An evolutionary survey from Monolingual Text Reuse to Cross Lingual Text Reuse in context to English-Hindi

Aarti Kumar*, Sujoy Das**

Abstract-With enormous amount of information in multiple languages available on the Web, mono and cross-language text reuse is occurring every day with increasing frequency. Near-duplicate document detection has been a major focus of researchers. Detecting cross-language text reuse is a very challenging task in itself and the challenge magnifies manifolds when it comes to translated, obfuscated and local text reuse. These difficulties and challenges are contributing to the most serious offence of plagiarising others' text. This paper presents an evolutionary overview of the various techniques being used to measure text reuse covering techniques for detecting reuse from mono-lingual to cross-lingual and from mono-script to cross-script with special emphasis on English-Hindi language pair.

Index terms- cross-lingual, cross-script, fingerprinting, mono-lingual, mono-script, obfuscated, pre-retrieval, TF-IDF, verbatim

1. INTRODUCTION

Web is flooded with large information of content that are easily accessible to the user. It prompts them to use it either in its original form or in paraphrased form for describing something that the user wants. The used content is referred as text reuse, plagiarism etc. It can also be referred as transformation of text to change its surface appearance. Duplicate or near duplicate document detection has been a major focus of researchers. Search engines needs to identify duplicate documents as they tend make these system less efficient because they consume considerable system resources [5]. Text reuse normally occurs when pre-existing texts or segments are used to create new once. It can be literal reuse of original sentences or reuse of facts and concepts, or it might be even reuse of style. Detecting literal uses may be easier to tackle if the contents are copied verbatim where as detecting facts, concepts or style is not a trivial problem to solve. Paul Clough [7] described text reuse as use of single or multiple number of known text sources either verbatim or otherwise in rewritten text. Detecting text reuse has got a vast application in different fields like automatic plagiarism detection, paraphrasing detection, detecting breach of copyright, news monitoring system etc.

Multilingual content are also proliferating on the web and due to this text reuse is now not limited to same language but has also crossed language boundary. The common text usage may translate the reused content and reproduce it either in a bit different style or with synonyms, antonyms etc. of that language. Therefore apart from the classification given by the authors reuse can also extend from mono-lingual to cross-lingual.

In this paper a survey is carried out to understand the different dimensions of research work that has been carried out to tackle the problem of text reuse. This paper traces the work of different authors in detecting text reuse from mono-lingual to cross-lingual and from cross-lingual mono-script to cross-lingual cross-script.

Rest of the paper is as follows: In Section 2 various types of text reuse is discussed, Section 3 discusses techniques used in detection of mono-lingual text reuse, section 4 discusses the techniques implemented in cross-lingual text reuse and Section 5 presents the concluding remarks.

*Corresponding Author.Research Scholar Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, India, E-mail: aartikumar01@gmail.com, Mob: +919303132828
**Associate Professor, Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, India, E-mail: sujdas@gmail.com, Mob: +919826345195

2. TYPES OF TEXT REUSE

In text reuse the modification can be at the level of words, phrases, sentences or even whole text by applying a random sequence of text operations such as change of tense, change of voice, shuffling a word or a group of words, deleting or inserting a word from an external source, or replacing a word with a synonym, antonym, hypernym or hyponym. The alterations normally should not modify the original meaning of the source text.

Based on the nature of the text [4],[6],[7] text reuse can be classified as (a)Verbatim or copy & paste : It is mostly falls in the category of direct and non-modified reuse and (b) Obfuscated/rewrite: In this the text is modified and its modified version is presented. The degree of obfuscation may low or high. The level of degree increases the complexity of reuse detection.

Jangwon Seo and W. Bruce Croft [5] identified six categories of reuses based on TREC newswire and blog collections. They are Most-Most, Most-Considerable, Most-Partial, Considerable-Considerable, Considerable-Partial, and Partial-Partial.

Researchers have classified text reuse based on authorship [8] as self reuse and cross reuse. In former author reuses his own work where as in latter someone else's work is reused. Categorizing text reuse as global and local is another perspective of looking at text reuse. In this either whole document has been reused i.e global reuse [3] or sentences, facts & passages have been reused and modified to produce local reuse [5]. Similar thing has been reported by Paul Clough et al. [6] in which newspaper articles has been classified as wholly, partially or non-derived based on degree of dependence upon, or derivation from.

Apart from this text reuse can be further classified based on the language of source and target document. It can be mono-lingual, cross-lingual or multilingual. The verbatim cross-lingual text reuse shall fall under the category of obfuscated text based on the level of translation. Level of obfuscation may also depend upon the quality of the translation. Cross-lingual reuse can have source and target documents in different languages but both these languages using the same script or both the language and the scripts of source document and target document may vary. The former can be classified as cross-lingual mono-script text reuse and later as cross-lingual-cross-script text reuse.

Although various tools and techniques are being used to detect reuse, still, cross-language text reuse detection has not been approached sufficiently due to its inherent complexity [28] whereas different methods for the detection of monolingual text reuse have been developed.

With so many languages spoken around the world, identifying cross language text reuse still remains a challenging task it becomes even tougher if one considers less resourced languages available around the world. Though few attempts have been made [20], [28],[29] by the researchers to tackle this problem. Fig. 1 gives a diagrammatic representation of various types of text reuse.

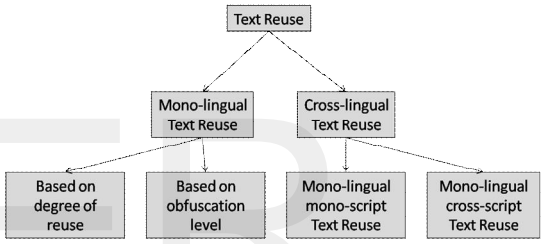

Fig.1 Types of text reuse

3. DETECTING MONO-LINGUAL TEXT REUSE

3.1 Techniques used to measure Verbatim Text Reuse

The detection of reuse in documents started with identifying verbatim reuse and was restricted to find the amount of words are similar in two documents.

The main technique for verbatim text reuse detection is to use document fingerprints [3],[5],[6],[7]. Fingerprints are the subset of hashed subsequences of words in documents called chunk or shingle, and are used to represent a document. Shared text is determined by finding containment of documents using containment ratio i.e. number of shared fingerprints that are common in the documents.

Another technique used for detecting verbatim reuse is the K-gram overlap method [3],[5],[6]. Normally a fixed window is defined and is slid over the source text to generate chunks and then fingerprints are compared. Number of fingerprints generated by using k-gram technique is enormous but it is than normal finger printing as more number of combinations can be compared. Approaches like

Winnowing[3],[5],[6], 0 mod p[3],[5],[6] and Hash Breaking[3],[5] are used to eliminate the insignificant fingerprints without losing the important ones.

Word ngram overlap measure finds shared text between Press Association articles and newspapers. To find overlap of words document ngrams are stored as unique entries as hash [7]. The value of the hash contains the number of occurrences of the ngram within the document.

Apart from fingerprinting and hashing approaches, [7] used a graphical approach called dot-plot to envisage patterns of word overlap between documents. The texts are split into ngrams and pairwise comparisons are made for all ngrams. A black dot is placed wherever a match exists. For example if the 7th ngram of one text matches the 9th in the other, a dot is placed at position (7, 9) in the dotplot. Ordered matching sequences appear as diagonal and unordered matches as square blocks of dots.

The main fingerprinting technique and its modified versions like k or n-gram for detecting text reuse fails in case of obfuscated text reuse, since the exact fingerprint no longer exists in the modified version of the text. The dot-plot approach appears successful in highlighting differences between derived and non-derived texts, and can also show the positions of word additions or deletions but may miss synonymous replacement of text.

Fingerprinting and hash-breaking is too sensitive to small modifications of text segments and are inefficient in terms of time and space complexity. As k-gram uses all chunks, it generally performs well but might be too high in context to time and space complexity.

3.2 Techniques used to measure Shuffled and obfuscated Text Reuse

Exact matching is not good for non-verbatim text reuse.Techniques devised for measuring verbatim text reuse normally does not performs well when word is reordered or shuffled or may be obfuscated with the use of synonyms, hypernyms or hyponyms.

Clough and Gaizauskas [6] proposed Greedy String Tiling technique in which substring is matched. It computes the degree of similarity between two strings and is able to deal with transposition of tokens. The GST algorithm performs a 1:1 matching of tokens between two strings and moves ahead with matching till a mismatch is found. The maximal length substrings which are matched from the other are called tiles. A minimum match length is used to avoid fake matches. But using overlapped and non-overlapped fingerprinting approach the same result can be obtained as GST. Another approach implemented for measuring obfuscated text reuse is cognate-based approach used by [6]. Here cognates are defined as pairs of terms that are identical, share the same stems, or are substitutable in the given context.

Whenever the content words are replaced by synonyms, string measures typically fail due to the vocabulary gap. Daniel Bar et al. [10] thus used similarity measures to capture semantic similarity between words. The document-level similarity is the average of applying this strategy in both directions, from source to target and vice-versa. Whereas the Cognate based approach could handle synonyms and word inflections, the directional similarity approach worked well in detecting semantic similarity between texts.

Maxim Mozgovoy [9] used the tokenization technique for measuring text reuse. In this technique the element names are substituted by the name of their class to which they belong. Like all numeric values can be replaced by its class signature “value”. In [9] the obvious difficulty concerns polysemantic words and homonyms. This technique seems to be the most advanced way of comparing structured documents, but the results in this direction are still very preliminary for any kind of evaluation. The tree matching procedure is still very experimental and Tokenization could produce many false positives because as per this technique “*Ram goes to Kashmir*” and “*Shyam comes from Rajasthan*” will be treated same because both these strings represent similar syntactic structure.

Researchers have tried to identify text reuse on the basis of concept of the document. [38] proposed Concept Map Knowledge Model based on this idea to find similarity among the non-verbatim documents. Creating concept map is a challenging task in itself. A very different text reuse detection technique based on the Semantic Role Labeling was introduced by Ahmed Hamza Osmana et al. [33]. They improved the similarity measure using argument weighting with an aim to study the argument behaviour and effect in plagiarism detection.

In text documents, the order in which words occur is an eminent aspect of the text's semantics in most of the languages. Few words always appear in association with some other word but change in their order might result either in a meaningless sentence or a sentence with changed semantics. Based on this assumption [3] proposed a fingerprinting algorithm called MiLe that utilizes the contiguity of documents and generates one fingerprint per document instead of a set of fingerprints.

Shivakumar and Garcia-Molina [7] designed a technique Stanford Copy Analysis Mechanism to detect plagiarism using a vector space model. In this the documents are compared using a variant of the cosine similarity measure. Not only content similarity, but also structural similarity, and stylistic similarity were used by [10] to measure text similarity. They used stopword n-grams, part of speech n-grams and word pair order to measure structural similarity.

The terms which appear only once in the document are known as hapaxlegomenon or hapax. Hapaxlegomena was used for measuring text reuse by [6],[9].

Many other authors have also worked upon automatic and local text reuse detection [5],[3] translation detection [37] and paraphrase detection [39] using similar techniques.

A few researchers worked on a subset of similar documents instead of processing whole corpora for similarity detection. They formulated efficient query formulation mechanism for such retrieval.

Bruno Possas et. al.[34] used data mining technique instead of syntactical and semantic techniques. They proposed association rules derive the Maximal Termsets. To select representative sub queries information of distributions is used and concept of maximal termsets is used for modelling.

Matthias Hagen and Benno Stein [32] also focused on query formulation problem as the crucial first step in the detection of text reuse and presented a strategy which achieves better results than maximal termset query.

These improved strategies worked well in case of mono-lingual text reuse but the question was to see whether these theory applies on cross-lingual as well? The answer lies in process of creating parallel corpora by converting the source language to target language and then comparing. The challenge is to devise techniques for detecting cross-lingual text reuse: both cross-lingual mono-script and cross-lingual cross-script.

4. MEASURING CROSS-LINGUAL TEXT REUSE

4.1 Measuring Cross-language Mono-scripts Text Reuse

An HMM-based approach for modelling word alignments in parallel texts in English and French was presented by Stephan Vogel et al.[36]. The characteristic feature of this approach is to make the alignment probabilities explicitly dependent on the alignment position of the previous word. Large jumps due to different word orderings in the two languages are successfully modelled using this approach.

Alberto Barr´on-Cede˜no et al. [16] compared the effectiveness of their approach with approach based on character n-grams and statistical translation. The language of their study is Basque, a less resourced language where cross language plagiarism is often committed from texts in Spanish and English.

Grozea and Popescu[31] evaluated cross-language similarity among suspected and original documents using a statistical model which finds the relevance probability between suspected and source document regardless of the order in which the terms appear in the suspected and original documents. Their method is combined with a dictionary corpus of text in English and Spanish to detect similarity in cross language.

While analysing European languages Bruno Pouliquen et al. [35] presented a system that identified translations and other similar documents among a large number of candidates, by representing the documents content with a vector of Thesaurus terms from multilingual thesaurus, and then by measuring the semantic similarity between the vectors.

Plagiarist commonly disguises academic misconduct by paraphrasing copied text instead of rearranging the citations, this motivated Bela Gipp et al.[15] to consider citation patterns instead of textual similarity for detecting text reuse. The technique is purely language independent.

4.2 Measuring Cross-language Cross-scripts Text Reuse

When it comes to measuring text reuse in cross-language cross-script, although a few more cross-script language have been studied but we focus on English –Hindi Language pair in this paper. This language draws our attention due to the fact that this is the language which is spoken by 4.46% of the world population and according to the number of native speakers, ranks fourth among the top ten languages of the world, following Mandarin, English and Spanish[1]. (Fig. 2)

Language	Native speakers (millions)	% of world population
Mandarin	935	14.1%
Spanish	387	5.85%
English	365	5.52%
Hindi	295	4.46%
Arabic	280	4.23%
Portuguese	204	3.08%
Bengali	202	3.05%
Russian	160	2.42%
Japanese	127	1.92%
Punjabi	96	1.44%

Fig. 2: Native speakers of top ten languages of world.

Identifying cross-language reuse in English-Hindi pair is a challenging as the scripts differs and Hindi stores information in morphemes where as English in positions of word also there is a vast distance between these two languages with regards to script, vocabulary and grammar. Being a low resource language, Hindi lacks properly developed translators and transliterators [28] to be translated to a parallel and comparable corpora and lot of challenges arise due to improper machine translation (Fig. 3) and transliteration (Fig.4).

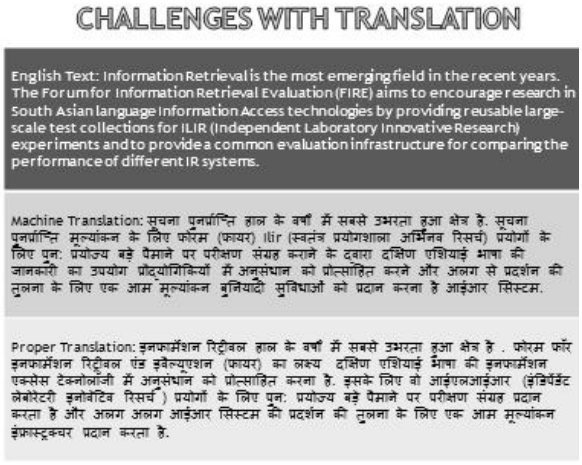

Fig 3. Mistranslated version of English to Hindi when Machine Translation is used

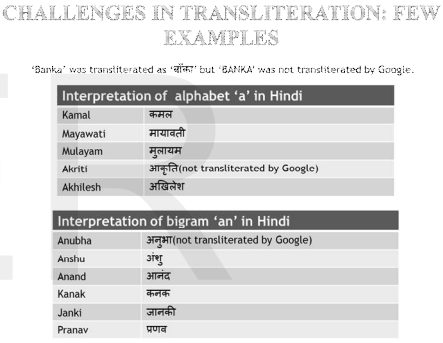

Fig. 4. Challenges in transliteration due to multi-interpretation of same unigrams and bigrams

Hindi also suffers from the fact that it has borrowed majority of its words from extremely inexhaustible vocabulary of the ancient languages Persian and Sanskrit[2]. Apart from these, it has also enriched its content with many loan words from other linguistic sources too. Forum for Information Retrieval and Evaluation (FIRE) has taken commendable initiative towards evaluation of South Asian languages. It provides reusable large-scale test collections for such languages and also provides a common evaluation infrastructure for comparing the performance of different IR systems for these. The work done towards detection of English-Hindi text reuse is, therefore, somewhat proportional to the tasks given by FIRE in the last five years.

[1] Source:http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[2] source: http://boards.straightdope.com/sdmb/showthread.php?t=623885

Towards cross-lingual and cross-script text reuse detection in English-Hindi language pair, Yurii Palkovskii and Alexei Belov [17] have used automatic language translation - Google Translate web service to translate one of the input texts to the other comparison language. Their ranking model includes six filters, each of which computes some similarity ranking points and the final score is a sum of all values. IDF, Reference Monotony and Extended Contextual N-grams IR Engine has been used by [26] to link English and Hindi News.

An unsupervised vector model approach and a supervised n-gram approach for computing semantic similarity between sentences were explored by [18]. Both approaches used WordNet to provide information about similarity between lexical items. Aniruddha Ghosh et al.[19] treated cross-language English-Hindi text re-use detection as a problem of Information Retrieval and have solved it with the help of WordNet, Google Translate, Lucene and Nutch, an open source Information Retrieval system. The uniqueness of their approach is that instead of using similarity score the dissimilarity score between each set of source and suspicious document is used for evaluation. n-gram Fingerprinting and VSM based Similarity Detection is used by [21] for Cross Lingual Plagiarism Detection in Hindi-English.

Aarti Kumar and Sujoy Das [28] used three pre-retrieval strategies for English-Hindi Cross Language News Story Search. They compared the performance of dictionary based approach with machine translation based approach with manual intervention.

Sujoy Das and Aarti Kumar [27] also compared the performance of dictionary based cross language information retrieval strategies for cross language English-Hindi news story search where the retrieval performance of short medium and long queries were evaluated. The simple strategies did not lead to good result but the strategies were able to capture text reuse across the language.

Parth Gupta and Khushboo Singhal [20] tried to see the impact of available resources like Bi-lingual Dictionary, WordNet and Transliteration mapping Hindi-English text reuse document pairs and used Okapi BM25 model to calculate the similarity between document pairs.

Prior to using Wikipedia-based Cross-Lingual Explicit Semantic Analysis, Nitish Aggarwal et al. [22] also performed heiuristic retrieval using publication date and vocabulary overlap to reduce the search space before applying their strategy.

To attain a very short and selective group of linked pairs instead of a long rank, enabling a very fast subsequent comparison, Torrejon et al.[26] used the High Accuracy Information Retrieval System engine, for indexing and selecting the best similar for every chunk of the Hindi translated versions of the English news, filtered by the reference monotony prune strategy to avoid chance matching.

Using the Lucene search engine identifying as many relevant documents as possible and then merging of document list followed by their re-ranking were the two-step procedure followed by Piyush Arora et al.[23] for measuring English-Hindi Journalistic text reuse.

Set-based Similarity Measurement and Ranking Model to Identify Cases of Journalistic Text Reuse is proposed by [24]. They compared the potential Hindi sources based on five features of the documents: title, the content of the article, unique words in content, frequent words in content, and publication date using Jaccard similarity.

GouthamTholpadi and AmoghParam[25] considered only those news stories pair which were published within a window of defined number of days around the date of publication of English news. Contrary to popular belief, they found that imposing date constraints did not improve precision.

All these techniques have been able to solve the problems of detecting cross-lingual cross-script text reuse detection in English-Hindi pair up to certain extent but a lot of work still needs to be done.

As per the analysis of the authors, Out of vocabulary words substitution, focus shifting, polysemy and phrasal handling are major problems in Hindi to be dealt with. The worst of all being the problem of identifying total rephrasing such as

a) Minister had already assured the House that all parties would be taken into confidence by the government on the issue.

b) महिला आरक्षण बिल पर सहमति कायम करने के लिए मंत्री ने सर्वदलीय बैठक बुलाई है।

Human brain can comprehend that these two are connected but it is difficult for automated strategies to treat the two as conceptually related text as obfuscation is multifold.

5. CONCLUSION

This paper presents an overview of the techniques applied to detect text reuse ranging from mono-lingual to cross-lingual and from cross-lingual mono-script to cross-lingual cross-script.

Success has been achieved in detecting verbatim reuse but techniques for detecting the use of synonyms, hypernym, and hyponym at the time of reuse needs further exploration.

Cross-lingual cross-script reuse detection especially in context of English-Hindi still needs manual interventions due to insufficient resources and requires further research to automate the process. Linguistically-motivated approaches to identify rewrites such as paraphrasing and obfuscation are still an open area for research.

ACKNOWLEDGMENT

One of the authors, Aarti Kumar, is grateful to her institution, Maulana Azad National Institute of Technology, Bhopal, India for providing her the financial support to pursue her Doctoral work as a full time research scholar.

REFERENCES

[1] C. D Manning., P. Raghavan andH. Schulz,*An Introduction to Information Retrieval*, Cambridge University Press

[2] R. Baeza-Yates andB. Ribeiro-Neto, *Modern Information Retrieval*, Pearson education

[3] A. Mittelbach, L. Lehmann, C. Rensing, and R. Steinmetz , "Automatic Detection of Local Reuse," *Proc. 5th European Conference on Technology Enhanced Learning*,no. LNCS 6383 p229-244 Springer Verlag, sep 2010 ISBN 3-642-16019-0

[4] Y. Palkovskii, I. Muzyka and A.Belov , "Detecting Text Reuse with Ranged Windowed TF-IDF Analysis Method," Available:http://www.plagiarismadvice.org/research-papers/item/detecting-text-reuse-with-ranged-windowed-tf-idf-analysis-method

[5] J. Seo and W. B. Croft, "Local Text Reuse Detection,"*SIGIR'08, July 20–24, 2008, Singapore.,*Copyright 2008 ACM 978-1-60558-164-4/08/07

[6] P.D.Clough, R. Gaizauskas, Scott S.L. Piao and Y. Wilks, "METER: MEasuringTExt Reuse*," Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 152-159.

[7] P. D. Clough," Measuring Text Reuse in Journalistic Domain," *http://ir.shef.ac.uk/cloughie/papers/cluk4.pdf*

[8] P. Gupta and P. Rosso, "Text Reuse with ACL(Upward)Trends,"*Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries,* pages 76–82,Jeju, Republic of Korea, 10 July 2012. c 2012 Association for Computational Linguistics

[9] M. Mozgovoy, V. Tusov and V. Klyuev, "The Use of Machine Semantics Analysis in Plagiarism Detection,"*http://web-ext.u-aizu.ac.jp/~mozgovoy/homepage/papers/mtk06.pdf*

[10] D. Bär, T. Zesch and I. Gurevych,"Text Reuse Detection Using a Composition of Text Similarity Measures" *https://www.ukp.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2012/COLING_2012_DaB_CameraReady.pdf*

[11] E. Barker andR.Gaizauskas," Assessing the comparability of news text,"*Proceedings of the Eigth International conference on Language Resources and Evaluation(LREC'12)*. European Language Resources Association(ELRA), Istanbul, Turkey(May 2012)

[12] P. Clough, "Measuring text reuse in a journalistic domain,"*Proc. 4th CLUK Colloquium*. pp. 53–63 (2001) DOI= http://ucrel.lancs.ac.uk/acl/P/P02/P02-1020.pdf

[13] M. Littman, S.T. Dumais andT. K. Landauer, "*Automatic cross-language information retrieval using latent semantic indexing*," In: Cross-Language Information Retrieval, chapter 5. pp. 51–62. Kluwer Academic Publishers (1998)

[14] S. Alzahrani, N. Salim and A. Abraham,"Understanding Plagiarism Linguistic Patterns, Textual Features and detection methods,"*IEEE Transactions on Systems, Man and Cybernatics-Part C: Applications and reviews, Vol. 42, No. 2, March 2012*

[15] B. Gipp, N. Meuschke, C. Breitinger, M.Lipsinki and A. Nurnberger, "Demonstration of Citation Pattern Analysis for Plagiarism Detection*," In: SIGIR'13, July 28-August 1, Dublin*, Ireland. ACM 978-1-4503-2034-4/13/07

[16] A. B. Cedeno, P. Rosso, E. Agirre and G.Labak, "*Plagiarism Detection across Distant Language Pairs*" DOI=http://delivery.acm.org/10.1145/1880000/1873786/p37-barron-cedeno.pdf?ip=14.139.241.84&id=1873786&acc=OPEN&key=045416EF4DDA69D9%2ECD3B1CD2041702DD%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=496177549&CFTOKEN=88742299&__acm__=1404212851_89aa6058922451631ae41dc9b5b48690

[17] Y. Palkovskii and A. Belov, "Using TF-IDF Weight Ranking Model in CLINSS as Effective Similarity Measure to Identify Cases of Journalistic Text Reuse," *Springer-Verlag Berlin Heidelberg 2011*

[18] S. Biggins, S. Mohammed, S. Oakley, L. Stringer, M. Stevenson and J. Priess, "Two Approaches to Semantic Text Similarity," *Proc. First Joint Conference on Lexical and Computational Semantics, pages 655-661, Montreal, Canada, June 7-8,2012*

[19] A. Ghosh, S. Pal and S. Bandyopadhyay, "Cross-Language Rext Re-Use Detection Using Information Retrieval," *In: FIRE 2011 Working Notes*

[20] P. Gupta and K. Singhal, "Mapping Hindi-English Text Re-use Document Pairs," *In: FIRE 2011 Working Notes*

[21] Y. Palkovskii and A. Belov, "Exploring Cross Lingual Plagiarism Detection in Hindi-English with n-gram Fingerprinting and VSM based Similarity Detection," *In: FIRE2011 Working Notes*

[22] N. Aggarwal,K. Asooja, P. Buitelaar, T. Polajnar and J. Gracia, "Cross-Lingual Linking of News Stories using ESA," *In:Working note for CL!NSS, FIRE ISI, Kolkata,India(2012)*

[23] P. Arora and J.F., Jones, "DCU at FIRE 2013: Cross-Language !ndian News Story Search," *In: FIRE 2013 Working Notes*

[24] A. Pal andL. Gillam, "Set-based Similarity Measurement and Ranking Model to Identify Cases of Journalistic Text Reuse," *In: FIRE 2013 Working Notes*

[25] G. Tholpadi, and A.Param, "Leveraging Article Titles for Cross-lingual Linking of Focal News Events," *In: FIRE2013 Working Notes*

[26] D.A.R. Torrejon, and J. M. M. Ramos, " Linking English and Hindi News by IDF, Reference Monotony and Extended Contextual N-grams IR Engine," *In:FIRE2013 Working Notes*

[27] S. Das, and A. Kumar, "Performance Evaluation of Dictionary Based CLIR Strategies for Cross Language News Story Search," *In: FIRE 2013Working Notes*

[28] A. Kumar and S. Das, "Pre-Retrieval based Strategies for Cross Language News Story Search," *Proc. ACM FIRE '13 conference proceedings. FIRE '13, December 04-0-2013, New Delhi, India, http://dx.doi.org/10.1145/2701336.2701640*

[29] P. Gupta, P. Clough, P. Rosso, M. Stevenson and R.E. Banchs,"PAN@FIRE 2013: Overview of the Cross-Language !ndian News Story Search (CL!NSS) Track," *In:FIRE2013 Working Notes*

[30] I. Androutsopoulos, and P. Malakasiotis, "*A Survey of Paraphrasing and Textual Entailment Methods*," Journal of Artificial Intelligence Research, 38(1), 135-187, 2010

[31] C. Grozea, and M. Popescu, "*ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection*," Stein, Rosso, Stamatatos, Koppel, Agirre (Eds.): PAN'09, pp. 10-18, Donostia, Spain, 2009.

[32] M. Hagen and B. Stein, "*Candidate Document Retrieval for Web-Scale Text Reuse Detection?*,"Extended version of an ECDL 2010 poster paper [10]. M. Hagen and B. Stein. Capacity-constrained Query Formulation. Proc. of ECDL 2010 (posters), pages 384–388.

[33] A. H. Osman,N. Salima, M.S. Binwahlanc, R. Alteebd, and A. Abuobiedaa"An Improved Plagiarism Detection Scheme based on Semantic Role Labeling," *Journal of Applied soft computing 12(2012) 1493-1502 doi:10.1016/j.asoc.2011.12.021*

[34] B. Possas,N.Ziviani, B. Ribeiro-Neto and W. Meira Jr. "Maximal Termsets as a Query Structuring Mechanism," *Technical Report TR012/2005, Federal University of Minas Gerais, Belo Horizonte-MG, Brazil*

[35] B. Pouliquen, R. Steinberger and C. Ignat"Automatic Identification of Document Translations in Large Multilingual document Collections," *Proc. International Conference Recent Advances in Natural Language Processing (RANLP '03), pp 401-408*

[36] S. Vogel, H. Ney, and C. Tillmann, "HMM-Based Word Alignment in Statistical Translation," *Proc. 16th conference on Computational linguistics (COLING '96) vol. 2, pp. 836-841, Association for Computational Linguistics.doi:10.3115/993268.993313*

[37] N. A. Smith, "From Words to Corpora: Recognizing Translation," *Proc. Conference on Empirical Methods in natural Language Processing, Philadelphia, July 2002, pp. 95-102. Association for Computational Linguistics.*

[38] A. Valerio, D. Leake and A.J. Ca˜nas"Automatically Associating Documents with Concept Map Knowledge Model"*http://www.cs.indiana.edu/~avalerio/papers/2007clei.pdf*

[39] F. Sánchez-Vega, E.Villatoro-Tello, M. Montes-y-Gómez, L. Villaseñor-Pineda and P. Rosso"Determining and characterizing the reused text for plagiarism detection," *Journal of Expert Systems with Applications 40 (2013) 1804–1813. http://dx.doi.org/10.1016/j.eswa.2012.09.021*