Cross Lingual Text Reuse Detection Using Key phrase Extraction and Similarity Measures

by Rambhoopal K. Vasudeva Varma
Overview

- Introduction
- About CL!TR
- Related work
- Approaches
- Experimental Evaluation
- Conclusions and Future Work
Introduction

- What is Text Reuse?
  - Text reuse is an imitate of phrases from others text documents and presenting them as their own.

- How hard is the problem?
  - Reused text is commonly modified aiming to hide the plagiarism.
  - Not all similar text sections are examples of Text Reuse.
Introduction

- Text reuse detection across languages is even more harder.
- Detection of text reuse between distant language pairs is most difficult of all.
Related work

- Most approaches have addressed text reuse issue by measuring
  - Lexical and structural similarity of documents:
    - Individual words.
    - Fixed length substrings (e.g. n-grams).
    - Variable-length substring.
Related work

- Drawbacks of these approaches:
  - Their decision/classification based on one single feature such as degree of overlap.
  - Common domain-specific word sequences causes an overestimation of their overlap.
About CL!TR

- CL!TR (Cross language Indian Text Reuse) task is one variant of FIRE 2011 (Forum for Information Retrieval Evaluation).
- CL!TR task is to identify the text reuse documents given suspicious documents in Hindi and source document in English.
- We have used CL!TR task data collection for all our experiments to detect text reused documents across languages.
Approaches

- We implemented three approaches in text reuse detection
  - First: n-grams and cosine similarity.
  - Second: Features denoting common text relevance and fragmentation.
  - Third: Automatic extraction of Key phrases and new measure for similarity while using an open source search engine.
First Approach

- Tri-grams extracted using sliding window for both suspicious and source documents.
- Cosine Similarity between tri-grams of each suspicious document against all tri-grams of all source documents.
- Top similarity scored source document as text reuse source for creating suspicious document.
- J48 Decision tree classifier using WEKA tool with cosine similarity score as single feature for classification of documents.
Second Approach

- Set of features denoting common text fragmentation and relevance.
- Fragmentation feature is based on two basic assumptions
  - Longer the sequence the greater the evidence of text reuse.
  - More common the sequence, greater the evidence of text reuse (since long sequence are very rare)
Second Approach

- Relevance Feature is based on a measure of relevance of frequent sequences.
- This measure of relevance has two parts
  - We evaluate how frequent is the given sequence in the suspicious document to penalize frequent sequences because of their probability of being domain specific.
  - Castigates the sequence formed by words that are frequent in both documents.
Third Approach

- Use keyphrase instead of n-grams
  - Keyphrase are sequence of words that captures the main topics in the document.

- For automatic keyphrase extraction we use following techniques
  - n-gram filtration technique.
  - term weighting technique.
Third Approach

- **n-gram filtration technique**
  - use LZ78 data compression techniques with simple refinement steps.
  - use pattern filtration techniques.

- **term weighting technique**
  - use position of phrase in the sentence.
  - use position of sentence of the given phrase in the document.
Third Approach

- Use an open source search engine *Nutch* for similarity measure.
- Index all source documents, use Opic-scoring algorithm to calculate the document score.
Third Approach

- Source documents as index and keyphrases of given suspicious document as queries we retrieve all the relevant source documents.

- 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.
Third Approach

- 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 considered source of its text reuse.
- We use minimum threshold of 31(frequency score) to consider a document as source for text reuse.
- Threshold is based on the development corpus.
Experimental Evaluation: Corpus

• We used CL!TR data collection for all our experiments.

• Training data contains 198 suspicious documents in Hindi and 5032 source documents in English.

• Testing data contains 190 suspicious documents in Hindi and 5032 source document in English.

• Training data of 198 suspicious documents is as follows:
  • 130 are given as positive examples of text reuse documents.
  • 68 are given as negative examples of text reuse documents.
Experimental Evaluation

- We used Google translator for translating all suspicious documents in training and testing data.
- We translated Hindi suspicious documents to English.
- We used English stop word list in all our approaches to remove them from training and testing documents while pre-processing the documents and contains 173 entries.
- We used English Porter stemmer for first and second approaches.
Experimental Evaluation

- Evaluation
  - CL!TR testing data given 190 testing suspicious documents following are the number of text reused documents considered by our approaches.
    - First approach considered 117 suspicious documents as text reused.
    - Second approach considered 125 suspicious documents as text reused.
    - Third approach considered 147 suspicious documents as text reused.
Experimental Evaluation

- There were total of 15 runs submitted for CL!TR task in FIRE-2011.
- Following are the results of our approaches
  - Our third approach secured first rank with F-measure of 0.649 in CL!TR 2011 task.
  - Second approach secured third rank with F-measure of 0.608.
  - First approach secured 7th place.
Differentiator: Usage of keyphrases instead of n-grams and a new measure for similarity for text reuse detection.

Features denoting relevance and fragmentation of common sequences was also good but with less/poor recall.

Future work: Explore how semantic text feature applied across languages between distant language pairs could improve the F-measure.
Thank You