Performance Evaluation of Dictionary Based CLIR Strategies for Cross Language News Story Search

Sujoy Das[1] and Aarti Kumar[2]

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

{ sujdas, aartikumar01 }@gmail.com

Abstract. Research on text reuse has gained momentum due to the fact that enormous amount of data is freely available on the web and can be used by anyone to express ideas. Cross language Journalistic text reuse is one such area where same news story may be presented in number of different languages and identifying test reuse across the language is a challenging task. This year in CLINSS focus was on journalistic text reuse between texts written in different languages. The test data comprised of 25 target English news stories with 50691 source Hindi news stories. This is our first participation at CLINSS task and we have submitted two sets of runs as MANIT-1 and MANIT-2. The idea of MANIT-2 runs is to test the performance of established simple dictionary based Cross Language Information Retrieval (CLIR) strategy on CLINSS task. The dictionary based approach has performed fairly well and has given a best performance of 0.5 for NDCG@1. The performances of all the strategies are in the range of 0.5 to 0.32 for different NDCG level.

Keywords: Cross Language Information Retrieval (CLIR), Dictionary based approach, transliteration, Part of Speech Tagger

1. Introduction

Research on text reuse has gained momentum due to the fact that enormous amount of data is freely available on the web and can be used by anyone to express ideas. Cross language Journalistic text reuse is one such area where same news story may be presented in number of different ways and also in different languages and there is a need to link such news stories as it offers number of benefits [1]. Abundance of such multilingual text can be of significant use for deriving number of language resources such as bilingual dictionaries, parallel corpora etc. This year in CLINSS focus was on journalistic text reuse between texts written in different languages. The test data comprised of 25 target English news stories with 50691 source Hindi news stories. The CLINSS task is to identify potential source news stories, written in Hindi, with same news and focal event from a set of target news stories that are written in English. It is our first participation at CLINSS task and we have submitted two sets of runs as MANIT-1 and MANIT-2. Dictionary, Parallel Corpora or Machine Translation based approach is primarily used for retrieving documents in the field of Cross Language Information Retrieval CLIR. The idea of MANIT-2 runs is to test the performance of established simple dictionary based Cross Language Information Retrieval (CLIR) strategies on CLINSS task.

2. Approach

Cross language news story search was studied by [1-3] in the last years CLINSS task. In this study dictionary based approach of CLIR is used to link English news stories with Hindi news stories. Performance of dictionary based Cross Language Information Retrieval System particularly across English-Hindi language was studied by [4]. The steps carried out to formulate the query before retrieving Hindi News Story is as follows:

Step i: Tokenization is applied on English news story, punctuations are removed and query is formulated using different strategies.
Step ii: Formulated query is translated using the Translation Module of English-Hindi dictionary based CLIR system.
Step iii: Translated query is submitted to Terrier retrieval system and top 100 Hindi News Stories are retrieved.

3. Experiment

CLINSS dataset of 2013 comprised of training and test data set. The training data set comprised of 50 English new stories whereas test data set comprised of 25 English news stories. The corpus contains 50691 Hindi source news stories. The task was to retrieve top 100 potential Hindi News stories. In all the runs the query was formulated using either <*title*> or using both <*title*> and <*content*> field of the target document. The source documents i.e. Hindi news stories were indexed using Terrier 3.5 TF-IDF ranking model [8].

Step 1 Preprocessing: Punctuations are removed at the time of tokenization before submitting the tokens to automated dictionary based English Hindi Cross Language Information Retrieval System.

Step 2 Query Formulation: Query is formed using either <*title*> *or* <*title*> and <*content*> of the target document at the time of query formulation. The strategies used for three runs are as follows:

Run 1: In this run query is formulated using only <*title*> field of the target document i.e. English documents. The tokens are submitted to dictionary based CLIR system. The query is formulated using following steps

Step i: Stop words are removed before query formulation.

Step ii: Remaining words are translated by English-Hindi dictionary based CLIR system using Shabdanjali dictionary [5]. The first available translation in dictionary is retrieved; if translation is not available in the dictionary then it is stemmed using Porter stemmer [6] before resubmitting it to dictionary based translation module.

Step iii: If word is still not translated using dictionary based translation module in (Step ii) then it is transliterated using transliterator developed by us. The developed transliterator is not mature and we are still working on it.

Step iv: Translated and transliterated query is submitted to Terrier retrieval system and top 100 documents are retrieved.

Run 2: In this run query is formulated using both <*title*> and <*content*> field of target document i.e. English document. The idea is to form query using content words that might be present in <*content*> field apart from the <*title*> field of the target document. In this run also stop word is removed before formulating the query. It goes through all the steps of Run-1 (Step i to iv).

Run 3: In this run query is formulated using only <*title*> of the target document i.e. English documents. This run is different from Run-1 as the query is tagged using Stanford part of speech tagger [7] before submitting it to the dictionary based translation module. The dictionary contains more than one translation for many of the English word(s) therefore the idea is to retrieve right meaning of the word (in right context) before submitting it to retrieval system. In this run stop word is not removed.

4. Result

The comparative evaluation performance at NDCG@1, NDCG@5 and NDCG@10 respectively is shown in Table 1. The performance reported for Run-1 is 0.32, 0.3654 and 0.3908 for NDCG@1, NDCG@5 and NDCG@10 respectively. The performance of Run-2 is 0.5, 0.4193 and 0.4626 and for Run-3 is 0.32, 0.3272 and 0.3544 for NDCG@1, NDCG@5 and NDCG@10 respectively. It is observed that in RUN2 in which both <*title*> and <*content*> are used for query formulation performed fairly well in comparison to RUN1 and RUN3.

Run	NDCG@1	NDCG@5	NDCG@10
run-1-manit2	0.32	0.3654	0.3908
run-2-manit2	0.5	0.4193	0.4626

run-3-manit2	0.32	0.3272	0.3544

Table 1. Comparative performance of the three runs

5. Conclusion

The dictionary based approach has performed fairly well and has given a best performance of 0.5 for NDCG@1. The performances of all the strategies are in the range of 0.5 to 0.32 for different NDCG level. The performance of Run-1 and Run-3 is more or less same. It is observed that dictionary based CLIR strategies are good for retrieving initial set of document from a large corpus but post processing techniques to link the exact news stories is needed to further improve the performance of the system.

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. 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)

2. 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)

3. 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).

4. Anurag Seetha, Sujoy Das, M. Kumar: Improving Performance of English-Hindi CLIR System using Linguistic Tools and Techniques. IHCI 2009: 261-271

5. Shabdanjali Dictionary Available at http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Shabdanjali/dict-README.html

6. M.F. Porter (1980). An algorithm for suffix stripping, in Program - automated library and information systems, 14(3): 130-137.

7. Part of Speech Tagger http://nlp.stanford.edu/software/tagger.shtml.

8. Terrier 3.5 available on http://terrier.org/download/