Cross Lingual Information Retrieval with SMT and Query Mining

Suneet Kumar Gupta[1] and Amit Sinha[2] and Mradul Jain[3]

[1]Department of IT ABES Engineering College, Ghaziabad, India

[2]Department of IT ABES Engineering College, Ghaziabad, India

Suneet0071@rediffmail.com

amitsinha@abes.ac.in

[3]Department of Computer Sc. & Engineering ABES Engineering College, Ghaziabad, India

mradul_ja@rediffmail.com

Abstract

In this paper, we have taken the English Corpus and Queries, both translated and transliterated form. We use Statistical Machine Translator to find the result under translated and transliterated queries and then analyzed the result. These queries wise results can then be undergone mining and therefore a new list of queries is created. We have design an experimental setup followed by various steps which calculate Mean Average Precision. We have taken assistance ship of Terrier Open Source for the Information Retrieval. On the basis of created new query list, we calculate the Mean Average Precision and find a significant result i.e. 93.24% which is very close to monolingual results calculated for English language.

Keywords

Information Retrieval, Mean Average Precision, Translation, Transliteration.

I. INTRODUCTION

When we retrieve the information from a given corpus it is known as Information Retrieval [1] and when our corpus and queries are in the same language it is known as mono-lingual information retrieval. In Cross–lingual Hindi- English Information Retrieval our Corpus is in English and queries are transliterated from Hindi to English [1][2].

Information retrieval (IR) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis, and technologies. IR is interdisciplinary, based on computer science, information science, linguistics and statistics.

Translation can be performed manually as well as with the help of machine. There are various tools available which is performed the translation these tools are known as "Statistical Machine Translator" as example "GIZA++"[3],"MOSES"[4] and Google Translator is also a very good Statistical Machine Translator(SMT) [6]. Transliteration means converting the statements word

DOI : 10.5121/acij.2011.2504

by word in other languages for the above statement “Anita eats mango”, the transliterated in the Hindi language is “अनीता खाती हैं आम”.

Monolingual retrieval always gives better results. It can because here the corpus and queries in the same language. Its higher probability that there are no out of vocabulary (OOV) words. In transliteration system obtained results is poor with respect to monolingual system due to some out of vocabulary words exist in our query due to transliteration.

II. DATA SET

We have experimented on data set taken from the FIRE (Forum for Information Retrieval and Evaluation) [5].three data files Corpus, Query file and Query Relevance file available are needed for our experiments.

A. Corpus

As written in section two corpus data, which has nearly 1.25Lakhs files with no typographical error, is taken from FIRE. These files are created from well known and reputed magazine TELEGRAPH and with the following format

```
<DOC>
<DOCNO>1041207_atleisure_index.utf8</DOCNO>
<TEXT>
</TEXT>
</DOC>
```

Where DOCNO tag represents the document number and our information is placed between TEXT tags.

B. Queries

Again we have downloaded the query file from FIRE and maintain results under 50 queries. The format of queries is as follows:

```
<top>
<num>76</num>
<title>Clashes between the Gurjars and Meenas</title>
<desc>
Reasons behind the protests by Meena leaders against the inclusion of Gurjars in the Scheduled Tribes.
</desc>
<narr>
The Gurjars are agitating in order to attain the status of a Scheduled Tribe. Leaders belonging to the Meena sect have been vigorously opposing this move. What are the main reasons behind the Meenas' opposition? A relevant document should mention the root cause(s) behind the conflict between these two sects.
</narr>
</top>
```

Here we require two queries- first in Hindi language and other is in English language for the cross lingual results.

C. Query Relevance

The query relevance, downloaded from FIRE is useful to know that relevant query and the corresponding file of corpus.

The format is:

```
76  0 1040901_nation_story_3702283.utf8 0
```

III. EXPERIMENTAL SETUP

We have implemented the experiment to find cross lingual result with respect to Terrier Open Source [7]. Terrier is a power full tool for Information Retrieval. At this stage we have corpus data query file and query relevance file with us. The following steps (section 3.1-3.7) are the necessary and continuous part of our experiment.

A. Translation of Hindi Queries into English Queries With the help of Statistical Machine Translator

In our experiment we have selected two forms of queries- first is transliterated form of query and another one is translated query. Here translation means we convert a given query list

Into its proper grammatical correct query. For example, for a sentence “अनीता आम खाती हैं”The English translation is “Anita eats mango”.

As above we have discussed about the format of query, there are mainly three things TITLE, DESCRIPTION, NARRATION. We have translated all the parts of queries from Hindi to English with the help of Google translator. One of which as follows:

<top>

<num>76</num>

<title> Clashes between Gujjars and Meena community </title>

<desc> Meena ST classified the Gujjars to protest leaders to appear </desc>

<narr>Gujjar community movement to make his scheduled tribes are classified ? Meena community leaders protest against it objected to the principal causes are Meenaoan ? Relevant documents both of these communities should be mentioned the main causes of conflict</narr>

</top>

B. Calculation of results under English Corpus and English Queries

In this step we have processed English query with relevance judgment under English corpus and calculate the following results. It is also known as base line results under English Corpus in monolingual system. For calculating the result we have used BB2C ranking model. Under this ranking model our Mean Average Precision is 0.3993. (see Table 1).

TABLE1. Result under English language/all without proper stemmer and stop-words list

Model	Mean Average Precision
InL2c	0.3865
BM25	0.3863
In_expC2c	0.3897
PL2c	0.3680
DFR_BM25	0.3859
IFB2c	0.3919
In_expB2c	0.3946

TF_IDF	0.3858
BB2C	**0.3993**

C. Calculation of results under English Corpus and Transliterated version of Hindi Queries

In this step we processed Transliterated queries [5] with relevance judgment under English Corpus and calculated the following results (see table 2). These queries are manually transliterated form of English query. Under the BB2C [7] ranking model the Mean Average Precision is 0.2299, (see Table 2).

TABLE 2.Results under Hindi to English transliterated query

Model	**Mean Average Precision**
PL2c	0.3680
DFR_BM25	0.3859
IFB2c	0.3919
In_expB2c	0.3946
TF_IDF	0.3858
BB2C	**0.3993**

D. Calculation of results under English Corpus and translated Queries generated by Google translator

In this step we first translate the query from Hindi to English with the help of one of SMT, e.g. Google Translator. There are some errors in all the queries during the translation. These queries are not same as existing queries of English language. Here we calculate the results of translated query with relevance judgment under English Corpus. The Mean Average Precision in this case is 0.3578. (See Table 3).

TABLE 3.Results under Hindi to English translation of query with SMT

Model	**Mean Average Precision**
InL2c	0.1968
BM25	0.2012
In_expC2c	0.2317
PL2c	0.2054
DFR_BM25	0.2006
IFB2c	0.2288
In_expB2c	0.2297
TF_IDF	0.2034
BB2C	**0.2299**

E. Analyze the results (Query wise) obtained in Sections 3.3 and 3.4

In Terrier (open source) there is a facility for analyzing our results query wise. As an Example in our

experiments there are 50 queries. Some of them give better result under Transliterated version of query and some of giving better result under translated version of query. We obtained the results. (See Figure1and Figure 2).

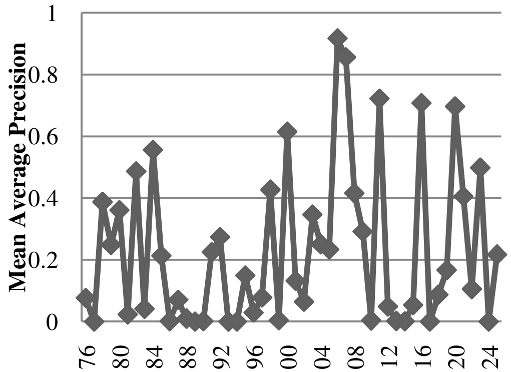

Figure 1.Querywise analysis for BB2C ranking model under transliterated queries

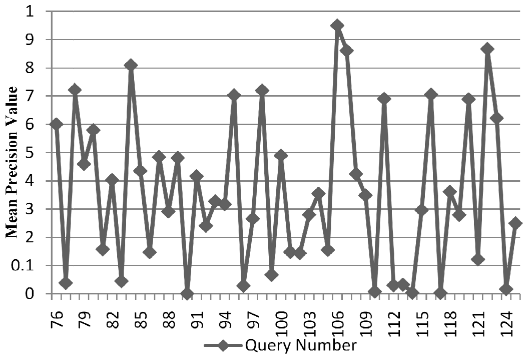

Figure 2. Querywise analysis for BB2C ranking model under translated queries

F. Apply the manual mining on results obtained in 3.5 and creation of new query file

We create a new query file. In this query file we place only those queries which had given the better results between transliterated query (see Figure 1) and translated query (see Figure 2). In given figure (see Figure 3) we can see that query number 76 is giving better result under translated condition. Similarly we can see for other queries. Now we create a new query file where we place some translated form of query and some transliterated form of query. This new query file is used for the results calculation, (see Figure 3).

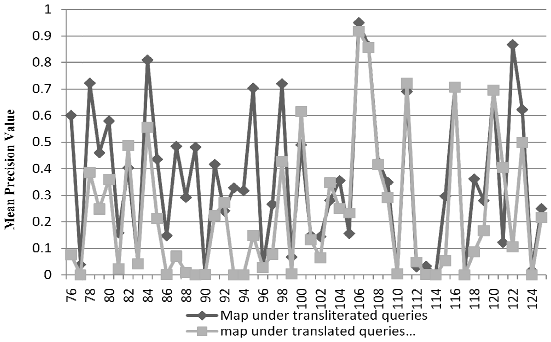

Figure 3.Result comparison between transliterated queries and translated queries

G. Calculation of results under English Corpus and Query list generated in 3.6

In this step we carried out result under English corpus, query list generated in previous step with query relevance file, we find that for the ranking model BB2C Mean Average Precision is .3723. Under the monolingual Information Retrieval for English Language Mean Average Precision is .3993.

I. APPLICATION AREA

India is the country of multilingual people means there are various languages as well as various culture. For communicating with others or sharing the information it is mandatory that both entities must communicate in common medium. In India, most of the people talk and share the information in local languages but simultaneously they know or rather understand English language. Suppose there is a problem with respect to the language we take the input from user (sender) passed the information through above proposed methodology then user(s) (receiver) receives, the information in respective language, which is very close to the monolingual retrieval under Hindi to English. This can be useful to other countries like India.

II. CONCLUSION AND FUTURE WORK

We worked in English monolingual and cross lingual tracks- Hindi-English .We experimented our system and then analyzed the results. our basic CLIR system is improved significantly by the two methodologies for handling OOV words – transliteration generation and mining. Significantly, we show that our cross lingual retrieval performance (that is enhanced with transliteration generation or mining) is nearly equal to that of our monolingual performance, validating our methodologies for handling OOV terms in the cross lingual retrieval.

The result of our experiment shows that if we have transliterated and translated form of the query then applying our proposed methodology, we can obtain much closer to monolingual results. The results can be improved by using the stemmer for transliterated queries. We can also propose disambiguation in transliterated queries to improve the result.

REFERENCES

[1] Yates R, Neto B .. Modern Information Retrieval. Addison Wesley 1999.

[2] CLEF. www.clef.org.

[3] www.fjoch.com/**GIZA++**.html.

[4] http://www.statmt.org/

[5] http://www.isical.ac.in/~fire/data_download.html

[6] www.google.com/translator

[7] www.terrier.org

[8] Khapra, M., Kumaran, A. and Bhattacharyya, P. 2010. Everybody loves a rich cousin: An empirical study of transliteration through bridge languages. In proceedings of NAACL 2010.

[9] NTCIR: http://research.nii.ac.jp/ntcir/.

[10] Majumder, P., Mitra, M., Pal, D., Bandyopadhyay, A., Maiti,S.,Mitra, S., Sen, A. and Pal, S. 2008. Text collections
for FIRE.Proceedings of SIGIR 2008.

[11] Li, H., Kumaran, A., Pervouchine, V. and Zhang, M. 2009.Report of NEWS 2009 Machine Transliteration Shared Task. Proceedings of the ACL 2009 Workshop on Named Entities (NEWS 2009), Association for Computational Linguistics, August 2009.

[12] Jagarlamudi, J. and Kumaran, A. 2007. Cross-Lingual Information Retrieval System for Indian Languag-es. Working Notes for the CLEF 2007 Workshop.

[13] Peter F. Brown , Vincent J. Della Pietra , Stephen A. Della Pietra , Robert L. Mercer, The mathematics of statistical machine translation: parameter estimation, Computational Linguistics, v.19 n.2, June 1993

[14] David Hull, Using statistical testing in the evaluation of retrieval experiments, Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, p.329-338, June 27-July 01, 1993, Pittsburgh, Pennsylvania, United States_ [doi>10.1145/160688.160758]

[15] Allan, J., Callan, J., Feng, F-F, and Malin, D. 2000. "INQUERY at TREC8." In TREC8 Proceedings, Special publication by NIST, 2000.

[16]. Douglas W. Oard, A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval, Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup, p.472-483, October 28-31, 1998.

[17]. L. S. Larkey, L. Ballesteros, and M. E. Connell. *Light Stemming for Arabic Information Retrieval*, volume 38 of *Text, Speech and Language Technology*, pages 221{243. Springer
Netherlands, 2007. ISBN 978-1-4020-6045-8.

[18]. W. Magdy, K. Darwish, O. Emam, and H. Hassan. Arabic cross- document person name normalization. In V. Cavalli-Sforza and I. Zitouni, editors, *Proceedings of the 2007 Work-shop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 25{32, Prague, Czech Republic, June 2007. Association for Computational Linguis-tics.URL: http://www.aclweb.org/anthology/W/W07/W07-0804.

[19]. Ashish Almeida and Pushpak Bhattacharyya, Using orphology to Improve Marathi Monolingual InformationRetrieval, FIRE 2008. Kolkata, India.

[20]. Paul McNamee, Textual Representations for Corpus-Based Bilingual Retrieval, PhD Thesis, University of Maryland Baltimore County, December 2008.