Cross Lingual Text Reuse Detection based on Keyphrase Extraction and Similarity Measures

Rambhoopal Kothwal, Vasudeva Varma

International Institute of Information Technology, Hyderabad
bhupal_iiit@research.iiit.ac.in, vv@iiit.ac.in

Abstract

The information available on the web in various languages is growing very fast, and it can reuse to create new document in other language and present as an original document. The cross language text reuse is on rise and unfortunately hard to detect. There is not much research done to detect the cross language text reuse, especially for less resource languages and between distant language pairs, such as Arabic and Indian languages. In our work, we focus on detecting the suspicious documents created by text reuse of others documents across languages. We differ from other available approaches in two ways: (1) we make use of keyphrases instead of n-grams (2) we use a new measure for similarity while using an open source search engine for text reuse detection. Two approaches we proposed in this paper have secured the top rank and the third rank in CL!TR-2011 task.

1 Introduction

Text reuse is an imitate of phrases from others text documents and present them as their own document. As there is vast growth of information on the web in various languages and can easily access this information to create text reuse documents in other languages. This text reuse documents across languages is on rise and hard to detect. Identifying this text reuse documents manually is very difficult across languages and becomes infeasible on large-scale of documents. Thus automatic extraction of the text reuse detection attracts attention. In our work, we focus on detecting the suspicious document created by text reuse of others documents across languages. We differ from other available approaches by defining the usage of keyphrases of the document instead of n-grams and use of a new measure for similarity while using an open source search engine for text reuse detection. Keyphrases or important topics are sequence of words that captures the main topics covered in a document.

This work is partially funded by the European Commission as part of the WIQEI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework

Our paper illustrates the cross language text reuse detection. For the related work in this area we referred following works: *Cross-language Plagiarism Detection*, which introduces a comprehensive retrieval process for cross-language plagiarism detection and a large-scale evaluation of three retrieval models to measure cross-language similarity of text by [1], *Towards Document Plagiarism Detection based on the Relevance and Fragmentation of the Reused Text*, which propose to represent the common text with a set of features that denotes its relevance and fragmentation with conjunction of supervised learning algorithms for automatic detection of document plagiarism by[2], *External Plagiarism Detection*, which takes a moving window of four word sequence and use chunk ratio R for identifying plagiarism passages by [3], *Plagiarism Detection across Distant Language Pairs*, which based on machine translation and monolingual similarity measure by [4], *External Plagiarism Detection*, which compares different similarity measures by [5], *Automatic Keyphrases Extraction from Scientific Documents Using N-gram Filtration Technique*, which presents an automatic keyphrase extraction technique by [6]. The rest of our paper organized as follows: section 2 explains our approach, our experiments and results explained in section 3. We conclude our paper with section 4 Conclusions and future work.

2 Our Approach

Automatic detection of text reuse documents given suspicious and source documents in same language are being well illustrated in all the earlier works. CL!TR task training and testing data had suspicious documents in one language and source document in another language. To determine the text reuse documents is perplex, when they are in different languages. Translation of documents is only contingency to overcome this problem. Once all documents are in same language we use n-gram filtration and term weighting scheme techniques for automatic keyphrase extraction. Extraction of keyphrases is use in text document classification, text document clustering and summarization etc. This n-gram filtration technique extracts n-grams using data compression based technique and with simple refinements and pattern filtration algorithms. This n-gram filtration technique does not require any complex mathematics. In term weighting scheme, we have used importance of position of the sentence where given phrase occurs first in document and position of phrase in sentence. Distinct n-gram lists is use to collect the n-grams of different length from pre-processed document after applying the n-gram filtration algorithm. Term weighting scheme calculates the weight of collected n-grams.

After several observations of the earlier works in syntactic analysis, it is clear that position of the phrases in the sentences would extract an important measure to decide their role in the document. To calculate the term weight we consider phrase position in the sentence and sentence position where given phrase occurs first in the document. In every document, frequency of occurrence of n-grams differ. Generally lower n-grams are more frequent than higher n-grams and applying term weighting schemes show a bias towards n-grams having smaller value of ‘n’. To solve this problem a slightly different strategy applied, which separately treats the n-grams of different lengths in weight calculation, in final keyphrase selection phase. Before extracting the keyphrases we pre-process the documents which would require following steps. (1) separating sentences, (2) removing all punctuation marks and stopwords and (3) converting entire document to lowercase alphabets. To create n-gram list we used LZ78 [11],[12] data compression technique, with some simple modifications. Following are the simple changes made to LZ78 technique, words in place of characters and space is use as a delimiter between the words. Using LZ78 technique we first create a list of distinct n-gram patterns given pre-processed text document. But this extracted n-grams list have words which are not more valid or important to consider them as keyphrases of text document for example Single Alphabets, Verbs which are less important. Single alphabets replaced with "**"and Verbs which are invalid were remove using [8],[9]. We separate lists of n-grams of distinct length. The separate collection of n-grams of different length has helped us in two ways: (1) in deleting the n-gram which are bias towards n-grams, which are having smaller value of 'n'. (2) helped in filtering out the higher length n-grams earlier with proper replacement of other unique terms, if it satisfy certain frequency related score. The n-gram filtration and term weight techniques for automatic extraction of keyphrases from the words of text document was implementation of the work of [6].

Similarity identifier was base on comparative of all keyphrases extracted from the suspicious document with all source documents from the collection. For measuring similarity between suspicious and source documents we have used an open source web search engine called Nutch, which uses Lucene Java for the search and index component. Nutch is a complete open source web search engine package that aims to index the World Wide Web as effectively as commercial search services [7] and can use for intranet and campus network which can run all its components on a single server. Using Nutch we index all source documents, which uses Opic-scoring algorithm to calculate the document score. Source documents as index and keyphrases of given suspicious document as queries we retrieve all the relevant source documents. Finally we get several groups of retrieved source documents for all keyphrases of given suspicious document. We create a list of unique source documents with their frequency score by combining all the groups of retrieved source documents for all keyphrases of a suspicious document. For each suspicious document we create this list of unique source documents with frequency score. The highest frequent score source document from the list of a suspicious document is consider source of its text reuse.

3 Experiments and Results

In this section we describe our three approaches implemented for cross language text reuse detection and comparison of their results. CL!TR is a Cross language Indian Text Reuse task in FIRE -2011 where the task is to identify the set of suspicious documents in Hindi created by text reuse from the set of source documents in English. The CL!TR -2011 training and testing collection contains suspicious documents in Hindi language and source documents in English language. Training data contains 198 suspicious documents, out of which 130 documents are positive text reused examples and other 68 documents are negative text reused examples. Testing data contains 190 suspicious documents. Each of training and testing collection contains 5032 source documents in English language. We use CL!TR task data for all our experiments. Training and testing suspicious documents are translate using Google translator API. For our three approaches we use translated suspicious documents. Porter stemmer is use in some of our approaches and our English stopword list contains 173 words. The Table.1 illustrates CL!TR task results. Results of our approaches are highlight bold in the CL!TR task results table.

First Approach (1*st* Run): Similarity identifier we consider in this approach to find the similarity between suspicious documents and source document is Cosine Similarity. In pre-processing phase we remove stopwords and implemented stemmer. Tri-grams extracted using sliding window of word tri-grams for both suspicious and source documents. Cosine Similarity between tri-grams of each suspicious document against with all tri-grams of all source documents is measure. We consider the top similarity scored source document as text reuse source for creating suspicious document. For classification of documents used J48 Decision tree classifier using WEKA tool. We trained the J48 classifier model using training set of 130 documents as positive text reuse examples and 68 documents as negative text reuse examples with respect to the similarity scores obtained by this approach. The trained model when applied on the test collection of 190 suspicious document with similarity scores obtained with this approach, the classifier classified 117 documents are text reused. This approach by [3] has motivated us for its very good precision in PAN-2010 and implemented it with augmentation of stemmer and removed stopwords for better recall, but results disappointed us again with less recall.

Second Approach (2*nd* Run): This is base on set of features that denotes the relevance and the length and quantity of the word sequences. In this approach we made use of stemmer in augmented with the work of [2]. Relevance and Length and frequency of the extracted n-grams of the source and suspicious documents is measure to know the similarity scores. For classification of documents used J48 Decision tree classifier using WEKA tool. We trained the J48 classifier model using training set of 130 documents as positive text reuse examples and 68 documents as negative text reuse examples with respect to the similarity scores obtained by this approach. The trained model when applied on the test collection of 190 suspicious document with similarity scores obtained with this approach, the classifier classified 125 documents are text reused. This approach was rank third in CL!TR-2011 task announced results.

Third Approach (3*rd* Run): This approach is base on which we have illustrated in section 2. We use minimum thersold of 31(frequency score) to consider a document as source for text reuse. The thersold is base on the development corpus. Our approach given 190 suspicious documents, considered 147 documents as text reused whose top frequency score was above decided thersold. This approach secured first rank in CL!TR-2011 task.

The CL!TR task results table clearly illustrate that our third approach has outperformed the accuracy results achieved by other approaches in cross language text reuse detection task. The third approach secured first place and our second

Rank	F-measure	Recall	Precision	Run
1	**0.649**	**0.750**	**0.571**	**our third approach**
2	0.609	0.821	0.484	1
3	**0.608**	**0.643**	**0.576**	**our second approach**
4	0.603	0.589	0.617	1
5	0.596	0.804	0.474	2
6	0.589	0.795	0.468	2
7	**0.576**	**0.589**	**0.564**	**our first approach**
8	0.541	0.473	0.631	2
9	0.523	0.500	0.549	3
10	0.509	0.607	0.439	3
11	0.430	0.580	0.342	1
12	0.220	0.214	0.226	2
13	0.220	0.214	0.226	3
14	0.085	0.107	0.070	1
15	0.000	0.000	0.000	1

Table 1: Illustrate CL!TR Task Results

approach secured third place. The F-measure of 0.649 and 0.608 are obtained by our third approach and second approach respectively. The difference in F-measure between second rank system and third rank system (our second approach) is very less. When we compare our three approaches results, CL!TR results table clearly infer low recall for our first and second approaches.

4 Conclusions and Future Work

In our work in detecting the suspicious documents created by text reuse of others text documents across language, we differ from various approaches by defining the usage of keyphrases instead of n-grams and use of a new measure for similarity by using an open source search engine for text reuse detection. Our results show that out of our three approaches in CL!TR task our two approaches ranked in top three in the results table. we have ranked has first and third in the CL!TR task. The F-measure obtained by our third approach which ranked first is 0.649 and the second approach which ranked third is 0.608. For future work it would be interesting to further analyze how semantic text feature applied across languages with distant language pairs could improve the F-measure.

5 References

[1] Martin Potthast, Alberto Barron-Cedeno, Benno Stein and Paolo Rosso.: Cross-language Plagiarism Detection. In: Springer Science+Business Media B.V., 2010.

[2] Fernando Sanchez-Vega, Luis Villasenor-Pineda, Manuel Montes-y-Gomez and Paolo Rosso.: Towards Document Plagiarism Detection based on the Relevance and Fragmentation of the Reused Text. In: MICAI, 2010.

[3] Sobha Lalitha Devi, Pattabhi R K Rao, Vijay Sundar Ram and A Akilandeshwari .: External Plagiarism Detection. In: Lab Report for PAN at CLEF, 2010.
[4] Alberto Barron-Cedeno, Paolo Rosso, Eneko Agirre and Gorka Labaka.: Plagiarism Detection across Distant Language Pairs. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).
[5] Hoad, Timothy C. and Justin Zobel.: Methods for identifying versioned and plagiarized documents. In:Journal of the American Society for Information Science and Technology (JASIST) 54(3), 203215 (2003).
[6] Niraj Kumar and Kannan Srinathan.: Automatic Keyphrase Extraction from Scientific Documents Using N-gram Filtartion Technique. In: Published in the proceedings of ACM DocEng, 2008.
[7] Rohit Khare, Doug Cutting, Kragen Sitakar and Adam Rifkin.:Nutch: A Flexible and Scalable Open-Source Web Search Engine. In:CommerenceNet Labs Technical Report 04-04, 2004.
[8] English Vocabulary: Regular Verbs List (EnglishClub.com)
[9] Irregular verbs:English Wiktionary,
http://en.wiktionary.org/wiki/Appendix:Irregular_verbs:English
[10] Porter Stemming Algorithm for suffix stripping , web link
http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html
[11] Khalid Sayood.: Introduction to Data Compression. In: ELSEVIER, 2nd Edition 2000.
[12] Ida m. Pu.: Fundamental data Compression. In: ELSEVIER, 1st edition 2006.