Set-based Similarity Measurement and Ranking Model to Identify Cases of Journalistic Text Reuse.

Arpan Pal, Lee Gillam
arpan.pal010@gmail.com, l.gillam@surrey.ac.uk
Department of Computing
University of Surrey
United Kingdom

Abstract. In this paper, we describe our approach to linking news articles in a cross lingual environment, English and Hindi, as submitted for the Cross-Lingual Indian News Story Search (CL!NSS)[1] task at FIRE'13. In our approach, English documents are first converted to Hindi using Google Translate[2], and compared to 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. A weighted combination of the five individual similarity scores provides an overall value for similarity. Results are promising, with a best Normalized Discounted Cumulative Gain (NDCG) to ranks 1, 5 and 10 (NDCG@1, NDCG@5, NDCG@10) of 0.6600, 0.5579, and 0.5604 respectively. These place the system in third by organization, and 5th by run.

Keywords: cross, lingual, text, similarity, rewrite, reuse, detection, natural, language, processing

1. Introduction:

Text reuse occurs when pre-existing texts or text segments are used to create new texts. Some popular methods of reuse can be duplication i.e. re-using the entirety of the text with little change, or fragmentation as in re-using part of the text, specific sentences or paragraphs etc. or derivation, where one or more sources are compiled into a new document.[3]

Text reuse is not necessarily a new phenomenon, but modern technologies make it ever easier to copy or modify from a large collection of documents. One source for such reuse is online news. Of course, some journalists will reuse their own content in subsequent articles and there is ready reuse as stories develop. For the news agencies, one pernicious reuse is the repurposing and republishing of news with neither attribution nor appropriate payment. This is harmful in two ways: those producing the articles are not necessarily properly rewarded for their efforts, whilst those who are obtaining such news through appropriate channels will have to absorb higher costs of operation than those not doing so, distorting the market. Properly syndicated news is encouraged, which readily allows for sourcing or referencing a number of different texts into one or more articles. News agencies would be expected to have at least two applications of interest here: checking proper syndicated uses, and determining improper usage.

Text reuse is readily exemplified elsewhere, but mostly either frowned upon or punishable. Genuine text reuse – with appropriate referencing – is key to the lineage of science. When attempts are made to mask such reuse, eventual discovery can lead to a variety of consequences[4]. For business, the reuse of intellectual property has been reported to have an impact of some $300bn per year, although such figures are entirely speculative[5]. Often, reuse with an attempt to mask goes beyond merely copy and paste, to involve translating, paraphrasing, summarizing or re-ordering in varying degrees to make the new text divergent from the original(s) – referred to by some as obfuscation[6]. A variety of approaches have been attempted in order to address detection in the face of such obfuscation.

An additional complication to such detection is translation. Automatic translation systems will, to varying degrees of success, convert between pairs of languages – and the chaining of pairs can produce a final (target) text somewhat divergent from the original (source). Such chaining can help to bring texts towards languages where cross-language resources are relatively scarce, but can hinder the detection capability. As such, identifying and linking news stories across languages becomes of interest for such detection, but also for being able to provide more information to the interested reader. Further, for countries with a number of regional languages (such as India), a national event covered in multiple languages becomes a great source of parallel or comparable data and as such becomes useful for NLP and IR tasks. The emergence of Cross-Lingual Information Retrieval (CLIR) suggests such possibilities and, given such linking, the possibility would then exist to be able to identify whether two arbitrary texts in two arbitrary languages had various shared characteristics, and so at a minimum to create better translation systems. The CL!NSS initiative looks to be a beneficial step in such a direction.

For CL!NSS, the challenge is to identify and link two articles which have identical or similar content, but are produced in two different languages in this case, English and Hindi. Related stories may have multiple authors and different perspectives of the same event, and so texts in the same language would be expected to have a number of similar words. Once such similarities are identified, the task is to link or group the articles according to their similarity. Once translated to a common language, depending on the quality of the translation, the task may resemble heavily obfuscated simulated or artificial plagiarism detection. The CL!NSS data-set for 2013 comprises 25 news stories in English and 50691 news stories in Hindi, and in the remainder of this paper we describe how we approached this task. In Section 2 we provide some additional background in relation to the Dataset task. Section 3 describes the subcategorisation of the task as done by CL!NSS. Section 4 describes our approach in detail, followed by our experiments and fine tune-ins in Section 5 that helped improve the results. We conclude this paper in Section 6 with our findings and possible improvements to the system in future.

2. CL!NSS Dataset:

The news recorded by press agencies can be written as being either about a single event or a follow up of an ongoing event. Similarly for news agencies, any news can be published in only one article, or a series of articles that describe the event as it develops. Thus any article can be categorised as below:

One-off Events: Events that occur only once, and are described by a single article.

Running Events: Events that continue throughout a certain timespan, reported on multiple articles.

To compare such articles, a common assumption would be that articles stemming from the same events are more comparable to each other. The 25 news stories in English and 50691 news stories in Hindi that make up the CL!NSS data-set for 2013 are classified as:

Focal Event: The main event of a singular/series of event(s) that provides detailed and specific information is considered the focus of the event, mostly being the very first article published on that event. Also, this kind of events are mainly written from a specific perspective.

Background Event: The role of this kind of events is to provide the context for the focal events, and also providing enough supporting information to help the user better grasp the perspective. These events include related event that are considered to be the causes of the focal event, similar events that occurred in past, and definitions or explanations of those things that play an important role in the event.

News Event: The whole of the event is considered as a complete news event. This includes all the focal, background and related events that may have been reported in multiple articles throughout a certain time limit. This is the interpretation of any real-world news covering all of a large event. Any and all articles that diverge from a particular event or topic share the same news event. Together they provide all the focus, background and context of the published event as a whole complete knowledge on that topic.

3. Task Categories:

The main goal of the task is to identify the same news event across multiple languages, and categorise the articles accordingly, i.e. extracting related documents or text segments, and furthermore to identify the level of co-derivation. Any two news articles can be compared if they belong to the same news story, but they may be describing either the same focal event, or two different ones. If they describe the same focal event, then we should expect some similarity between them, and the task extends to identifying the parallel content.

The scheme chosen by CL!NSS divides the task into the following categories:

Story Detection: Given the target document, finding a list of all other sources that cover the same incident, but in a different language.

Fragment Detection: Given a pair of similar (comparable) reports, the task is to extract parallel text fragments.

Story/Fragment Classification: Finding cases of co-derivation, i.e where a new report is uses another report as its source.

4. Approach

Numerous approaches exist that measure text similarity, and some have been applied in previous iterations of CL!NSS, including TF-IDF ranking models, key-phrase extraction, longest common subsequence (LCS), amongst others. Common techniques also include (sliding window) n-grams and semantic similarity matching. Our approach involves automatic translation and a ranking model based on measuring similarities between specific properties of the text: (i) title, (ii) the word content, and from the word content, we extract (iii) unique words, and (iv) frequent words; (v) publication date is used as a filter. In this approach, the documents can be indicated as similar by the unique words, and frequent words determine the subject of the document. The whole system comprises of three stages as described below.

1.1 Pre-Process:

This phase prepares the documents for matching and consists of two parts:

Translation: It is common in cross-lingual IR to address a common language and adapt back out. We decided to translate all the target documents into Hindi, hence subsequent matching is performed for Hindi. We relied on Google Translate for to provide this function.

Publication date filtering: News texts around an event will tend to have a similar publication date unless the event periodically evolves, is repeated, or is part of a larger continuous event, and several other reports are needed to describe all of the event. For the first situation, every re-occurrence of the event usually references the previous occurrence but has its own supporting data, and is considered repetitive, but a separate event nonetheless. For the second scenario, the new event will definitely have more information about itself simply because it is relatively new, and the old event is used for support. Thus it can be assumed that two news articles that have their publication dates close together have at-least some possibility of relating to the same event, or of being derived from such. For the experiment, each target news publication date was matched against all the source document publication dates, and differences logged. This information can be used as a score-boosting mechanism as well as a threshold value to reduce search space. In the end we decided to use this to boost the scores of documents that were published within 8 days of the target document.

1.2 Candidate Selection:

In this phase, documents are analysed to generate a list of candidates for each target. For each test, the target documents are compared with all of the source documents and the similarity scores logged. This process gathers candidates for the next phase. The tests in this phase are described below:

Title similarity: We assume that commonality in document titles suggests that two documents may originate from the same focal event, depending on the extent of commonality. For each document, the title is treated as a bag of words and similarity between titles is determined using the Jaccard Coefficient, i.e.

$$commonwords = words_{target} \cap words_{source}$$

$$totalwords = words_{target} \cup words_{source}$$

$$score = \frac{commonwords}{totalwords}$$

Fig 1 : Jaccard coefficient Similarity measure

Content similarity: Similar to title similarity, we assume that if a source document has a number of common words to the target document currently being analysed, they too may originate from the same news event. For each target document, the content was extracted and broken into list of words, with a Hindi stopwords list used to filter this, and then tested against source documents using the Jaccard coefficient.

Unique words: As a variant on the above, we compare only words that occur uniquely in both source and target documents, again using the Jaccard Coefficient.

Frequent words: As a further variant on content similarity, similarity is assessed between the frequent words. From a target document, a number of frequent words are extracted and then checked against the frequent words in each of the documents in the corpus, measuring similarity as above.

1.3 Post Processing:

After generating all the scores for each of the tests, the date difference, and the four types of similarity, namely the $score_{title}$, $score_{content}$, $score_{unique}$, and $score_{frequent}$, a single score is generated from which a ranked list is produced. A weighted average of all the scores is taken, similar to:

$$score_{final} = \frac{score_{title} + score_{content} + score_{unique} + score_{frequent} + value_{date}}{docs_{title} \cup docs_{word} \cup docs_{unique} \cup docs_{frequent}}$$

Fig 2 : Combining the scores

Also in this stage we generate the CL!NSS format results for evaluation.

5. Experimental Tuning and Results

Using the training data, we investigated the best formulation for our approach. For similarity, we also tried Dice and Cosine but decided that Jaccard was both sufficient and computationally light for our system. For publication date, the score is boosted by 1.0 if dates are within eight days. Variants of the overall scoring equation were tried, looking to provide optimal ranked match values.

For first and second test runs, we used the scoring as above and submitted the top 50 and top 100 documents. As would be expected, both scored identically (0.62 for NDCG@1, 0.5005 for NDCG@5 and 0.5221 for NDCG@10).

$$score_{final} = \frac{score_{title} + score_{content} + score_{unique} + score_{frequent} + value_{date}}{docs_{title} \cap docs_{word} \cap docs_{unique} \cap docs_{frequent}}$$

Fig 3 : Combining the scores penultimate equation

For our third run, we weighted the equation with the individual accuracy of each score, thus limiting and compensating the effect of each score. The new equation becomes:

$$score_{final} = \frac{accu_{title} \times score_{title} + accu_{content} \times score_{content} + \ldots}{docs_{title} \cap docs_{word} \cap docs_{unique} \cap docs_{frequent}}$$

Fig 4 : Combining the scores final equation with weighting

With help of the evaluation script[7] and qrels[8], we measured the accuracy of each of the filters, and replaced the values in the above equation with the NDCG@50 scores of the tests, as shown below:

$$score_{final} = \frac{0.16 \times s_{title} + 0.25 \times s_{content} + 0.24 \times s_{unique} + 0.31 \times s_{frequent} + 0.04 \times v_{date}}{docs_{title} \cap docs_{word} \cap docs_{unique} \cap docs_{frequent}}$$

Fig 5 : Combining the scores final equation with weighting values

This run achieved NDCG@1, NDCG@5, NDCG@10 of 0.6600, 0.5579, and 0.5604 respectively, our best scores, with reasonable computational efficiency.

6. Conclusions

The problem of journalistic text reuse stands out from numerous other kinds of reuse for one reason: there are several reasons why it can be entirely acceptable, and even encouraged. Hence news texts provide a rich seam for the investigation of such reuse. This extends to the acceptability of translation and publication elsewhere. Such collections similar characteristics to heavily obfuscated texts, to the point where source and target could become so diverged that they look like inherently different texts.

In our system for CL!NSS, simple arithmetical or set theory based measurements have been used to generate a set of candidate matches with minimal computation. Using overlap, the task of retrieval is readily converted into filtering given a threshold to generate the results required. Future work in this direction would initially involve systematically investigating the information gain afforded by the five scores to explore how far the approach could be pushed, and subsequently to determine whether other scoring approaches are more effective.

References:

[1] Cross-Language !ndian News Story Search - http://users.dsic.upv.es/grupos/nle/clinss.html [accessed 14/11/13]

[2] Google Translate - http://translate.google.com/ [accessed 14/11/13]

[3] PAN@FIRE'11 Overview - http://memex2.dsic.upv.es/workshops/2011/clitr/downloads/slides/PAN_FIRE-2011-overview-slides.pdf [accessed 14/11/13]

[4] Retraction Watch: http://retractionwatch.wordpress.com/ [accessed 14/11/13]

[5] The National Bureau of Asian Research (2013)“The IP Commission Report: The report on the theft of American Intellectual Propery”, online: http://www.ipcommission.org/report/IP_Commission_Report_052213.pdf [accessed 14/11/13]

[6] Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso (2010), “An Evaluation Framework for Plagiarism Detection”. Proceedings of the 23rd International Conference on Computational Linguistics (COLING). Beijing, China, Association for Computational Linguistics.

[7] CLINSS Evaluation perl script - http://users.dsic.upv.es/~pgupta/clinss/downloads/clinss12-eval.pl [accessed 14/11/13]

[8] CLINSS Evaluation qrel file - http://users.dsic.upv.es/~pgupta/clinss/downloads/clinss12-en-hi.qrel [accessed 14/11/13]