Pre-Retrieval based Strategies for Cross Language News Story Search

Aarti Kumar[1] and Sujoy Das[2]

[1]Research Scholar, Department of Computer Applications, MANIT, Bhopal
[2]Associate Professor, Department of Computer Applications, MANIT, Bhopal

{ aartikumar01, sujdas }@gmail.com

Abstract: The task of measuring text reuse and linking target and source documents becomes challenging if reused text has been translated in other language and even more challenging if translation language uses a different script and syntactical structure. This year, in CLINSS, focus was on journalistic text reuse between texts written in two different languages. This working note focuses on identifying text reuse between texts written in two different languages-English and Hindi and aims at evaluating the CLINSS dataset by linking 25 target English news stories with top 100 Hindi news stories out of corpora of 50691 stories. In this study two pre-retrieval strategy 1) Query formed using Proper Noun and 2) Query formed using words whose frequency is equal to or higher than average frequency are used to formulate the query. Query is translated using using either dictionary based or machine translation CLIR based approach before retrieving the documents. In Run 1 and 2 pre-retrieval strategies clubbed up with CLIR's dictionary based approach is used to link English news stories with Hindi news stories. In Run-3 instead of using dictionary based approach, freely available online Google translate [7] and online Changathi Hindi transliterater [8] is used as Hindi resources for translating/transliterating the query. This is our first attempt at CLINSS, and the attempt was to evaluate the performance of two pre-retrieval strategies and to compare the existing dictionary based and machine translation based approaches of CLIR. It is observed that dictionary based approach clubbed up with proper noun based pre-retrieval strategy performed better.

Keywords. Text reuse, transliteration, stopword removal, query formulation

1. Introduction

Text reuse has mainly been studied in Newswire or blog collections. Measuring accurately the amount of reused text between documents has got number of applications such as plagiarism detection, summarization etc [1]. The task of linking reused source documents to original target documents becomes challenging if reused text has been translated in other language and even more challenging if translation language uses a different script and syntactical structure. This year, in CLINSS, focus was on journalistic text reuse between texts written in two pairs of different languages: English-Hindi and English-Gujrati.

The Evaluation test corpus of CLINSS 2013 consisted of a set of 50691 potential source news stories S, written in Hindi, and a set of 25 target news stories T, written in English. The task given was to find and link news stories s in S which has the same news event and the same focal event as of corresponding news story t in T. This working note focuses on identifying text reuse between texts written in two different languages- English and Hindi and aims at evaluating the CLINSS dataset by linking 25 target English news stories with Hindi news stories and retrieving top 100 potential Hindi News stories.

2. Approach

In this study pre-retrieval strategy 1) Query formed using Proper Noun and 2) Query formed using higher frequency words whose frequency is equal to or higher than average frequency are used to retrieve the Hindi news stories. Once the query is formulated then either dictionary based or machine translation based approach is used to translate the query to Hindi. In Run 1 and 2 pre-retrieval strategies clubbed up with CLIR's dictionary based approach is used to link English news stories with Hindi news stories. In Run-3 instead of using dictionary based approach, freely available online Google translate [7] and online Changathi Hindi transliterater [8] is used as Hindi resources for translating/transliterating the query. The results of these strategies are compared in Section 4. Terrier 3.5 retrieval engine is used to index and retrieve the documents.

3. Experiment

The test data of CLINSS dataset of 2013 has been used for experiment for experimental studies. The corpus contains 50691 Hindi source news stories and 25 English target news in test data set and 50 in training data set. The whole corpus of source documents was included for indexing and retrieval purposes.

3.1 **Preprocessing:** For both the phases and all the three runs that we have submitted, all the words of <*title*> and <*content*> were extracted from each of the English document. Punctuation was removed at the time of tokenization and stop-words were also removed from <*content*> part only. No preprocessing was done on the <*title*> and all the words were taken as it is at the time of query formulation. Stop-words, verbs and adverbs were removed using a list of 430 stop-words [5] and 1514 verbs and adverbs [6] which were compiled from the web. Dates and numbers were also removed at time of preprocessing from both <*title*> and <*content*> as it was observed at the time of trial runs that query started drifting if one considers them.

3.2 **Query Formulation**: Query Words of <*title*> is common for each run and different set of query words based on the pre retrieval strategy is selected from the <*content*> for each of the runs at the time of query formulation. The pre-retrieval strategies of Run1 and 2 have been fully automated.

3.2.1 Pre-Retrieval Strategies and Dictionary Based approach for Run1 and 2

Run 1: In this run only Proper nouns are extracted from <*content*> of the English news story. The grammar rule, that proper noun begins with a capital letter, has been used to identify Proper nouns instead of using part of speech tagger.

The idea behind choosing proper nouns for formulating queries to retrieve the source documents is that they are the ones that are never changed while translating text and more so in news stories as they are important entities in any news.

Run 2: In this run only those words whose frequency is greater than or equal to the average word frequency of the <*content*>, has been selected at the time of query formulation. Taking words having greater than or equal to average word frequency for forming query words is considered in view of the fact that words which appear more than average number of times, some of the words out of those words must be of importance in catching the linked documents.

In both of these runs dictionary based approach is used for translating query in Hindi. Porter Stemmer [10] is used for stemming. The Shabdanjali dictionary [9] is used for translating English tokens to Hindi and only the first Hindi translation of each word is considered. The words that didn't have Hindi equivalent in Hindi Shabdanjali dictionary were transliterated using a transliterator developed by us. Our self developed transliterator is not a mature one as still it has not developed completely. The translated queries are submitted to Terrier retrieval engine [11] and top 100 documents are retrieved.

3.2.2 Pre-Retrieval Strategy and Machine Translation Based approach for Run3

Run 3: This run is same as that of Run-1 but instead of using dictionary based approach for translating query words machine translation based approach is used. Freely available i) online Hindi Google Translate [7] and ii) online Changathi Hindi transliterator [8] are used to translate/transliterate English query words to Hindi. Although some transliteration was also done by Google translate [7] but it failed to transliterate a few of the words that needed it. For those words Changathi Hindi transliterator [8] was used. The process was carried out manually. This manual intervention was with the purpose of getting the correct Hindi words and then comparing the results thus obtained, with our fully automated approaches used for Run 1 and 2.

3.3 **Indexing and retrieval:** Indexing of Hindi documents and retrieval of linked news stories in Hindi for each English document has been done using Terrier 3.5 [11] using TF-IDF ranking model.

4. Result

The comparative evaluation performance at NDCG@1, NDCG@5 and NDCG@10 respectively is shown in Table 1. Run-1 gives performance of 0.6, 0.545 and 0.5388 for NDCG@1, NDCG@5 and NDCG@10 respectively. Run-2 gives performance of 0.56, 0.4521 and 0.4828 for NDCG@1, NDCG@5 and NDCG@10 respectively. Run-3 gives performance of 0.5, 0.4803 and 0.4867 for NDCG@1, NDCG@5 and NDCG@10 respectively. It is observed that proper noun based pre-retrieval strategy clubbed with dictionary based CLIR approach has performed fairly well.

Run	NDCG@1	NDCG@5	NDCG@10
run-1-manit1	0.6	0.545	0.5388
run-2-manit1	0.56	0.4521	0.4828
run-3-manit1	0.5	0.4803	0.4867

Table 1. Comparative performance of the three runs

5. Conclusion

It is observed that dictionary based approach clubbed up with proper noun based pre-retrieval strategy performed better than other two runs in all the three cases. Run-3 which aimed at getting the right translation and transliteration for given query words, did not show a good performance at NDCG@1 level. In this study some of the pre-retrieval strategies to retrieve a subset of source Hindi documents from large corpus has been studied. The post processing techniques to link the exact news stories shall be studied in future.

Acknowledgement

We are thankful to Terrier group for providing us Terrier Retrieval Engine to carry out our research work.

One of the authors, Aarti Kumar, is thankful to Maulana Azad National Institute of Technology, Bhopal for providing her the financial support to pursue her Doctoral work as a full time research scholar.

References

1 Paul D. Clough, Department of Computer Science University of SheÆeld, England : Measuring Text Reuse in Journalistic Domain

2 Parth Gupta, Paul Clough, Paolo Rosso, Mark Stevenson: PAN@FIRE: Overview of the Cross-Language !ndian News Story Search (CL!NSS) Track. In:Forum for Information Retrieval Evaluation, ISI, Kolkata,India(2012)

3 Yurii Palkovskii, Alexei Belov: Using TF-IDF Weight Ranking Model in CLINSS as Effective Similarity Measure to Identify Cases of Journalistic Text Re-use In: Overview paper CLINSS 2012, Forum for Information Retrieval Evaluation, ISI, Kolkata,India(2012)

4 Nitish Aggarwal, Kartik Asooja, Paul Buitelaar, Tamara Polajanar, Jorge Gracia: Cross-Lingual Linking of News Stories using ESA. In:Overview paper CLINSS 2012, Forum for Information Retrieval Evaluation, ISI, Kolkata,India(2012)

5 List of Stopwords Available on `http://www.ranks.nl/resources/stopwords.html`, `http://norm.al/2009/04/14/list-of-english-stop-words/`, `http://www.webconfs.com/stop-words.php`, `http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop`

6 List of Verbs and Adverbs Available on `http://www.englishclub.com/vocabulary/regular-verbs-list.htm`, `http://www.momswhothink.com/reading/list-of-verbs.html`, `http://www.linguanaut.com/verbs.htm`, `http://www.acme2k.co.uk/acme/3star%20verbs.htm`, `http://www.enchantedlearning.com/wordlist/verbs.shtml`, `http://www.enchantedlearning.com/wordlist/adverbs.shtml`

7 `http://translate.google.com/?prev=hp&hl=en&text=&sl=en&tl=hi#en/hi/` -

8 Changathi Transliterator Available on `http://hindi.changathi.com/`

[9] Shabdanjali available on `http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html`

[10] Porter stemmer available on `http://ir.dcs.gla.ac.uk/resources/linguistic_utils/porter.java`

[11] Terrier 3.5 available on `http://terrier.org/download/`