Ad-hoc Information Retrieval focused on Wikipedia based Query Expansion and Entropy Based Ranking

Utsab Barman Pintu Lohar Pinaki Bhaskar Sivaji Bandyopadhyay
Department of Computer Science and Engineering, Jadavpur University
Kolkata – 700032, India
{utsab.barman.ju, pintu.lohar, pinaki.bhaskar}@gmail.com, sivaji_cse_ju@yahoo.com

ABSTRACT

This paper presents the experiments carried out at Jadavpur University as part of the participation in the Forum for Information Retrieval Evaluation (FIRE) 2012 in ad-hoc mono-lingual information retrieval task for Bengali Hindi and English languages. The experiments carried out by us for FIRE 2012 are based on query expansion and entropy based ranking. The document collection for Bengali, Hindi and English contained 4, 57,370 , 3,31,599 and 3,92,577 documents respectively. Each query was specified using *title, narration and description* format. 100 queries were used for training the system while the system was tested with 50 queries in Bengali.

Keywords

Information Retrieval, Query Expansion, Ranking.

1. INTRODUCTION

The Forum for Information Retrieval Evaluation (FIRE) is mainly focused on Indian languages. In the Ad-hoc monolingual Information Retrieval research the systems accept queries in natural language and make an attempt to respond to the queries by providing relevant documents that contain the related information. The document could be text documents, images, audio or video documents. In this task, we have concentrated only on news text documents. The FIRE 2012 organizers provided the corpus and the query sets.

Various techniques have been used so far in the area of Monolingual Information Retrieval systems. These techniques can be broadly [1] classified as controlled vocabulary based and [2] as free text based systems at very high level. Very little work has been done in the past in the areas of IR and CLIR involving Indian languages. It is very difficult to create, maintain and scale a controlled vocabulary for general purpose IR systems in a general domain for a large corpus. Some of the earlier systems that were developed for Indian languages include cross language Hindi headline generation and English to Hindi question answering system [3].Our aim was to find out the most optimum algorithm for the ad-hoc mono-lingual information retrieval. Conceptually a general preprocessing technique like stemming, stop word removal, phrase detection, paragraph detection etc. was done for the query set and document set both. In our previous participation in Cross Language Evaluation Forum (CLEF 2007) [4] we proposed a semi-automatic query term list preparation. The International Institute of Information Technology (IIIT) in Hyderabad, India built a monolingual web search engine for various Indian languages, which is capable of retrieving information from multiple character encodings [5]. The Government of India has initiated a consortia project titled “Development of Cross–Lingual Information Access System” [6], where the query would be in any of the six different Indian languages (Bengali, Hindi, Marathi, Telugu, Tamil, Punjabi) and the output would be also in the language desired by the user. In our previous participation in FIRE 2008 [7] and FIRE 2010 [8] systems were proposed based on stemming, zonal indexing, theme identification, TF-IDF based ranking model and positional information. This time we add some extra features – query expansion using Wikipedia and entropy based ranking on the top of our previous retrieval models.

2. RELATED WORK

Query to an information retrieval system is the shortest way to express human information need. There are several ways to expand a query using global or local methods [9]. The use of WordNet [10] and sometimes POS tagging with it [11] had been used as a thesaurus earlier. There are several works based on thesaurus or automatically derived thesaurus but recent trends show us different ways of query expansion using Wikipedia as a resource. Wikipedia is one of the biggest human made knowledge resources available freely. Its structured way of representation allowed researchers to develop several methods of query expansion using Wikipedia. Some of these focused to the category information [12] of its articles. The link information [13] is also became a part for query expansion. As an example in cross lingual information retrieval the link analysis became an important part [14] and also in domain specific information retrieval [15]. In our work we used Wikipedia links as well as text part to gather information or related terms to expand a query. The processes are described in sec 5.

3. CORPUS STATISTICS

The corpus for ad-hoc mono-lingual retrieval was made available by the FIRE 2012 organizers.

3.1 Test Data

The data consists of ten consecutive years of news from the archives of the reputed newspapers: Anandabazar Patrika from Kolkata (Bengali) and BDNews24 from Bangladesh. The copyright issues for the newspaper corpus were taken up with the publishers by the FIRE 2012 organizers. Corpus was sub divided into many other sub-divisions like District, State, Sports, Editorial, and Government etc. The objective is to evaluate the effectiveness of retrieval systems in retrieving accurate and complete ranked lists of documents. The FIRE 2012 ad-hoc task focuses specifically on South Asian languages. We participated in the ad-hoc monolingual tasks for Bengali, Hindi and English. Corpus statistics for Bengali, Hindi and English are shown below in the table 1, table 2 and table 3 respectively.

Table 1. Bengali Corpus Statistics

Source Name	Anandabazar Patrika and BDnews24
Source URL	http://www.anandabazar.com/ and http:/www.bdnews24.com
Time-Period	1st JAN 2001 - 31st DEC 2010
Encoding Type	UTF-8
Number of documents in the corpus	4,57,370
Corpus Size(MB):	2.6 GB
Markup	<DOC> Starting tag of a document. <DOCNO> Contains document identifier. </DOCNO> : <TEXT> Contains document text. </TEXT> </DOC> Ending tag of a document.

Table 2. Hindi Corpus Statistics

Source Name	AmarUjala and Navbharat Times
Source URL	http://www.amarujala.com/ and http://navbharattimes.indiatimes.com/
Time-Period	1st JAN 2001 - 31st DEC 2010
Encoding Type	UTF-8
Number of documents in the corpus	3,31,599
Corpus Size:	1.2 GB
Markup	<DOC> Starting tag of a document. <DOCNO> Contains document identifier. </DOCNO> : <TEXT> Contains document text. </TEXT> </DOC> Ending tag of a document.

Table 3. English Corpus Statistics

Source Name	BDNews24 and The Telegraph
Source URL	http://bdnews24.com/ and www.telegraphindia.com/
Time-Period	1st JAN 2001 - 31st DEC 2010
Encoding Type	UTF-8
Number of documents in the corpus	3,92,577
Corpus Size:	1.0 GB
Markup	<DOC> Starting tag of a document. <DOCNO> Contains document identifier. </DOCNO> : <TEXT> Contains document text. </TEXT> </DOC> Ending tag of a document.

3.2 Topic

In FIRE 2012, different topics are present. Each of these topics is subdivided into four different parts: query identifier (num),a title (title), description (desc), and more details about the topic (narr). Table 4, table 5 and table 6 present a sample topic of the Bengali, Hindi and English respectively from FIRE 2012.

Table 4. FIRE 2012 Bengali Topic Number 177

<num>177</num>
<title>সঙ্গীতশিল্পীদের ভারত রত্ন</title>
<desc> সঙ্গীতশিল্পীদের ভারত রত্ন প্রদান করা হয়েছে, এই সম্বন্ধে তথ্য </desc>
<narr> প্রাসঙ্গিক নথিতে বিখ্যাত সঙ্গীতশিল্পীদের (গায়ক এবং বাদক যেমন রবি শংকর, এম. এস সুব্বুলক্ষ্মী এবং লতা মঙ্গেশকার) ভারত রত্ন পুরস্কারে ভূষিত হওয়ার বিষয়ে তথ্য এখানে প্রাসঙ্গিক। এই সঙ্গীতশিল্পীদের সম্পর্কে প্রবন্ধ (যেমন সংক্ষিপ্ত জীবনী, সঙ্গীতানুষ্ঠান পর্যালোচনা) সম্বন্ধে তথ্য অপ্রাসঙ্গিক, যদি না সেই নথিতে সঙ্গীতশিল্পীদের ভারত রত্ন গ্রহণ করার বা করবার উল্লেখ থাকে। </narr>

Table 5. FIRE 2012 Hindi Topic Number 177

<num>177</num>
<title> संगीतकारों को भारत रत्न</title>
<desc> संगीतकारों को भारत रत्न से सम्मानित करने की सूचना</desc>
<narr> प्रासंगिक प्रलेख में प्रसिद्ध संगीतकारों (गायकों और वादकों जैसे रवि शंकर, एम.एस.सुब्बालक्ष्मी, लता मंगेशकर) का भारत रत्न से सम्मानित होने से सम्बंधित सूचनाएं होनी चाहिये। इन संगीतकारों के बारे में लेख (जैसे संक्षिप्त जीवनी, संगीत कार्यक्रम की समीक्षा) प्रासंगिक नहीं हैं जब तक कि विशेष रूप से उल्लेख न हो कि संगीतकार ने भारत रत्न प्राप्त किया (या प्राप्त करेंगे)। </narr>

Table 6. FIRE 2012 English Topic Number 177

<num>177</num>
<title Musicians Bharat Ratna</title>
<desc>Information about the Bharat Ratna being awarded to musicians</desc>
<narr> Relevant documents should contain information related to famous musicians (including both vocalists and instrumentalists such as Ravi Shankar, M.S. Subbalakshmi and Lata Mangeshkar) being awarded the Bharat Ratna. Articles about these musicians (e.g. brief biographies, concert reviews) are not relevant unless they specifically mention that the musician received (or will be receiving) the Bharat Ratna.</narr>

4. PRE-PROCESSING

The documents and queries are preprocessed separately. The FIRE 2012 data is structured with well-defined tags. The preprocessing method includes tag removal, stop word removal and well defined topic/query word extraction.

4.1 Query Preprocessing

Three components of a query - title, description and narration are considered in this step. These components were processed to extract bag of words or extended query terms. Like our previous systems removal of stop words, suffix stripping [4] are the basic parts of the process. It ends up with a bag of words for each three parts of a query. In addition the named entities were also identified using a list look-up method.

4.2 Corpus Preprocessing

The news corpus made available by FIRE 2012 organizers is in XML format. A cleaning process was applied on the news corpus to extract the *title* and the *news body* from every document. The process of stop word removal first removes the stop words from the document and the suffix stripping module then removes the suffixes from every word for Bengali and keeps them in the original order of their occurrence in the document.

5. QUERY EXPANSION

In our experiment we considered that the exact expansion of a query would increase the relevance of retrieval. The query input in our IR system was a bag of words retrieved from the query pre-processing module (sec. 4.1). In the descriptive and narrative portion of a query the information need was specified, from where preprocessing module collected the bag of words. Hence the focus was to expand the query using the descriptive and narrative terms to find more relevant terms. The new terms with the previous would eventually represent the maximum relevance to the information needed.

Given a set of query terms (Q_i), a set of narrative query terms (Q_N) and another set of descriptive query terms (Q_D), we will try to find a set of related terms (Q_F), which will contain more keywords closely related to Q_i and ($Q_N \cup Q_D$). In other way it can be said Q_F is a representation of the keyword level distance between Qi and ($Q_N \cup Q_D$).

To obtain Q_F we considered the following steps -

1. Similar Term (Q_{ST}) extraction i.e. ($Q_i \cap Q_N \cap Q_D$).
2. Remaining Term (Q_{RT}) extraction i.e. $(Q_N \cup Q_D) - Q_{i.}$
3. According to our consideration Q_F is the distance between Q_{ST} and $Q_{RT.}$ To find the elements of Q_F we obtain the following technique –

   We used the Wikipedia (anchor text, external links etc.) as a resource for related term generation.
4. For the above method the target was to find the occurrence of any elements of Q_{RT} provided the initial search was started with an element of Q_{ST}. In the travel we will choose the top N relevant paths of new terms to reach from Q_{ST} to Q_{RT}.

The terms are then weighted and top K terms are chosen for the elements of Q_F.

5.1 Wikipedia Link Analysis

A query term can be expanded if there is a Wikipedia page with the query term as a title or in the page URL. As a matter of fact the structure of Wikipedia page gives us a good facility to retrieve some specific and relevant data related to the page topic/title. Here we consider - any hypertext and meaningful data in the content of a wiki page is related to the page topic i.e. the page title. We used the hyperlinks and some text portion of Wikipedia directly or by analyzing them as a resource of query expansion.

(Note: The header part of the html source of a Wikipedia page contains some SEO specific keywords. It is a script variable named WgCategory. During query expansion these keywords are blindly stored as expanded terms. They directly convey very much relevant information about the topic of a Wikipedia page.)

5.1.1 Wiki Query Generation

The parts of a named entity in query terms were considered as a single term. Rest of the terms was used as unigram and bi-grams with the named entities.

Let $Q_{ST} = \{Q_N^1, Q_N^2{}_, Q^3{}_, Q^{4,} Q^5\}$ where $\{Q_N^1, Q_N^2\}$ are the parts of a named entity. The queries to the Wikipedia to find a page would be as follows –

$\{Q_N^1, Q_N^2\}$

$\{Q_N^1, Q_N^2\} + Q^3$

…

$\{Q_N^1, Q_N^2\} + Q^5$

$\{Q_N^1, Q_N^2\} + Q^3{}_+ Q^4$

…

$\{Q_N^1, Q_N^2\} + Q^4{}_+ Q^5$

We ignored the immediate lowest gram if a higher gram with named entity returned a successful result. Any named entity found in the query terms would be considered separately for wiki search.

5.1.2 *Hypertext Chains (HC)*

After wiki query generation from Q_{ST} we did the following process for obtaining the elements of Q_F -

```
Initially   Term Count (T) = 0;
            Depth Count (DC) = 0;
            Depth Constant (DK) = Integer Constant;
            Link Chain = Empty;

            For a wiki-query if a wiki page is found repeat the
            following steps until DC ≥ DK.
    1.      Store the hyperlinks of the page.
    2.      For each links do the following
            a.   If a hyper-text or page contains any element of $Q_{RT}$
                     T = Count of unique $Q_{RT}$ terms;
                     Add the hypertext in Link Chain
                     DC++;
            b.   Else visit the link and repeat step 1.
```

When the process is finished for all the generated wiki queries from Q_{ST}, we have a set of hypertext chains, with count of depth (VC) and count of Q_{RT} terms (T) at each level of depth. From these chains we will choose the elements of $Q_{F.}$ The hypertext chains are then weighted as follows –

$$Chain\ Weight\ \left(C_{W\_link}\right) = \sum_{i=1}^{n} \frac{T_i}{|Q_{RT}|*DC_i}$$

i = the no. of depths

$|Q_{RT}|$ = No. of terms in Q_{ST}

T_i = No. of Covered terms (coverage based on Q_{ST}) at depth i

DC_i = depth at i'th level

Form the weighted chain top N chains are selected to represent $Q_{F.}$ Not all the terms of the chains are selected for Q_F. We propose another process to refine the search for the elements of Q_F.

5.2 Wikipedia Text Analysis

The process is an extended part of link analysis. Many hyperlinks in Wikipedia stay in the content portion as a part of a sentence. While storing the hyperlinks in the previous process we also stored those sentences in which the links were found. With the construction of link chain – automatically the sentences were chained in the same manner. Again like previous process we have a set of sentence chain for the generated wiki queries.

5.2.1 *Sentence Chain (SC) For Bengali & Hindi*

For a wiki-query (generated from Q_{ST}) if a wiki page is found, we did the following to construct a sentence chain for a single wiki-query –

```
Initially   $Q_{RT}$ Term Count (T) = 0;
            $Q_{ST}$ Term Count (L) = 0
            Depth Count (DC) = 0;
            Depth Constant (DK) = 1;
            Sentence Chain = Empty;
            Store the hyperlinks of the page.
Repeat the following steps until DC≥ DK.
    1.      For each link do the following
            a.   If a hyper-text or sentence contains any element of
                 $Q_{RT,}$ store the sentence (containing the link) into
                 Sentence Chain. Remove Stop words. Do Suffix
                 Stripping to obtain keywords.
                 T = Count unique $Q_{RT}$ terms present in the
                 sentence;
                 L= Count unique $Q_{ST}$ terms present in the sentence;
                 DC++;
            b.   Else visit the link and repeat step 1.
```

Likewise the previous after finishing for all the queries we have a set of sentence chains. The chain weight is calculated as -

$$Chain\ Weight\ \left(C_{W\_sentence}\right) = \sum_{i=1}^{n} \frac{T_i*L_i}{|Q_{RT}|*|Q_{ST}|*DC_i}$$

i = the no. of depths

$|Q_{RT}|$ = No. of terms in Q_{RT}

$|Q_{ST}|$ = No. of terms in Q_{ST}

T_i = No. of Covered terms (coverage based on Q_{RT}) at depth i

L_i = No. of Covered terms (coverage based on Q_{ST}) at depth i

DC_i = depth at i'th level

We considered both Q_{RT} and Q_{ST} here. The reason behind it is simple. During link analysis the generated chain will contain very little number of keywords in compression with sentence chain. A sentence might also have keywords that are not closely related with the query. To narrow the scope of relevance we considered the elements of both sets. It would be also tough to handle a huge number of expanded terms in our IR system.

Another point of argument was depth constant (DK). The reason behind DK =1 (i.e. single depth) was to minimize the time and the complexity of the system. Visiting more depth would make a huge number of expanded query terms, which was very difficult to handle with the system.

From the weighted chains again top K chains are obtained to represent $Q_{F.}$

5.2.2 *Co-Reference Chain (CC) For English*

Anaphora resolution and resolution of co-reference of a text gives us valuable information about the entities in the text. As an example considers the following text – "Manmohan Sing is the PM of India. The prime minister is now having a meeting." After resolution the co-reference we will have the keywords 'PM of India' and 'prime minister of India' as a referral terms of the entity 'Manmohan Sing'. Thus resolution of co-reference will eventually give us more related and relevant keywords of an entity. In our experiment we choose to do the co-reference resolution at depth 1 of a Wikipedia page. We used the tool Stanford Core-NLP to get the Co-Reference Chains of a document. An example of sample text and the Co-reference chain output is presented below.

Table 7. Example of Co-Reference Chain

CHAIN4-["the satanic verses controversy , also known as the rushdie affair ," in sentence 1, "the satanic verses controversy" in sentence 1, "the heated and frequently violent reaction of some muslims" in sentence 1, "the rushdie controversy" in sentence 7] CHAIN10-["the publication of salman rushdie 's novel the satanic verses , which was first published in the united kingdom in 1988" in sentence 1] CHAIN12-["salman rushdie 's novel the satanic verses , which was first published in the united kingdom in 1988" in sentence 1] CHAIN15-["the satanic verses , which was first published in the united kingdom in 1988" in sentence 1, "the satanic verses" in sentence 1]

As shown in the table 7, a chain starts with the entity as chain head, followed by the terms which refer to the entity. Here we select some chains based on the coverage of the elements in Q_{RT} and Q_{ST}. The co-reference chain selection strategy of our experiment is as follows

> Top K chains which have covered the maximum elements in Q_{RT} and Q_{ST} are selected.

Likewise the previous section we also calculate a co-reference chain weight as follows –

$$Chanin\ Weight\ (C_{W\_coref}) = \frac{T_i * L_i}{|Q_{RT}| * |Q_{ST}|}$$

From the weighted chains again top K chains are obtained to represent Q_F. As there are no such tools present for Bengali and Hindi, we followed the process only for English.

The obtained terms from chains are simply added to the set of expanded query terms.

5.3 Wikipedia Junk Filter

For the above processes especially in the link analysis some junk texts (default Wikipedia links) are automatically incorporated with the chains. A list of 56 junk texts was prepared. The chain elements are removed which contain the junk texts. The filtration was executed just after a chain creation.

5.4 Chain Merging

Now we have 2 types of weighted chains of related elements with Q_{ST} and Q_{RT}. One is sentence chain and another is hyper-link chain.

5.4.1 Hybrid Chains (HyC = SC+ HC)

We will construct a hybrid chain if any element of HC resides in any element of SC. We constructed the sentence chain with the sentence that had hyperlinks. So it was quite obvious that in the elements of SC some of the elements of HC would reside. Those would not reside, were collected from another depth. We would join only those elements of HC that were not present in SC.

The construction of a HyC will give us a strong chain with relevant and related elements. It would be as described below -

Let a SC = $\{Sr^1, Sr^2 \ldots Sr^i \ldots Sr^n\}$

And any HC = $\{Hc^1 \ldots Hc^i, Hc^{i+1} \ldots. Hc^n\}$

Let $(Hc^1 \ldots Hc^i)$ is present in the elements of SC.

$(Hc^{i+1} \ldots. Hc^n)$ is not present in SC.

Then we will construct a hybrid chain

HyC = $\{Sr^1 \ldots Sr^n, Hc^{i+1} \ldots. Hc^n\}$

Obviously the process modifies the chain weight of SC defined as follows –

$$C_{W\_hybrid} = (C_{W_{sentence}} + \sum\nolimits_{j=i+1}^{n} \frac{T_j}{|Q_{RT}| * DC_j})$$

After weighting, top weighted chain was retrieved and the elements are of the chains are selected as the elements of Q_F. The collected elements are then used as a bag of words in our IR system.

6. RETRIEVAL SYSTEM

The retrieval model we used is an extended part of our previous work. In our previous system theme cluster based technique was incorporated with the lucene IR framework. This time we added entropy based ranking functionality, which will re-rank the obtained results from lucene based on the average probability distribution of query, words and the expanded query terms.

6.1 Lucene Based Retrieval

Lucene is a standard IR system of Apache software foundation. It works well with a bag of words instead of a question or a query in natural language. The set of expanded query term is a fine representation of a bag of words. The process to retrieve them is described in sec. 5. For each modified query a list of documents were retrieved. We strictly did AND search in lucene to obtain the most relevant documents because expanded query terms to the IR system represent the most relevant set of keywords to the information need. The results obtained were re-ranked with an entropy based ranking model stated below

6.2 Entropy Based Ranking

During navigational search the common attitude of a user is to navigate around all search results to fulfill the information need. A prediction about the amount of unique information of a webpage in the search result will reduce the human effort. The navigation will be selective rather than visiting all results. We propose a novel method and a model is to find the distribution and measure of unique information of a result obtained from a search. The measure of information is based on the theory of entropy in Information Theory proposed by Claude E Shannon [16].

The information in a document is mostly related to different entities, keywords, concepts and events present in the document. Here these simply refer to the expanded query terms present in a document. Two different documents having occurrence of same expanded query terms may have the same information to express. Our target is to distinguish and sort the search results obtained from lucene[1], based on the presence of unique information (expanded query terms).

The obtained search results from the IR System are re- ranked with an entropy-based model. The model is simply calculates the average of probability distribution of query terms or expanded query terms for each result. The classic model of entropy is used here.

Let D be the document. The probability of occurrence of an expanded query $term\ (Qi)$ is $P(Qi)$. According to Shannon's theory the information obtained from this query term is

$$log\left(\frac{1}{P(Qi)}\right)$$

[1] http://lucene.apache.org/

Now for N observations of a query term (Qi), it will occur (N*P (Qi)) times. Hence the total information -

$$I = N * P(Qi) * log\left(\frac{1}{P(Qi)}\right)$$

$$\frac{I}{N} = \sum_{i=1}^{n} P(Qi) * log\left(\frac{1}{P(Qi)}\right)$$

$$H(D) = \sum_{i=1}^{n} P(Qi) * log\left(\frac{1}{P(Qi)}\right)$$

This is the measure of average information per query term or expanded query term of the document or the entropy of the document - $H(D)$.

Among the N retrieved results from a search engine, K of them are/is sufficient enough to fulfill the information need of a query Q. In (N-K) results same information are present in repetitive or distributive manner. The goal is to distinguish K results. We simply followed the process to get the K documents.

Initially

Expanded Query Term Storage (EQTS) = NULL;

Do the following for each result.

1. Calculate H(D) based on the presence of expanded query terms.
2. Ignore the terms in calculation that are present in EQTS.
3. Add the new terms in EQTS.

After finishing the process sort the N results based on their H(D) score.

By this process we re-ranked the obtained results. The re ranking is simply based on the average distribution of information of unique expanded query terms.

7. EVALUATION AND RESULTS

The runs for Bengali, Hindi and English were submitted as the part of ad-hoc monolingual retrieval task. Three Bengali runs, two Hindi runs ad two English run were submitted. The submitted runs were evaluated with TREC evaluation tool. The evauation scores of the best run of each of the three languages are shown in table 8. Only the following evaluation metrics have been listed for the run: mean average precision (MAP), geometric mean average precision (GM-AP), R-Precision (R-Prec) and binary preferences (Bpref).

Table 8. Bengali Evaluation Result

Language	Qrel	MAP	GM MAP	R-Prec.	BPref.
Bengali	v1	0.0438	0.0015	0.0639	0.0976
	v2	0.0428	0.0017	0.0654	0.0809
Hindi	v2	0.1149	0.0162	0.1437	0.1351
English	v1	0.1104	0.0035	0.1333	0.1523
	v2	0.1109	0.0040	0.1347	0.1422

8. CONCLUSION & FUTURE WORK

Clearly after doing query expansion the strict AND search was the main reason for low scores. The avarage document retrived in our all the seven run are 415 documents. As the qrel cointains 1000 document for each query, we did not able to retrive such amount of document with the AND search.

The approach of list based named entity detection and filtering is also another important area to look through. However the expansion technique is useful for some of the queries. But as part of this ad-hoc mono lingual retrieval it needs more improvement. The sentence chain method can be replaced with any anaphora or co reference resolution technique to obtain more specific referral terms to an entity from wiki text. Such good tool or system is not present in Bengali and Hindi language. In future we will try to develop such tool for Bengali. The hyperlinks often contain some unavoidable junks to identify. It caused also junk sentences addition in sentence chain. The structural information of wiki text could be incorporated with the expansion technique and could increase the performance of the system. Identification of temporal data in query and in document could have a positive effect in results. As there is a machine translation system present in our group, in future we will try to incorporate these techniques with the MT-System and we will take part in cross lingual retrieval tasks.

9. ACKNOWLEDGMENT

The work has been carried out with support from Department of Electronics and Information Technology (DeitY), MCIT, Govt. of India funded Project Development of “Cross Lingual Information Access (CLIA)” System Phase II.

10. REFERENCES

[1] Oard, D.: Alternative Approaches for Cross Language Text Retrieval. In: AAAI Symposium on Cross Language Text and Speech Retrieval, USA (1997)

[2] Dorr, B., Zajic, D., Schwartz, R.: Crosslanguage Headline Generation for Hindi. ACM Transactions on Asian Language Information Processing (TALIP) 2(3), 270–289 (2003)

[3] Sekine, S., Grishman, R.: Hindi-English Cross-Lingual Question-Answering System. ACM Transactions on Asian Language Information Processing(TALIP) 2(3), 181–192 (2003)

[4] S. Bandyopadhyay et al.: Bengali, Hindi and Telugu to English Ad-Hoc Bilingual Task at CLEF 2007

[5] Pingali, P., Jagarlamudi, J., Varma, V.:Webkhoj: Indian Language IR from Multiple Character Encodings. In: WWW 2006: Proceedings of the 15th International Conference on World Wide Web, pp. 801–809 (2006)

[6] CLIA Consortium: Cross Lingual Information Access System for Indian Languages. In: Demo/Exhibition of the 3rd International Joint Conference on Natural Language Processing, Hyderabad,India, pp. 973–975 (2008)

[7] S. Bandyopadhyay, A. Das and P. Bhaskar: English Bengali Ad-hoc Monolingual Information Retrieval Task Result at FIRE 2008. In: Forum for Information Retrieval Evaluation (FIRE), Kolkata, India (2008)

[8] S. Bandyopadhyay, P. Pakray, A. Das and P. Bhaskar: Theme Based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010. In: Forum for Information Retrieval Evaluation (FIRE), Kolkata, India (2010)

[9] Manning, C. D., Raghavan, P., and Schutze, H. (2008). Introduction to Information Retrieval. Cambridge University Press

[10] Voorhees, E. M. (1994a). Query expansion using lexical-semantic relations. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '94, pages 61-69, New York, NY, USA. Springer-Verlag New York, Inc.

[11] Smeaton A. F., Kelledy F., and O'Donnell R. (1995). Thresholding posting lists, query expansions with wordnet and pos tagging of spanish. In The Fourth Text Retrieval Conference TREC-4, pages 373-390.

[12] Li, Y., Luk, W. P. R., Ho, K. S. E., and Chung, F. L. K. (2007). Improving weak ad-hoc queries using wikipedia as external corpus. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, pages 797{798, New York, NY, USA. ACM.

[13] Kaptein, R. and Kamps, J. (2009). Advances in focused retrieval. Chapter Finding Entities in Wikipedia Using Links and Categories, pages 273-279.Springer-Verlag, Berlin, Heidelberg.

[14] Hsu, C. C., Li, Y. T., Chen, Y. W., & Wu, S. H. (2008). Query expansion via link analysis of wikipedia for clir. Proceedings of NTCIR-7.

[15] Müller, C., & Gurevych, I. (2009). Using wikipedia and wiktionary in domain-specific information retrieval. Evaluating Systems for Multilingual and Multimodal Information Access, 219-226.