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

Presented by:
Sujoy Das & Aarti Kumar

Associate Professor & Research Scholar

Department of Computer Applications
MANIT, Bhopal
The CLINSS 2013 Task

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

A set of 50691 source Hindi news stories

A set of 25 target English news stories in test Data and 50 target English news stories in training Data Set.
Objective of Study

- To study dictionary based CLIR approach for CLNSS Task.
- To evaluate the performance of these strategies.
APPROACHES FOR CLIR

- Dictionary Based Approach
- Parallel Corpus Based Approach
- Machine Translation Based Approach
Dictionary Based Approach for CLINSS

Preprocessing

English Documents

Dictionary based CLIR System

Hindi Documents

Retrieval Engine

Retrieved Hindi Documents Top 100

Formulated Query

Stopword

Stemmer

Tagger

Dictionary

Language Resources
Dictionary Based Approach for CL!NSS

**Step i:** Tokenization applied on English news story, punctuations removed and query formulated using different strategies.

**Step ii:** Formulated query translated using the Translation Module of English-Hindi dictionary based CLIR system.

**Step iii:** Translated query submitted to Terrier retrieval system and top 100 Hindi News Stories retrieved.
Experiment

- Preprocessing
- Query Formulation
- Indexing and retrieval
Story is tokenized and punctuations are removed at the time of tokenization before submitting the tokens to Automated Dictionary Based English Hindi Cross Language Information Retrieval System.
MANIT-2-Runs

MANIT-2-Run1 Query Formulation using only Title Field and Dictionary Based approach.

MANIT-2-Run 2 Query Formulation using only Title and Content Field of News Story and Dictionary Based approach.

MANIT-2-Run 3 Query Formulation using only Tagged Title Field and Dictionary Based approach.
Query is formulated using only `<title>` field of the target document i.e. English documents.

Stop words are removed before query formulation.

Remaining words are translated by English-Hindi dictionary based CLIR system using Shabdanjali dictionary [5].

The first available translation in dictionary is retrieved for the given key word.

If translation is not available in the dictionary then it is stemmed using Porter stemmer [6] before resubmitting it to dictionary based translation module.

If word is still not translated using dictionary based translation module in (Step ii) then it is transliterated using transliterator developed by us.
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 MANIT 2-Run 1 (Step i to iv).
In this run query is formulated using only `<title>` of the target document i.e. English documents.

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.

This run is different from MANIT 2-Run 1 as the query is tagged using Stanford part of speech tagger [7] before submitting it to the dictionary based translation module.

In this run stop word is not removed.
Indexing and Retrieval

- Indexing of Hindi documents is done using Terrier 3.5[11]
- Translated and transliterated query is submitted to Terrier retrieval system and top 100 Hindi documents are retrieved using Terrier 3.5[11] TF-IDF ranking model.
## Comparison Relevance 2

<table>
<thead>
<tr>
<th>Story</th>
<th>manit-2 run-1</th>
<th>manit-2 run-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Result

The performance reported for MANIT 2-Run 1 is 0.32, 0.3654 and 0.3908 for NDCG@1, NDCG@5 and NDCG@10 respectively.

The performance of MANIT 2-Run 2 is 0.5, 0.4193 and 0.4626 and for MANIT 2-Run 3 is 0.32, 0.3272 and 0.3544 for NDCG@1, NDCG@5 and NDCG@10 respectively..

It is observed that in MANIT 2-Run 2 in which both <title> and <content> are used for query formulation performed fairly well in comparison to MANIT 2-Run 1 and MANIT 2-Run 3.
## Comparative performance

<table>
<thead>
<tr>
<th>Run</th>
<th>NDCG@1</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>run-1-manit2</td>
<td>0.32</td>
<td>0.3654</td>
<td>0.3908</td>
</tr>
<tr>
<td>run-2-manit2</td>
<td>0.5</td>
<td>0.4193</td>
<td>0.4626</td>
</tr>
<tr>
<td>run-3-manit2</td>
<td>0.32</td>
<td>0.3272</td>
<td>0.3544</td>
</tr>
</tbody>
</table>

*Table 1. Comparative performance of the three runs*
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 MANIT 2-Run 1 and MANIT 2-Run 3 is more or less same.

- At some places spelling variations created problem.
- The transliterator is to be improved.
Contd...

- Overstemming and understemming also created problem.
- 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 presenters, 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


References continued...

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


- Terrier 3.5 available on http://terrier.org/download/
THANKS