Plagiarism Detection in Regional Languages – Its challenges in context to Punjabi documents

Rajeev Puri[1], R.P.S. Bedi[2], Vishal Goyal[3]

[1]Research Scholar, Punjab Technical University, Kapurthala Road, Jalandhar.
[2]Research Supervisor, Punjab Technical University, Kapurthala Road, Jalandhar
[3]Assistant Professor, Dept of Comp. Sc, Punjabi University, Patiala.
[1]rpuri@davjalandhar.com, [2]bedirps2000@yahoo.com, [3]Vishal.pup@gmail.com

ABSTRACT:

Plagiarism detection has always been a challenging task for researchers and developers across the world. A number of researchers have contributed to this task by suggesting methods to detect plagiarism in research publications, books and software. Subsequently some applications based on these methods are also available in open source as well as commercial marketplace. Majority of these applications are performing well with English language documents, but fail to give satisfactory results with documents written in regional languages. This paper deals with investigating the challenges in plagiarism detection in regional languages, with specific case study of documents written in Punjabi language.

1. Introduction

Plagiarism can be defined as the unauthorized use or reproduction of the work originally done by someone else and thereby presenting it as one's "Own and Original" work. The copied work may be from some researcher's research work or it may be a copy or partial copy of some literary articles, piece of code originally written by someone else.

The plagiarism has been around since the human beings started documenting their research and literary works. The problem of plagiarism has increased to a great extent due to increased use of digital media for storage, retrieval and communication of the information. Plagiarism may be intentional when the author makes a copy of someone else's work intentionally, or it may be un-intentional when the author does not know the rules and regulations for research publications and fails to give credits/citations to the original author.

2. Types of Plagiarism

2.1. Copy & Paste Plagiarism

Picking sentences or phrases or paragraphs from the original work done by oneself or by someone else and not giving any reference/due credit to the original author is known as copy and paste plagiarism. It is the simplest form of plagiarism and is very easy to detect using the inspection method or by using the available automated tools.[1]

2.2. Word Switch Plagiarism

Picking the sentences or phrases or paragraphs from the original work done by oneself or by someone else , making a few changes in the words by writing alternative words (synonyms) to hide the originality and not giving due credits to the author is known as Word Switch Plagiarism.

2.3. Idea Theft

Picking the idea from someone else's work without giving a due credit to the original source of idea is known as idea theft.

2.4. Missing Citations

This is similar to idea theft, where someone copies the text from someone else's research findings and fails to give a citation to the original author.

2.5. Invalid citations

In this category, the author gives fake or invalid references related to his/her research work.

2.6. Self Plagiarism

Copying one's own work and present it as something new, without giving any reference to the previous work is self – plagiarism.

2.7. Dual Submission

Submitting the same work to two or more journals is also a type of plagiarism.

2.8. Translation

Translation is one of the most popular forms of plagiarism, where the author translates the original work done by someone else into some other language and presents as his own work, without giving a reference to the original author. This type of plagiarism is very hard to detect.

A number of other categories of plagiarism have also been identified by many researchers in various research publications [2],[3].

3. Plagiarism Detection

The plagiarism detection methods proposed by researchers involve a variety of tasks to be performed starting from document procurement, text preprocessing, document indexing, keywords identification, similarity detection to final calculation of the degree of plagiarism in the query document. The subsequent sections of this paper deal with the challenges identified in each of these tasks required for detecting plagiarism in Punjabi text documents.

3.1. Repository Building

Unlike English language, the regional languages such as Punjabi and Hindi have multiple keyboard layouts. The same key character from keyboard is used to represent more than one letters depending on their layouts and typefaces used to show the characters. For example, two most popular layouts for Punjabi text are Phonetic keyboard layout and inscript / revised inscript layout [Fig 1]. An in-compatible typeface may change the meaning of words or may even show incorrect spellings of the words. The documents collected from different sources need to be identified first for the keyboard layouts used in writing those documents. Further the documents need to be converted to a uniform encoding scheme (UNICODE) format for storage and subsequent retrievals during the plagiarism detection process. The documents already written in UNICODE may skip this conversion routine.

3.2. Stop word removal

The most frequently used words from the language, such as articles, prepositions, punctuation marks etc should be removed from the documents before comparison. The presence of stop words in the document sources increases document vector space for comparison and also may influence the research findings. A dictionary of such stop words for Punjabi documents needs to be built by analyzing the frequency of the words in a large corpus.

3.3. Synonyms replacement

Use of synonyms is the most common method for plagiarizing the texts. In plagiarism detection tasks, the document comparison results can be more accurate if all the synonyms of a particular word are replaced by the indexed synonym word. For this, a complete, indexed dictionary along with the synonym words needs to be available for the source language. No such accurate synonyms dictionary database is available for the Punjabi language.

3.4. Word Stemming

Stemming is the process that conflates the morphologically similar terms into a single root word, that can be used for improving the efficiency of the plagiarism detection task. The stemming process also reduces the size of the document, making the information retrieval process faster. Researchers have suggested different approaches for stemming the text. The Brute force approach used by Frakes , Baeza Yates [4] uses a

table lookup method for stemming. This approach suffers from its limited scope, which is proportional to the size of the lookup dictionary. A statistical approach suggest by Hafer and Weiss [5] called "Successor Variety Stemmer" identifies the morpheme boundaries based on the available lexicon and decides where the words should be broken to get a stem. A suffix stripping approach was proposed by Lovins [6] and Porter [7] that uses the longest match first suffix stripping for finding the root word. All these algorithms are best suitable for less inflectional languages like English. A Lightweight Stemmer for Hindi by Ramanathan and Durgesh D Rao[8], YASS - Yet Another Suffix Stripper for Bangali by Prasenjit Majumder et all [9], Rule based stemmer for Bangali by Sandipan Sarkar and Sivaji Bandyopadhyay[10], Punjabi language stemmer for nouns and proper names by Vishal Gupta and Gurpreet Singh Lehal[11] have also been developed for Indian languages. These stemming approached have demonstrated good results, but are yet far away from the state of the art stemmers.

3.5. Keyword Identification

Keyword identification plays an important role in plagiarism detection. The keywords retrieved from a document may help in building a list of suspected documents, where there are high chances of finding the matching contents. The keyword identification techniques proposed by the researchers can be classified as under –

3.5.1. Statistical Techniques

These techniques do not require any training data. The statistical information of the words in the document can be used to identify the keywords relevant to the document. Cohen[12] has used a language and domain independent technique that uses N-Grams for finding the important terms in the document. An extension of this technique has been tested on English, Spanish, German and some other language texts, but its relevance to the Punjabi text has to be established yet. Some other approaches include the use of TF-IDF for finding the relevant keywords for a document. This approach also eliminates the need of removing stop words for keywords detection, which otherwise was a big challenge as there is no such standard list of stop words available for Punjabi text.

3.5.2. Linguistic approaches

The language dependent approaches have also shown good results in automatic keyword detection. Gonenc Ercan and Ilyas Cicekli[13] have described keyword extraction as a supervised learning task and have used Lexical Chains for keywords extraction from a document. A. Hulth [14] demonstrated that by adding linguistic knowledge to the representation, a better result is obtained after extracting NP-chunks and by adding the POS tag(s) assigned to the term as a feature. Since these approaches are language dependent, their direct use in Punjabi text keyword identification is not possible. The standard Punjabi annotated text needs to be developed before using linguistic approaches.

3.5.3. Machine Learning Approaches

Machine learning approach uses training documents to learn a model and applies the model to find keywords from new documents. This approach includes Naïve Bayes [15], SVM [16], CRF[17] etc. All these approaches need good amount and quality of training data for providing reasonably accurate results. The training data needs to be developed for Punjabi language.

4. Similarity calculation

A number of schemes have been proposed for similarity calculations. These schemes are categorized as under.

4.1. String based comparisons

It is the most common method for document comparison. In this approach, a word for word check is made between the suspected and the query documents. This approach is language independent. It also matches any misspellings in suspected and query documents. A number of algorithms such as Rabin-Karp, KMP String matching, Boyer-Moore algorithms have been proposed for faster string matching. These algorithms can be easily used for comparing Punjabi text documents.

4.2. Term frequency comparisons

The algorithms under this approach are based on the frequency of occurrence of words in the source as well as query documents. Algorithms such as Bag of words, nGram comparison etc use this approach to estimate the similarity between documents. These algorithms are also language independent and can be used for Punjabi text.

4.3. Vector space model

The VSM can be used to measure the similarity between source document and query documents. The similarity is measured using cosine similarity of the term frequencies of the documents being compared. This model is language independent and hence can be applied to the Punjabi language documents. The documents can be pre-processed to reduce the vector space of the source as well as query document.

4.4. NGram Overlap

This method uses n-Gram approach for detecting the similarity between two documents. This method in combination with VSM is known to show improved results.

A few other similariy measures such as Jaccard Similarity , Dice similarity, overlap similarity , Hoad and Zobble similarity can also be used for similarity calculation. All these approaches can be comfortably used with Punjabi text.

5. The proposed model

The [Fig 2] shows the proposed architecture for designing a system for detecting plagiarism in Punjabi Text documents. This system uses a mix of statistical as well as linguistic approaches for its tasks. The language dependent components have been kept separate from the language independent components, so that by just replacing the language dependent components for one language to other language, the system should work well.

6. Conclusion

A detailed study of the tasks involved in plagiarism detection broadly suggests two approaches viz. statistical approach and linguistic approach for text pre-processing. The current plagiarism detection tools perform well with English language text, however very poor performances are observed when the tools are posed with Punjabi text. The statistical approaches used in these tools are language independent, so the real problem behind failure is the language dependent components used in these tools. The linguistic approaches require huge language resources such as dictionaries, synonyms databases, stop words database and stemming rules etc. The development of these linguistic resources for plagiarism detection is the need of the hour. Further, keeping the language dependent components replaceable in the tool, a unified plagiarism detection tool can be developed that automatically detects the language of the document and use the appropriate linguistic resources for plagiarism detection.

References:

1. Romans Lukashenko, Vita Graudina, Janis Grundspenkis (2007)- Computer-Based Plagiarism Detection Methods and Tools: An Overview - International Conference on Computer Systems and Technologies - CompSysTech'07 Pages IIIA.18-(1 – 6)
2. Hermann Maurer, Frank Kappe, Bilal Zaka (2006) Plagiarism - A Survey , Journal of Universal Computer Science, vol. 12, no. 8 (2006), 1050-1084
3. Mathieu Bouville (2008) - Plagiarism: Words and ideas , Journal of Science and Engineering Ethics, vol 14, issue 3 , 311-322
4. Frakes, W. B. and R. Baeza-Yates. Information Retrieval: DataStructures & Algorithms. Prentice Hall. (1992)
5. M. Hafer and S. Weiss. Word Segmentation by Letter Successor Varieties,Information Storage and Retrieval, 10, (1974) 371-85.
6. Julie Beth Lovins. Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2, March and June 1968
7. Porter, M.F. An algorithm for suffix stripping, Program, Vol. 14 No.3 (1980), pp.130-137

8. Ananthakrishnan Ramanathan and Durgesh D Rao. A Lightweight Stemmer for Hindi In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL), on Computational Linguistics for South Asian Languages(Budapest, Apr.) Workshop(2003),pp 42-48.
9. Prasenjit Majumder et all. YASS: Yet another suffix stripper, ACM Transactions on Information Systems (TOIS)Volume 25 Issue 4, October 2007 Article No. 18
10. Sandipan Sarkar , Sivaji Bandyopadhyay. Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages
11. Vishal Gupta and Gurpreet Singh Lehal. Punjabi Language Stemmer for nouns and proper names, Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, Chiang Mai, Thailand, pp. 35–39. (2011).
12. J. D. Cohen. Highlights: Language and Domain-independent Automatic Indexing Terms for Abstracting. Journal of the American Society for Information Science, 1995, 46(3): 162-174.
13. G. Ercan, I. Cicekli. Using Lexical Chains for Keyword Extraction. Information Processing and Management, 2007, 43(6): 1705-1714.
14. A. Hulth. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In: Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing, Sapporo, Japan, 2003: 216-223.
15. E. Frank, G. W. Paynter, I. H. Witten. Domain-Specific Keyphrase Extraction. In: Proceedings of the 16th International Joint Conference on Aritifcal Intelliegence, Stockholm, Sweden, Morgan Kaufmann, 1999: 668-673.
16. K. Zhang, H. Xu, J. Tang, J. Z. Li. Keyword Extraction Using Support Vector Machine. In: Proceedings of the Seventh International Conference on Web-Age Information Management (WAIM2006), Hong Kong, China, 2006: 85-96.
17. Chengzhi ZHANG et al. Automatic Keyword Extraction from Documents Using Conditional Random Fields. Journal of Computational Information Systems 4:3(2008) 1169-1180

Research Cell: An International Journal of Engineering Sciences ISSN: 2229-6913 Issue Dec. 2011, Vol. 5

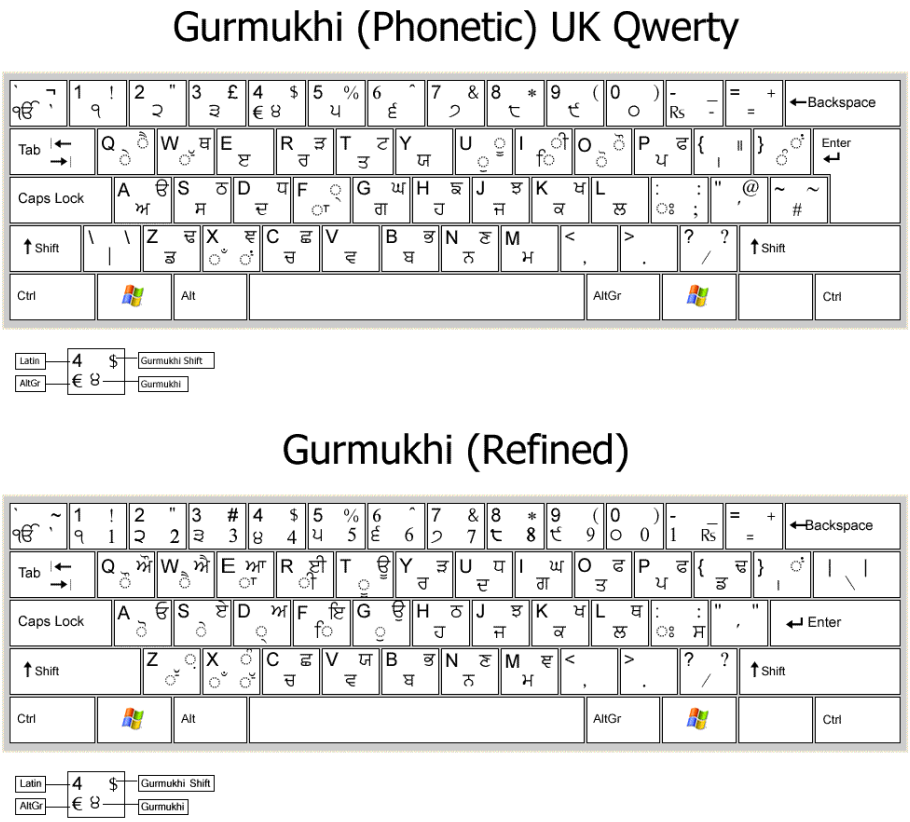

Fig 1. Phonetic and Inscript (Refined) keyboard layouts for Gurmukhi Script.

Research Cell: An International Journal of Engineering Sciences ISSN: 2229-6913 Issue Dec. 2011, Vol. 5

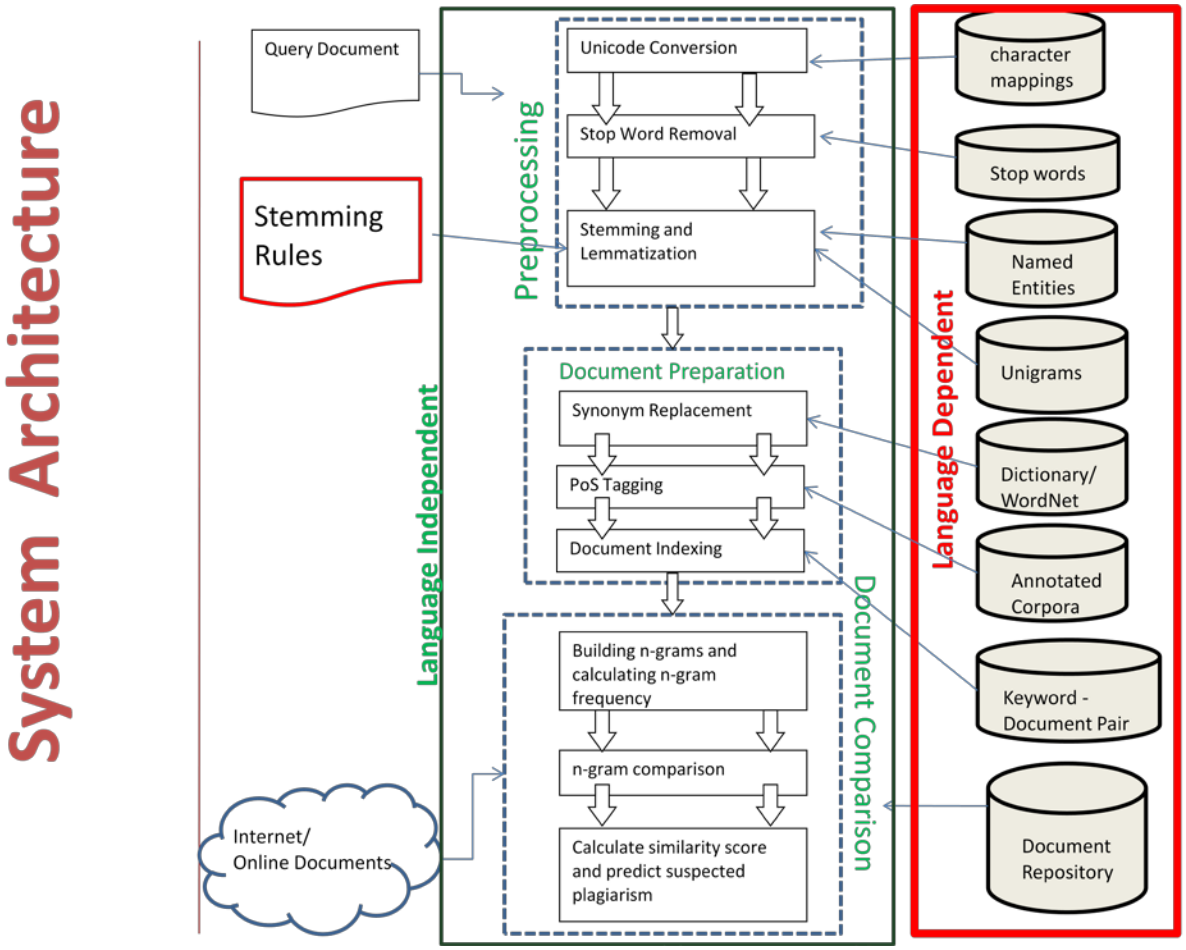

Fig. 2. Detailed Architecture of the proposed model for plagiarism detection

