to write answers in English. The approach was adapted for CL!TR@FIRE to create a cross-lingual version: participants were provided with source text in English and asked to provide answers in Hindi. Volunteers were allowed to use automatic translators when generating some of the cases, either modifying the resulting translation or not.

The corpus was divided into training and test partition. In both partitions the collection of Wikipedia articles (D_{en}) is the same one. The collection D_{hi} was divided into two sub-collections. The training partition was composed of 198 documents, whereas the test partition contained 190 documents. The distribution of simulated-plagiarism and original documents in D_{hi} is shown in Table 2.

3 Proposed Task

The focus of CL!TR is on cross-language text re-use detection. This year we target two languages: Hindi and English. The potentially re-used documents are all written in Hindi, whereas the potential source documents are written in English (cf. Section 2).

The task is to identify those documents in D_{hi} that were created by re-using fragments from a document $d \in D_{en}$. It can be described as follows:

> Let D_{en} be a collection of documents (Wikipedia articles). Let $d_q \in D_{hi}$ be a re-used document. Given d_q, retrieve those documents $d \in D_{en}$ that are likely source texts of d_q. Afterwards determine whether the pair $p(d_q, d)$ compose a case of re-use together with its source.

This is a document level task; no specific fragments inside of the documents are expected to be identified. Determining either a text has been re-used from its corresponding source is enough. Specifying the kind of re-use (Exact, Heavy, or Light) was not necessary.

For the training phase we provided an annotated corpus. The actual cases of re-use (re-used and source document) were labelled, as well as the specific kind of re-use they composed. During the test phase no annotation or hints about the cases were provided.

4 Submissions Overview

Six teams from five different countries (India, Spain, Ireland, Hong Kong, and Ukraine) participated in the competition. They were allowed to submit up to three runs in order to encourage them to considering different approaches or parameters. A total of 15 text re-use detection runs were submitted.

Most of the participants opted for a "traditional" cross-language information retrieval approach. They translated the suspicious documents in D_{hi} into English in order to perform a monolingual similarity estimation [3, 14, 15, 22, 28]. Most of these approaches exploit the Google or Bing translation services.

The prototypical —information retrieval— process that follows the language normalisation is as follows. D_{en} is indexed into a search engine (most of the participants use Nutch/Lucene) and a document d_{hi} is queried to the search engine in order to retrieve the most similar documents $d \in D_{en}$.

Following we describe the different particularities of the information retrieval process followed in three approaches.

[3] does not perform any no pre-processing to the documents in D_{en}, which are directly submitted to the index. Afterwards, the documents d_{hi} are queried to the index and the most relevant retrieved document is considered a candidate of being the source of d_{hi}.

[14] splits the documents in D_{en} into paragraphs and expands their vocabulary on the basis of WordNet relationships (hyponyms, hypernyms and synsets). The enriched representation of each paragraph is fed to the index. The sentences in a d_{hi} are queried to the index and the top 10 source paragraphs are retrieved. The best matches are considered in order to select pairs of re-used and source (entire) documents.

[28] tried an information retrieval process for their run 3. After indexing D_{en}, key phrases were extracted from d_{hi} in order to independently query the index. The most frequently retrieved document $d_{en} \in D_{en}$ by the different key phrases in d_{hi} is selected as the source document.

Instead of translating the documents, [15] uses a bilingual dictionary in order to map Hindi to English words. Those words for which no possible translation exists in the dictionary are transliterated. Afterwards, a similarity estimation is carried out between the representations of d_{hi} and d_{en}. [15] submitted three runs proving a scale up set of settings: (*i*) for run 1, only dictionary based mapping is applied to d_{hi}; (*ii*) for run 2 mapping and transliteration are applied to d_{hi}; and (*iii*) for run 3, additionally to the mapping and transliteration processes, a minimal similarity threshold has to be surpassed in order to consider that d_{hi} is re-used from d_{en}.

Instead of assessing the similarity between the vocabulary in d_{hi} and d_{en}, [22] applies a fingerprinting model in order to detect exact string matches. After discarding non alpha-numeric characters, chunks of 5 words with a sliding window of 4 are hashed as in [23]. All the matches between d_{en} to d_{hi} are merged and used to estimate whether a case of re-use is at hand. The three runs of [22] consider different parameters for the fingerprinting process. The best settings are the just described.

Additionally to the aforementioned search engine-based approach of [28], this team tried two more techniques, based on machine learning. The model is based on a J48 decision tree classifier. For run 1 the features for the classifier were composed of the cosine similarity estimated over stemmed word 3-grams. For run 2 stopwords were removed and key phrases extracted. The relevance and length of the sequences compose the features for the classifier.

The approach of [2] is based on machine learning as well. This approach uses an SVM classifier considering features of statistical machine translation and sentence alignment models. The features for the classification process are three: (*i*) and (*ii*) are the score of the most likely alignments at sentence and paragraph level between d_{hi} and d_{en}, respectively. These scores were computed with the length based alignment algorithm proposed by [13]. (*iii*) is a lexical feature: A Hindi-English dictionary was used to gloss the Hindi documents and calculate an idf-based cosine similarity between suspicious and potential source documents.

5 Evaluation

The success of a text re-use detection model was measured in terms of Precision (P), Recall (R), and F_1-measure (F_1) —the harmonic mean of P and R— on detecting the re-used documents together with their source in the test corpus. A detection is considered correct if the re-used document d_{hi} is identified together with its corresponding source document d_{en}. For the P, R and F_1 computation, we consider three sets:

- *total detected* is the set of suspicious-source pairs detected by the system,
- *correctly detected* is the subset of pairs detected by the system which actually compose cases of re-use, and
- *total re-used* is the gold standard, which includes all those pairs which compose actually re-used cases.

P, R and F_1 are defined as follows:

$$P = \frac{\text{correctly detected}}{\text{total detected}} \qquad R = \frac{\text{correctly detected}}{\text{total re-used}} \qquad F_1\text{- measure} = \frac{2 \cdot R \cdot P}{R + P}$$

F_1-measure is used in order to compose the competition ranking. The evaluation results are presented in Table 3.

Table 3. Evaluation Results. Additionally to rank and evaluation of the runs, we show the number of suspicious documents identified as re-used.

Rank	F_1	R	P	Detections	Run	Ref.	Participant
1	0.649	0.750	0.571	147	3	[28]	Rambhoopal K.
2	0.609	0.821	0.484	190	1	[3]	N. Aggarwal
3	0.608	0.643	0.576	125	2	[28]	Rambhoopal K.
4	0.603	0.589	0.617	107	1	[22]	Y. Palkovskii
5	0.596	0.804	0.474	190	2	[15]	P. Gupta
6	0.589	0.795	0.468	190	2	[3]	N. Aggarwal
7	0.576	0.589	0.564	117	1	[28]	Rambhoopal K.
8	0.541	0.473	0.631	84	2	[22]	Y. Palkovskii
9	0.523	0.500	0.549	102	3	[22]	Y. Palkovskii
10	0.509	0.607	0.439	155	3	[15]	P. Gupta
11	0.430	0.580	0.342	190	1	[15]	P. Gupta
12	0.220	0.214	0.226	106	2	[14]	A. Ghosh
13	0.220	0.214	0.226	106	3	[14]	A. Ghosh
14	0.085	0.107	0.070	172	1	[14]	A. Ghosh
15	0.000	0.000	0.000	98	1	[2]	K. Addanki

5.1 Discussion

The most successful approaches for this task are based on standard cross-language information retrieval techniques. After translating the suspicious documents into English and building a search engine, [28] compose the queries by selecting a set of key phrases from the suspicious document. By means of this approach a good balance between recall and precision was obtained. The second best approach opts for a full representation of d_{hi} when generating the queries to the search engine [3]. The recall of this approach is the highest among all the participants (0.82), but with a high cost: a low precision value (0.484). The reason for this result is that [3] decided to assume that every document in D_{hi} was re-used and by simply retrieving the most similar document in D_{en} the problem would be solved. This assumption was made by [15] as well.

On the other side, the best precision values —with still competitive recall— are obtained by [22]. The reason may be that fingerprinting models are very strict to modifications and use to identify exact matches only.

The bad results obtained by the approach of [2] may be due to the nature of constructing re-use cases. As aforementioned, the documents in D_{hi} contain, in general, one single paragraph. For the re-used partition, this paragraph has been extracted from entire Wikipedia articles, causing the length factor to be less expressive (even if the length factors used are at sentence and paragraph level).

The remarkable results obtained by some of the approaches have to be read with caution. Most of them perform a language normalisation based on the Google translator. When generating the cases, the volunteers were allowed to

use this and other automatic tools to translate the contents they had selected to answer a given question and further modify it.

In order to make a more "realistic" task, the documents in D_{en} provided to the participants, coming from Wikipedia, included Wikitext annotation. Nevertheless none of the participants reports having performed any pre-processing to eliminate this noisy information.

6 Final Remarks

In this paper we presented an overview of the Cross-Language !ndian Text Re-Use Detection Competition. The challenge consisted of identifying, among a set of short documents written in Hindi, those texts that had been generated by re-use and the corresponding source document written in English.

Taking advantage of the first text collection of this nature, fifteen approaches were compared. Most of them were based on standard cross-language information retrieval and some other on statistical machine translation and machine learning techniques.

Acknowledgements

We would like to thank Pattabhi RK Rao and Mona Parakh from AU-KBC for their help in setting up the necessary resources for the competition. We also thank all the volunteers from the Indian School of Mines (ISM), Dhanbad, other institutions, and AU-KBC that helped to manually creating the re-use cases in the CL!TR corpus.

This research work is partially funded by the WIQ-EI (IRSES grant n. 269180) and ACCURAT (grant n. 248347) EC research projects. The research work of the first author is funded by the CONACyT-Mexico 192021 grant. The research of the second author is in the framework of the VLC/Campus Microcluster on Multimodal Interaction in Intelligent Systems and partially funded by the MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (plan I+D+i). The research from AU-KBC Centre is supported by the Cross Lingual Information Access (CLIA) Phase II Project.

References

1. FIRE 2011 Working Notes (2011)
2. Addanki, K., Wu, D.: An Evaluation of MT Alignment Baseline Approaches upon Cross-Lingual Plagiarism Detection. In: FIRE 2011 Working Notes [1]
3. Aggarwal, N., Asooja, K., Buitelaar, P.: Cross Lingual Text Reuse Detection Using Machine Translation & Similarity Measures. In: FIRE 2011 Working Notes [1]
4. Alegria, I., Forcada, M., Sarasola, K. (eds.): Proceedings of the SEPLN 2009 Workshop on Information Retrieval and Information Extraction for Less Resourced Languages. University of the Basque Country, Donostia, Basque Country (2009)

5. Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A.: On Cross-lingual Plagiarism Analysis Using a Statistical Model. In: Stein, B., Stamatatos, E., Koppel, M. (eds.) ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008). pp. 9–13. CEUR-WS.org, Patras, Greece (2008)
6. Bendersky, M., Croft, W.: Finding Text Reuse on the Web. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining. pp. 262–271. ACM (2009)
7. Ceska, Z., Toman, M., Jezek, K.: Multilingual Plagiarism Detection. In: Proceedings of the 13th International Conference on Artificial Intelligence. pp. 83–92. Springer Verlag, Berlin Heidelberg (2008)
8. Clough, P., Stevenson, M.: Developing a Corpus of Plagiarised Examples. Language Resources and Evaluation 45(1), 5–24 (2011)
9. Clough, P.: Plagiarism in Natural and Programming Languages: an Overview of Current Tools and Technologies. Research Memoranda: CS-00-05, Department of Computer Science. University of Sheffield, UK (2000)
10. Clough, P.: Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service (2003), http://ir.shef.ac.uk/cloughie/papers/pas plagiarism.pdf
11. Clough, P., Gaizauskas, R.: Corpora and Text Re-Use. In: Lüdeling, A., Kytö, M., McEnery, T. (eds.) Handbook of Corpus Linguistics, pp. 1249—1271. Handbooks of Linguistics and Communication Science, Mouton de Gruyter (2009)
12. Comas, R., Sureda, J.: Academic Cyberplagiarism: Tracing the Causes to reach Solutions. In: Comas, R., Sureda, J. (eds.) Academic Cyberplagiarism [online dossier], Digithum. Iss, vol. 10, pp. 1–6. UOC (2008), http://www.uoc.edu/digithum/10/dt/eng/cyberplagiarism.pdf
13. Gale, W., Church, K.: A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics 19, 75–102 (1993)
14. Ghosh, A., Pal, S., Bandyopadhyay, S.: Cross-Language Rext Re-Use Detection Using Information Retrieval. In: FIRE 2011 Working Notes [1]
15. Gupta, P., Singhal, K.: Mapping Hindi-English Text Re-use Document Pairs. In: FIRE 2011 Working Notes [1]
16. Head, A.: How today's college students use Wikipedia for course-related research (Mar 2010), http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2830/2476
17. IEEE: A plagiarism FAQ. http://www.ieee.org/web/publications/rights/plagiarism_FAQ.htm (2008), [Online; accessed 3-March-2010]
18. Kulathuramaiyer, N., Maurer, H.: Coping With the Copy-Paste-Syndrome. In: Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007 (E-Learn 2007). pp. 1072—1079. AACE, Quebec City, Canada (2007)
19. Lee, C., Wu, C., Yang, H.: A Platform Framework for Cross-lingual Text Relatedness Evaluation and Plagiarism Detection. In: Proceedings of the 3rd International Conference on Innovative Computing Information (ICICIC'08). IEEE Computer Society (2008)
20. Martínez, I.: Wikipedia usage by Mexican students. The constant usage of copy and paste. In: Wikimania 2009. Buenos Aires, Argentina (2009)
21. Maurer, H., Kappe, F., Zaka, B.: Plagiarism - A Survey. Journal of Universal Computer Science 12(8), 1050–1084 (2006)
22. Palkovskii, Y., Belov, A.: Exploring Cross Lingual Plagiarism Detection in Hindi-English with n-gram Fingerprinting and VSM based Similarity Detection. In: FIRE 2011 Working Notes [1]

23. Palkovskii, Y., Belov, A., Muzika, I.: Using WordNet-based Semantic Similarity Measurement in External Plagiarism Detection - Notebook for PAN at CLEF 2011. In: Petras, V., Forner, P., Clough, P. (eds.) Notebook Papers of CLEF 2011 LABs and Workshops. Amsterdam, The Netherlands (Sep 2011)
24. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler, M., Harman, D. (eds.) Notebook Papers of CLEF 2010 LABs and Workshops. Padua, Italy (Sep 2010)
25. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-Language Plagiarism Detection. Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis 45(1), 1–18 (2011)
26. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Huang, C.R., Jurafsky, D. (eds.) COLING 2010 Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 997–1005. Coling 2010 Organizing Committee, Beijing, China (August 2010)
27. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. pp. 1–9. CEUS-WS.org (2009), http://ceur-ws.org/Vol-502
28. Rambhoopal, K., Varma, V.: Cross-Lingual Text Reuse Detection Based On Keyphrase Extraction and Similarity Measures. In: FIRE 2011 Working Notes [1]
29. Weber, S.: Das Google-Copy-Paste-Syndrom. Wie Netzplagiate Ausbildung und Wissen gefahrden. Telepolis (2007)