Mapping Hindi-English Text Re-use Document Pairs

Parth Gupta[1], Khushboo Singhal[2]

[1] Natural Language Engineering Lab - ELiRF
Department of Information Systems and Computation
Universidad Politécnica de Valencia, Spain
http://users.dsic.upv.es/grupos/nle
pgupta@dsic.upv.es
[2] IR-Lab, DA-IICT, India.
http://irlab.daiict.ac.in
khushboo_singhal@daiict.ac.in

Abstract An approach to find the most probable English source document for the given Hindi suspicious document is presented. The approach does not involve any complex method of Machine Translation (MT) as a language normalization pre-processing step, rather it relies on standard cross-language resources available between Hindi-English and calculates the similarity using the Okapi BM25 model. We also present the further improvements in the system after the analysis and discuss the challenges involved. The system is developed as a part of CLiTR competition and uses the CLiTR-Dataset for the experimentation. The approach achieves the recall of 0.90 - the highest and F-measure of 0.79 - the 2^{nd} highest reported on the Dataset.

1 Introduction

Text re-use, here, tries to project the phenomenon of 'Plagiarism' where it refers to "the unauthorized use or close imitation of the language and thoughts of another author and the representation of them as one's own original work, as by not crediting the author"[1]. Easy access to the information with prolific World Wide Web makes it essential to check the authenticity of the work in certain texts like research papers, dissertations, student reports and so on. Cross-language text re-use is a special case where the information is taken from the the source which is in different language.

Recently, text re-use detection has attracted the attention of information retrieval (IR) and natural language processing (NLP) communities and the state-of-the-art is being advanced with evaluation campaigns like PAN (Uncovering Plagiarism, Authorship and Social Software Misuse)[2] at cross-language evaluation forum (CLEF)[3]. Text re-use system identifies the re-used text fragments in the given suspicious documents, if any, from the available source documents. The text re-use detection systems are broadly comprised of 4 steps [Potthast et al., 2009]:

[1] http://dictionary.reference.com/browse/plagiarism
[2] http://pan.webis.de
[3] http://clef-campaign.org/

1. pre-processing, which consists of the normalization of text, language identification and/or translation of documents;
2. selection of candidate documents, i.e. the selection of a small subset of a large collection as potential source of text re-use;
3. detailed analysis, which implies the investigation of suspicious and source documents in detail to identify the re-used text sections; and
4. post-processing, which consists of merging the detected parts of a single re-use case, removing detected cases which are properly cited.

Recently, a few approachs to address cross-language text re-use (CLTR) and plagiarism detection (CLPD) are reported. In [Ceska et al., 2008], the cross-language similarity between texts is calculated using multilingual thesaurus like EuroWordnet and language specific lemmatizer is used as a pre-processing step. In [Steinberger et al., 2002], cross-lingual conceptual thesaurus is used to map a list of concepts to the text in question, and later the similarity is measured in terms of number of matching concepts. The approach presented in [Potthast et al., 2008], exploits the vocabulary correlations in the comparable corpus. [Pinto et al., 2009] measures the cross-language similarity based on the statistical bilingual dictionary generated from the parallel corpus - similar to one used for machine translation. Some of the popular approaches used for cross-language information retrieval (CLIR) are also tested to detect CLTR like character n-gram model [McNamee and Mayfield, 2004] for the languages which share the same script and can be found in [Potthast et al., 2011]. Most popular approach, noticed in the recent editions of PAN , is to translate the documents into the language of comparison using any of the available machine translation (MT) service and then carry a monolingual analysis.

In our approach, we transform the Hindi documents in English comparable space by the means of available resources like bilingual dictionary, wordnet and transliteration engine. Thereafter, we calculate the similarity based on the probabilistic model Okapi BM25 [Robertson and Spärck Jones, 1994]. From our experiments and analysis in [Rao et al., 2011,Gupta et al., 2011], we believe that similarity based on words has an edge over that based on word n-grams because in the cross-language environment the sequential information of terms is generally not preserved and hence the latter does not produce the best results. We also notice that bilingual dictionary and transliteration engine produces the best results.

The problem statement, of CLiTR track, is to identify the most potential source document for the text re-use, if any, for the given suspicious document. The source documents are in the English while the suspicious documents are in Hindi. The system, developed to address the aforementioned problem, is described in Section 2. We report the results in Section 3 while in Section 4 we present the analysis of the results. Finally in Section 5 we conclude the work and talk about future activities.

2 Approach

In the standard setting of candidate retrieval, for each suspicious document d_i, a set of source documents S', from entire collection of source documents S, is retrieved, where $|S'| << |S|$. Then, each candidate document s_i in S' is further compared in detail with the d_i in the following stage to find the fragment level text re-use. In contrast to this, the

problem statement in CLiTR limits itself to find the source document of text re-use, if any, rather than finding the re-used fragments. Therefore, we consider it as a candidate retrieval step with aim to maximize Recall@1.

The strategy adopted has its roots in cross-language information retrieval (CLIR). The approach is described in Fig. 1. The resources involved are described below:

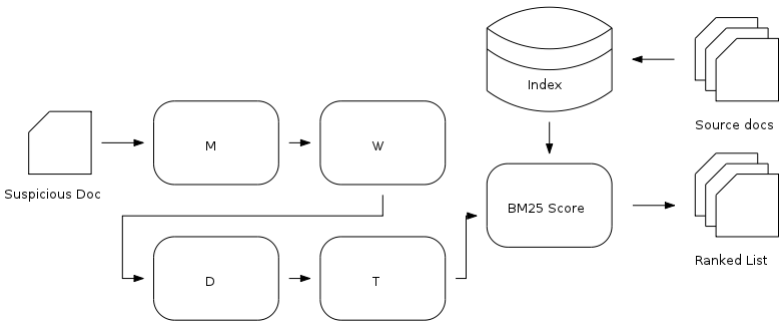

Figure 1. Block diagram of the approach

It is a two phase process where, first phase tries to bring the document in Hindi and English to comparable space by the means of available cross-language resources, while in second phase the similarity is calculated to find the source document of text re-use. These phases are described below

2.1 Resources

In order to compare the Hindi suspicious documents with English source documents, we use different natural language resources like morphological analyser, bilingual dictionary, wordnet and transliteration system. We have tested the impact of these resources on the detection and the results are reported in Section 3.

Morphological Analyser (M): In order to find the root of each term t_i in suspicious document d, we incorporate morphological analyser. The intension behind this step is to maximize the probability of getting an entry in the bilingual dictionary for term t_i. We use stemmer proposed in [Ramanathan and Rao, 2003], which largely handles inflectional morphology. It does not account for most of the derivational morphology of Hindi.

Wordnet (W): For each term t_i in suspicious document d, we retrieve all its senses and synonyms from the Hindi Wordnet [Narayan et al., 2002], if it has an entry in wordnet.

Bilingual Dictionary (D): We substitute each term t_i of suspicious document d by its corresponding English dictionary entry if any. We use The Hindi Universal Word (UW) dictionary[4], which contains total 134968 words.

[4] http://www.cfilt.iitb.ac.in/hdict/webinterface_user/index.php

Transliteration (T): If the term t_i of the suspicious document q has no entry in the dictionary then we transliterate t_i using Google Transliterate API[5].

2.2 Similarity Score

We index all the English source documents in S. Each suspicious document d_i in D is fired as a query on this index and the ranklist is retrieved. The s_i with the highest BM25 score for given d_i is considered to be the potential source of re-use for d_i.

We also introduce a similarity threshold θ. If a suspicious document does not have any source document with similarity score above θ, we consider it free from text re-use. Though we intend to introduce this mechanism in hope to take care of false positives, it is not the best strategy in the present form as discussed in Section 4.

3 Results

We tested the above mentioned strategies on the training & test data. Table 1 contains the results on training data.

Method	Precision	Recall	F-Measure
D	0.455	0.692	0.549
W+D	0.172	0.262	0.207
D+T	0.505	0.769	0.610

Table 1. Results on training data

After looking at the high value of recall we worked on improving the precision. In order to reduce the false positives, we introduced a threshold on similarity score. Fig. 2 describes the evaluation performance with different threshold values on training data.

It can be seen that setting the θ below 9.0 will hurt the precision without gaining in terms of recall, similarly, setting it above 20.0 will hurt recall greatly. So we set the θ = 15.0, between 9.0 and 20.0 based on empirical tuning, which achieves the maximum F-Measure on training data.

Table 2 shows the results achieved on the test data.

As can be seen in Table 2, the performance with θ is not in accordance with the improvement obtained on training data, hence we tune θ according to the test data and realise that θ = 64 produces best results. We also incorporate the morphological analyser and the results are depicted in Table 3, which involves the highest recall and 2^{nd} highest F-measure score achieved in the competition.

[5] www.google.com/transliterate

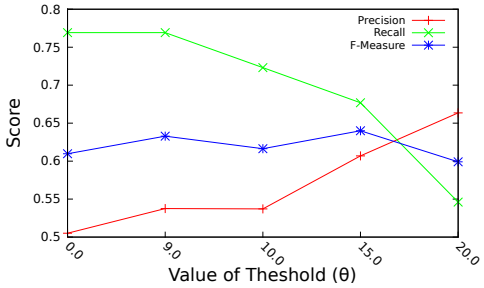

Figure 2. Behaviour of evaluation measures with threshold values

Method	Precision	Recall	F-Measure
D (Run-1)	0.342	0.580	0.430
W+D	0.263	0.343	0.298
D+T (Run-2)	0.474	0.804	0.596
D+T+θ (Run-3)	0.439	0.607	0.509

Table 2. Results on test data

4 Analysis

From the results depicted in Table 2, it is noticeable that bilingual dictionary (D) was itself capable to help in identifying 58% of re-used documents. The average length of suspicious documents is 358 and the average hit in dictionary look up is 104 terms. This signifies that there are a lot of terms, which do not participate in similarity, can improve the detection.

We also considered a fact that, if the word itself is not in the dictionary then we can find its all possible senses and synonyms from wordnet (W) to look in dictionary (D) with the hope to increase the similarity. Surprisingly the performance deteriorates as can be seen in Table 2, we believe incorporation of all the senses triggered the topic drift. Wordnet incorporation in this way is not useful and there has to be a logical criterion for inclusion of a particular sense for similarity, which we aim to investigate in future.

With our experience with named entities (NEs) in [Gupta et al., 2010], we considered to transliterate all the words, which could not be found in dictionary. These words have high probability to be an NE. The results in Table 2, achieved with transliteration (D+T), confirms this hypothesis and the recall is increased to 85%.

The similarity threshold, in order to abandon false positives, improves the performance as shown in Table 3. This threshold selection strategy is not comparable across corpora and hence is unreliable. As we discussed in Section 1, the main aim of our approach is to maximize the recall and provide the candidate list to the later stages, where the documents are compared in detail at sentence or fragment levels. These later stages usually take care of false positives and hence we did not handle it. Still it makes sense to incorporate a robust document level threshold, which we intend to investigate in future.

In further analysis of results, we notice that some of the Hindi terms are in their morphological form and hence could not get dictionary hit while their root terms exist in the dictionary. Hence we incorporate the morphological analyser which further improves the performance as depicted in Table 3 and achieves the recall as high as 90%.

Method	Precision	Recall	F-Measure
D+T+θ	0.850	0.739	0.783
M+D+T	0.695	0.904	0.786

Table 3. Improved results on test data

Table 4 presents the performance evaluation of the best run (M+D+T) for the different types of re-use cases (exact, heavy and light).

Type	Exact	Heavy	Light
Recall	1.000	0.907	0.855

Table 4. Performance evaluation based on different levels of re-use.

5 Conclusion and Future Work

The obtained results suggest that available resources are capable enough in finding the text re-use document pairs for Hindi-English. Transliteration helps in identifying the named entities and contributes to obtain a higher recall. Morphological analyzer provides better look-up in dictionary in terms of performance. In future, we wish to work on the precision of the system in order to take care of false positives. Moreover, we would also like to investigate the better way to incorporate wordnet.

6 Acknowledgment

The work of the first author has been partially funded by the European Commission as part of the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework.

References

[Ceska et al., 2008] Ceska, Z., Toman, M., and Jezek, K. (2008). Multilingual plagiarism detection. In *Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '08, pages 83–92, Berlin, Heidelberg. Springer-Verlag.

[Gupta et al., 2010] Gupta, P., Rao, S., and Majumder, P. (2010). External plagiarism detection: N-gram approach using named entity recognizer - lab report for pan at clef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.

[Gupta et al., 2011] Gupta, P., Singhal, K., Majumder, P., and Rosso, P. (2011). Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism. In *ICON 2011*, Chennai, India. Macmillan Publishers.

[McNamee and Mayfield, 2004] McNamee, P. and Mayfield, J. (2004). Character n-gram tokenization for european language text retrieval. *Inf. Retr.*, 7(1-2):73–97.

[Narayan et al., 2002] Narayan, D., Chakrabarti, D., Pande, P., and Bhattacharyya, P. (2002). An experience in building the indo wordnet - a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India*.

[Pinto et al., 2009] Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., and Rosso, P. (2009). A statistical approach to crosslingual natural language tasks. *J. Algorithms*, 64(1):51–60.

[Potthast et al., 2011] Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011). Cross-Language Plagiarism Detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 45(1).

[Potthast et al., 2008] Potthast, M., Stein, B., and Anderka, M. (2008). A wikipedia-based multilingual retrieval model. In *ECIR*, pages 522–530.

[Potthast et al., 2009] Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P. (2009). Overview of the 1st International Competition on Plagiarism Detection. In Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E., editors, *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 1–9. CEUR-WS.org.

[Ramanathan and Rao, 2003] Ramanathan, A. and Rao, D. D. (2003). A lightweight stemmer for Hindi. In *Computatinal Linguistics for South Asian Languages*, Budapest, Apr.

[Rao et al., 2011] Rao, S., Gupta, P., Singhal, K., and Majumder, P. (2011). External & intrinsic plagiarism detection: Vsm & discourse markers based approach - notebook for pan at clef 2011. In *CLEF (Notebook Papers/Labs/Workshop)*.

[Robertson and Spärck Jones, 1994] Robertson, S. and Spärck Jones, K. (1994). Simple, proven approaches to text retrieval. Technical Report UCAM-CL-TR-356, University of Cambridge, Computer Laboratory.

[Steinberger et al., 2002] Steinberger, R., Pouliquen, B., and Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In *CICLing*, pages 415–424.