Global Scientific JOURNALS

GSJ: Volume 5, Issue 11, November 2017, Online: ISSN 2320-9186
www.globalscientificjournal.com

SINDHI TO ENGLISH CROSS LANGUAGE INFORMATION RETRIEVAL SYSTEM

Naadiya Mirbahar , Mutee-U-Rehman, Saajid Hussain

Abstract

The process of Information Retrieval (IR) helps a prospective user to find the required knowledge either from www or from corpus. Cross Language Information Retrieval (CLIR) is a task of identifying documents written in language different than that of the user specified query language. In this globalization era and continued internationalization of internet, the growing multilingual contents, motivate researchers to cope up with the problem of CLIR. A number of systems have been developed over past few years for English and other European languages. However, no work has been carried out on Cross Language Information Retrieval System for Pakistani languages in general and Sindhi in particular. Thus searching in regional languages will undoubtedly lead to a more precise solution. The proposed system implements this concept. To do so, a Cross Language IR system for Sindhi and English is developed with Query translation based upon the bilingual dictionary.

The performance evaluation results show that developed model reduces the incorrectness of result. It is found that the proposed model successfully retrieve better results for the CLIR Sindhi to English than the existing search engines.

Keywords CLIR System. Stemming. Tokenization. Stop word removal. Stem Dictionary.

1. Introduction

Information retrieval (IR) is the process of getting relevant document out of a hoard of documents based on user's query. IR system makes it possible to obtain the documents which are likely to contain the required information related to the query (Wang and Oard, 2006). Since 1960s, keyword searching has been used for text retrieval. Two main methods i-e., Hypertext and Hypermedia were being used for crisscrossing database and identical words in the queries. The development of latest search engines such as Google, Ask, and Yahoo etc. on internet combined with keyword searching, NLP (Natural Language Processing) and hyperlinks has made it easier to retrieve information.

Cross Language Information Retrieval (CLIR) is concerned with the issue of obtaining information in a language which is different from that of user (Saraswathi et al., 2010). The approach of cross language retrieval explicitly removes the linguistic barriers from monolingual Information retrieval. The documents in multilingual storage are in various languages, and the text in these documents is usually in two or more languages. The approach is Cross lingual if it deals with just two languages, i.e. one source (Sindhi for example) and one target or document language (English for instance) or vice versa.

IR and CLIR systems with various approaches have been developed for different languages rich in resources, including English-Chinese CLIR system (Zhou et al., 2008), Telgu to English CLIR system (Pingali et al., 2006), Japanese-Chinese CLIR system (Hasan and Matsumoto 2000), English to Spanish (Sheridan and Schauble, 1997) with very high accuracies. However the languages with poor resources lack such systems. This leaves an open research area to work on such systems for poor resourced languages like Sindhi. The proposed Cross Language Information retrieval system will have positive impact on research and development of Sindhi IR systems. The remaining paper is organized as follows; Motivation, Design of proposed Sindhi to English CLIR system, Experimental scenarios and results. Finally conclusion and discussion followed by references.

2. Motivation

The Websites are growing in number with various languages on WWW with English content being dominant on web. Due to lack of CLIR systems users are unable to retrieve information written in required language other than English. It has been reported that over 4000 languages are being spoken in the world. Sindhi is an Indo Aryan-language, and according to the World Sindhi Institute, Sindhi is spoken by more than 40 million people, majority of whom live in Pakistan followed by India, and by Sindhi immigrants which live in several other Asian, European and North American countries. Most of the existing systems offer a search for the information in an outfit of limited languages leaving the user dissatisfied and wanting for further information in a prosperous manner.

CLIR system development considering Sindhi and English is not even initiated. Research and development efforts need to be initiated in Sindhi to English CLIR which will provide basis for further research and system development in the future.

3. Proposed model

The proposed CLIR emphasises on analysis and implementation of Sindhi-English Cross-Language Information Retrieval System based on dictionary based query translation, and translates the Sindhi query into English. The system uses a stemmed Sindhi –English dictionary to perform query translation. Stemmed Dictionary is basically developed to resolve the problem of stemming and translation of the query processed keywords (Sindhi) to target language (English).

Sindhi-English CLIR system is categorized into three modules.

I) Text processing stage, which deals with dividing Sindhi query into small tokens. Stop word removal is a step of removing pronoun and prepositions from the query. To do this, stop words list is also developed. Stemmer is built to obtain the root words, avoiding the inflected and derived words.

II) Verification module takes output of text processing module as input and then searches the processed source query terms in the bilingual stemmed dictionary. The exact matched words found in the dictionary are handed over to the translation module. This module is responsible for the formation of query in target language.

III) The query is sent to the IR engine and the result is retrieved.

Fig. 1 shows the architecture of proposed CLIR system. The module and sub-modules of this system are described in subsequent sections.

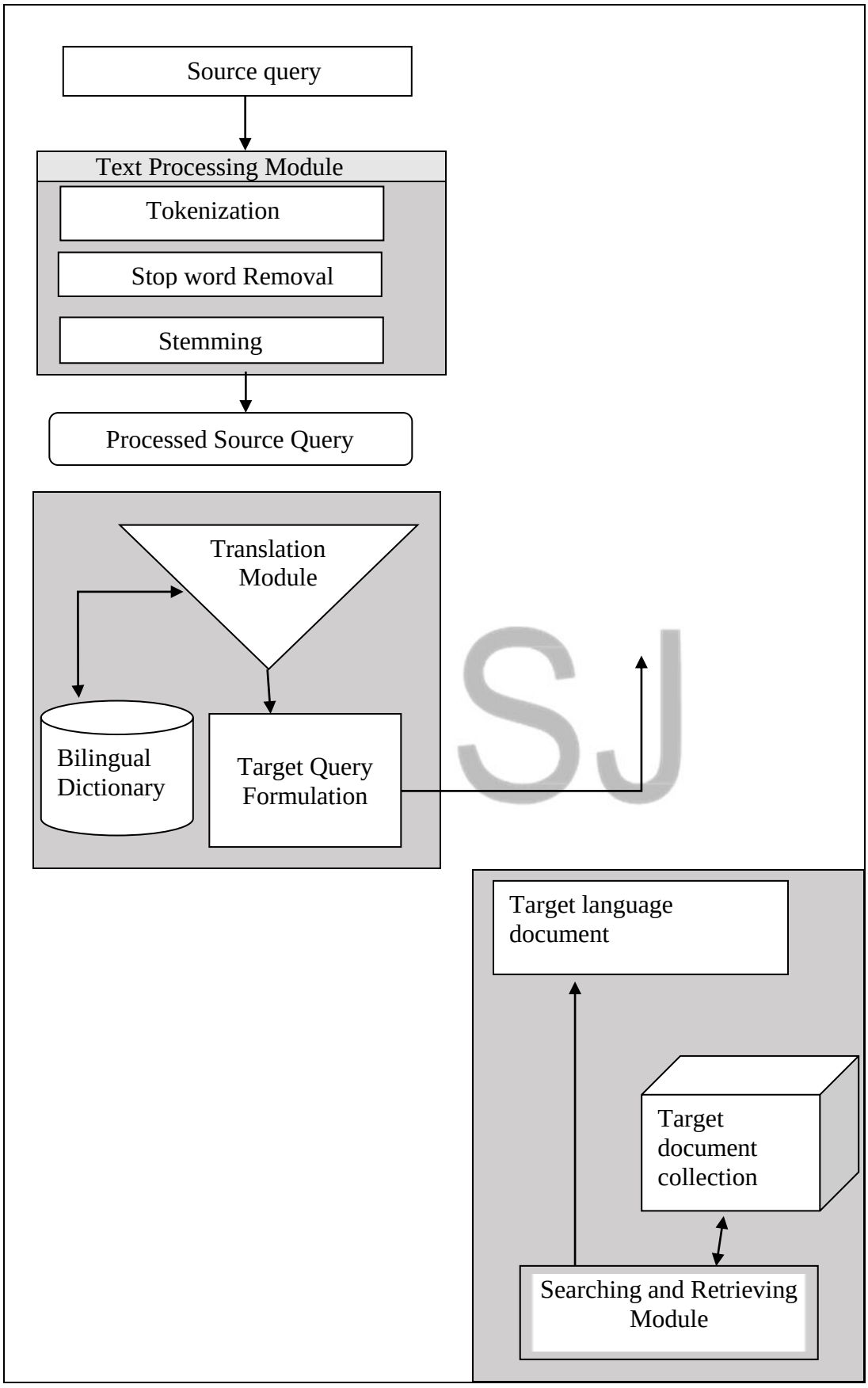

Fig: 1 Model of CLIR System (Sindhi to English)

3.1 Text Processing Module

This module aims at the pre-processing of query given by the user before the translation of the query keywords into the required language. Text processing modules include sub-modules of tokenization, stop words removal and stemming.

3.1.1 Tokenization

Tokenization is the process of dividing the query into chunks (pieces). These chunks are also called terms. Tokenization also eliminates punctuation marks (. , / ?). Tokenization of terms takes place by extracting the words/terms on the basis of word delimiters (spaces, tabs, punctuation marks). Tokenization of an example query can be seen in fig. 2.

Input Text

" كمپيوٽر ، اطلاع ۽ معلومات ذخير وكري تمام تيزيءَ سان جواب ڳولي سگهندڙ مشين آهي "

will be tokenized as:

[كمپيوٽر][اطلاع][۽][معلومات][ذخيرو][كري][تمام][تيزيءَ][سان][جواب][ڳولي][سگهندڙ][مشين] [آهي]

Fig 2: Tokenization of source query

3.1.2 Stop Word Removal

The process of eliminating frequent non-significant words (stop words) in a document or a request is normally done using called stop word list. Word lists have been used in information retrieval systems for the removal of high frequency words like prepositions, pronouns, articles, conjunction .In the Fig. 3 the stop word ۽, سان are thrown away from the source query in the stop word removal phase.

Input Text

" كمپيوٽر ، اطلاع ۽ معلومات ذخير وكري تمام تيزيءَ سان جواب ڳولي سگهندڙ مشين آهي "

Stop words removed as:

[كمپيوٽر][اطلاع] [معلومات][ذخيرو] [تمام][تيزيءَ] [جواب][ڳولي][سگهندڙ][مشين]

Fig 3: Stop word removal process

3.1.3 Stemming

Stemming is a task of removing suffixes from the word and return back a real/ root word. For example **سگهندڙ,سگهندي, سگهنديون , سگهندو,سگهون** are the inflected terms, and when expressed in source query, the stemmer removes its suffix and returns its stem word **سگھ .**

When text pre-processing of user given query is completed that query is now a collection of pre-processed source query terms as shown in fig. 4.

Input Text **"كمپيوٽر، اطلاع ۽ معلومات ذخير وكري تمام تيزيءَ سان جواب ڳولي سگهندڙ مشين**

will be stemmed as:

[كمپيوٽر][اطلاع] [معلومات][ذخيرو] [تمام][تيزَ] [جواب][ڳول][سگھ][مشين]

Fig 4: Stemming of the source query

3.1.4 Translation Module

This module accepts the processed source query terms as input and translate them into the targeted language (English) with the help of dictionary. The source language Sindhi query and the database is considered to be written in English. This module uses dictionary based translation method. A stem word dictionary has been developed, which make it possible to translate query words into source language (English). After text processing step, dictionary look up operation is performed for the each term of the source processed query (SPQ). SPQ terms are matched with words to identify the root words and other words. Remaining words that are not available in the dictionary are omitted as shown in fig. 5.

Input Text
"ڪمپيوٽر، اطلاع ۽ معلومات ذخير وکري تمام تيزيءَ سان جواب ڳولي سگھندڙ مشين "
will be translated through dictionary as:
[find][answer] [store][knowledge] [information][computer]

Fig 5: Translation process Sindhi to English

3.1.5 Retrieval Module

The purpose of this module is to search and obtain the related target documents in response to the user generated query that has been translated with the help of dictionary in targeted language. Now, target query is use to retrieve the documents from the document set or Internet by using a search engine as shown in fig. 6.

Now, this query will use for the retrieval purpose:
[find][answer] [store][knowledge] [information][computer]

Fig 6: Retrieval and searching documents

3.2 Problems and solutions

The problems that arise during translation process of Sindhi to English query are assessed by prescribed Sindhi-English CLIR System and complications are given with proper solution.

In the first technique, system takes query as an input, and split down the query in words. Splitted words are known as the tokens and this process is called tokenization. During this process certain characters such as punctuation marks are discarded. For this process, a tokenizer has been implemented in the system.

The second technique, is used to distinguish grammatical form of word when reduced to simple and root forms by sequential removal of word endings. This is called stemming and product is known as stem. For this purpose a stemmer has been implemented in the system.

In the third technique, the system does not manage phrase recognition and compound translation, because a compound word is formed when two words are joined to make a new word e.g. سنڌ يونيورسٽي(Sindh University), مٽي جو طوفان(dust storm). After examining compound words individually, if needed

Forth technique, i.e., Stop word list is created to frequently remove non-significant words in query. For the removal of high frequency words like articles, conjunction, article and preposition, the process is called “stop word removal”, which is utilized in Sindhi English Cross Language information to eliminate stop word in the source query.

4. Results and Experiments

As discussed in Chapter III, Sindhi to English CLIR System with query translation uses the dictionary based approach focusing on possible information seeking scenario. The test is set for evaluating the performance of a system. The performance result is evaluated in terms of two user effort measures, i.e., first 20 full precision and search length-i. Three information retrieval scenarios are considered, which include one cross lingual and two monolingual runs by formulating Sindhi and English queries in search engine (Google). Target retrieval contains Sindhi and English documents. Due to difficulties and complexities in Sindhi IR as discussed in chapter I, getting accurate results is highly unpredictable. The following are the results of different experiments, which include pre-processed monolingual English query, Sindhi query, and cross lingual Sindhi to English query.

4.1 Scenario 1: Monolingual Retrieval of Simple English Query

Four Computer Science students were selected. Each participants was asked to submit 5 queries to the search engine and rate them according to their relevance using five point scale (0-4). Where 0 indicates no keyword matching with the query words; 1 indicates irrelevant hit, a bad hit or duplicate link; 2 denotes somewhat relevant, like short mention of a topic with in a large page; 3 refers to less relevant or contains a link to a page of relevant information; and 4 for the most relevant document. In our study we considered Search Length-2. User also has to evaluate the search length, i.e., the number of links the user has to go through to find two highly relevant documents. A total of 20 queries were executed.

The graphical representation of first 20 full precision and Search lenght-2 of Monolingual English queries are given in Fig. 7. And Fig. 8 respectively.

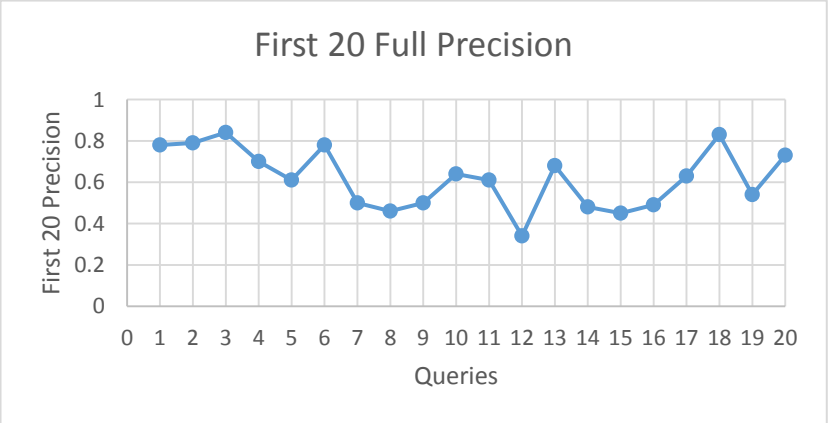

Fig 7: First 20 full precision of Monolingual English Queries

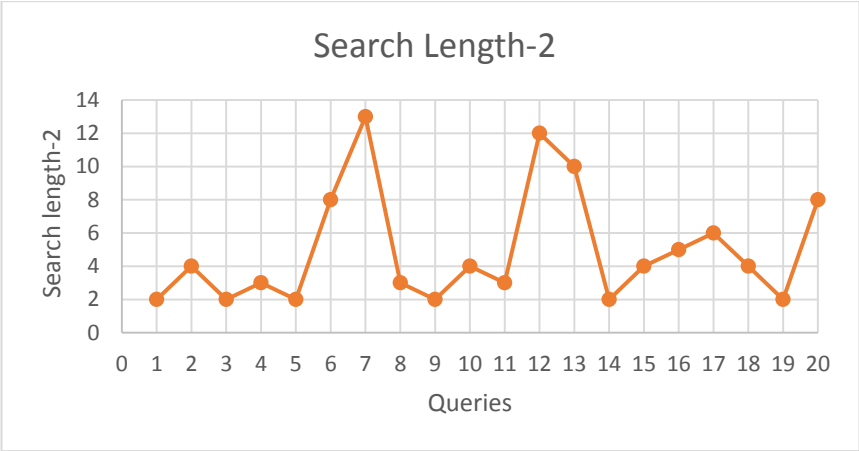

Fig 8: Search length -2 of Monolingual English Queries

Table 1 shows the results of first 20 full precision and search length 2 of twenty monolingual English queries.

Query number	First 20 full Precision	Search Length-2
1	0.78	2
2	0.79	4
3	0.84	2
4	0.70	3
5	0.61	2
6	0.78	8
7	0.50	13
8	0.46	3
9	0.50	2
10	0.64	4
11	0.61	3
12	0.34	12
13	0.68	10
14	0.48	2
15	0.45	4
16	0.49	5
17	0.63	6
18	0.83	4
19	0.54	2
20	0.73	8

Table 1: First 20 full precision and Search lenght-2 Monolingual English Queries

4.2 Scenario 2: Monolingual Retrieval of Simple Sindhi Query

As already described, each of the four participants submitted 5 queries to the search engine and rate them according to their relevance judgment with the help of relevancy scale (0-4). The result of this performance is comparatively far less than the English monolingual query retrieval where precision is higher than the Sindhi retrieval and search length is less. The returned documents are less relevant, because of the fact that search engines are not optimized for Sindhi search; therefore the search is based on exact word matching retrieval of Sindhi documents.

Table 2 shows result of first 20 full precision and search length-2 for this performance.

Query number	First 20 full Precision	Search Length-2
1	0.04	20
2	0	20
3	0.09	5
4	0.25	20
5	0.43	20
6	0.33	20
7	0.15	12
8	0.30	20
9	0.18	2
10	0.25	20
11	0.03	20
12	00	20
13	00	20
14	0.03	2
15	00	20
16	0.31	20
17	00	20
18	0.14	8
19	0.05	4
20	00	20

Table 2: First 20 full precision and Search lenght-2 of Monolingual Sindhi Queries

The graphical representation of first 20 full precision and search lenght-2 are shown in Fig. 9 and Fig. 10 respectively.

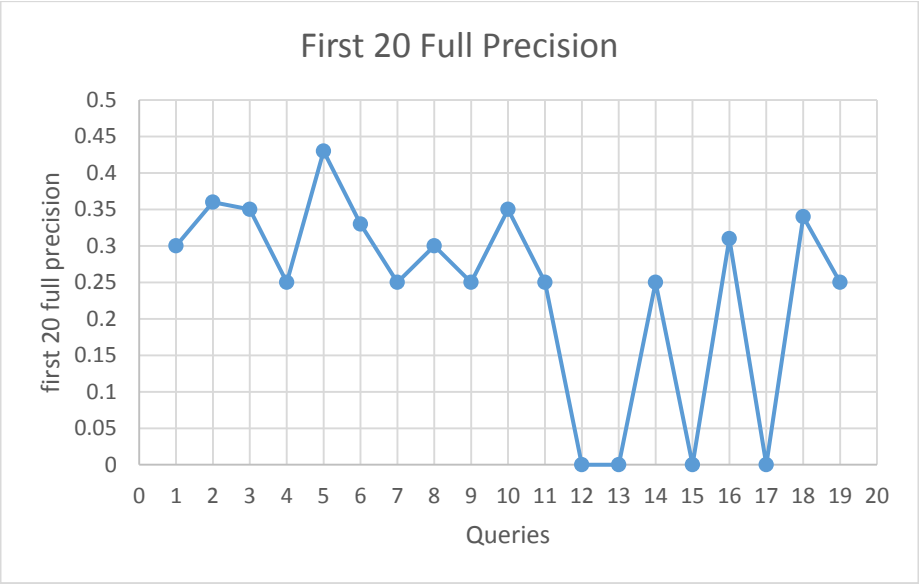

Fig 9: first 20 full precision of Monolingual Sindhi Search Queries

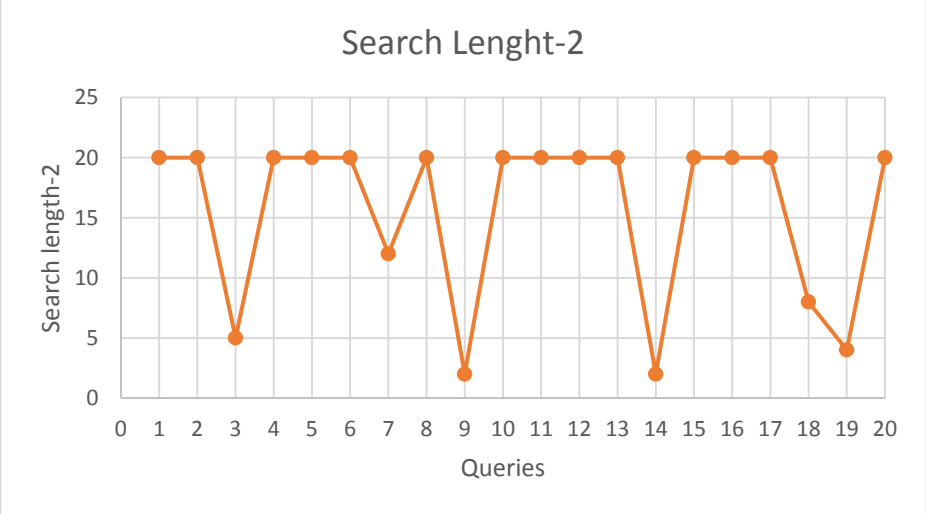

Fig 10: Search length-2 of Monolingual Sindhi Queries

4.3 Scenario 3: Cross language IR Sindhi to English Query

In this scenario, participants expressed same Sindhi query to the proposed system with already known situation, where relevance judgment based on five point scale (0-4) and search length-2. The results indicate that the given CLIR system retrieve the relevant document with high efficiency. The system processed Sindhi query translated into English, to retrieve the document in targeted language.

The results of first 20 full precision and search length2 of CLIR Sindhi to English queries are given in Table 3.

Query number	First 20 Precision	Search Length 2
1	0.68	5
2	0.79	6
3	0.80	2
4	0.65	5
5	0.55	5
6	0.68	4
7	0.58	12
8	0.61	4
9	0.59	4
10	0.48	7
11	0.43	16
12	0.33	3
13	0.58	5
14	0.48	2
15	0.33	18
16	0.39	11
17	0.68	2
18	0.50	6
19	O.53	7
20	0.68	8

Table 3: First 20 full precision and search length2 of CLIR Sindhi to English queries

The graphical representation of first 20 full precision and Search length-2 are given in Fig. 11 and Fig. 6 respectively.

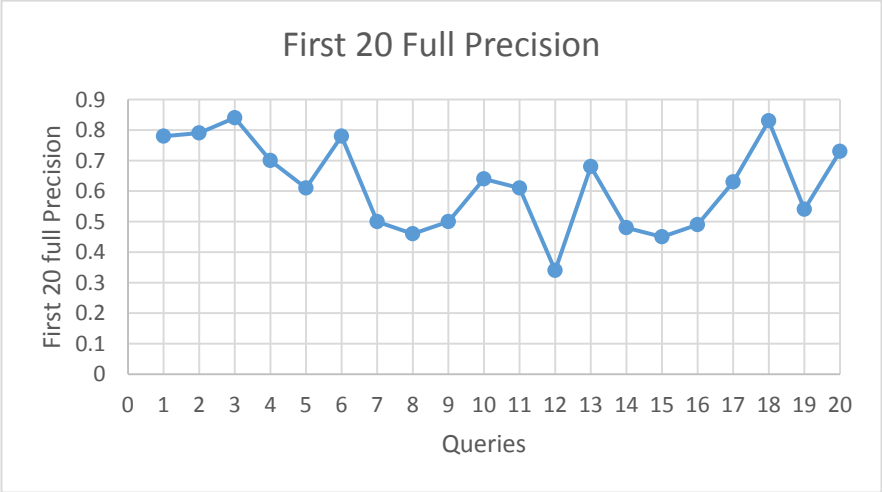

Fig 11: First 20 full precision of CLIR Sindhi to English Queries

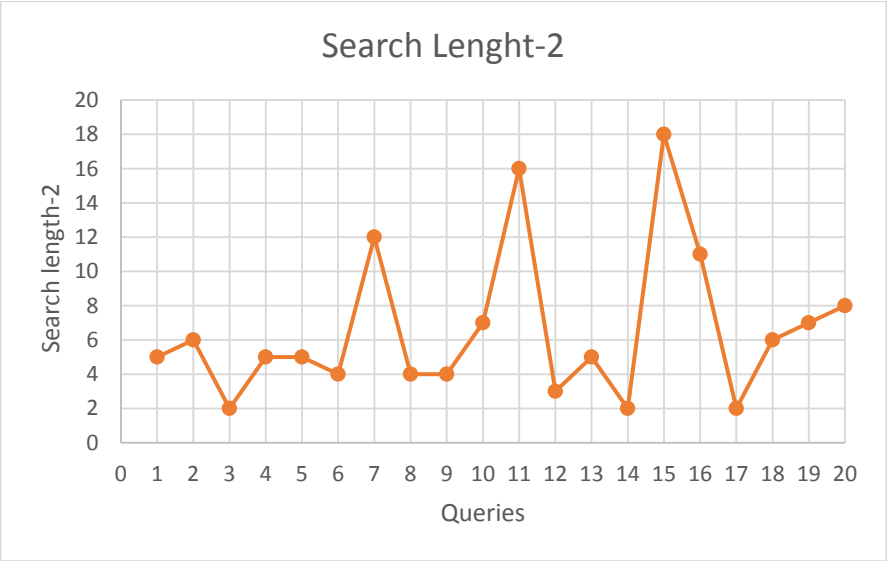

Fig 12: Search length-2 of CLIR Sindhi to English Queries

The visual representation of first 20 full precision of three different scenarios is shown in Fig. 13. It can be clearly seen that proposed model results fall either near to English or between Sindhi and English monolingual queries retrieval. It also concludes that proposed system shows the better results than Monolingual Sindhi query retrieval and lowers than or equivalent to the Monolingual English query retrieval.

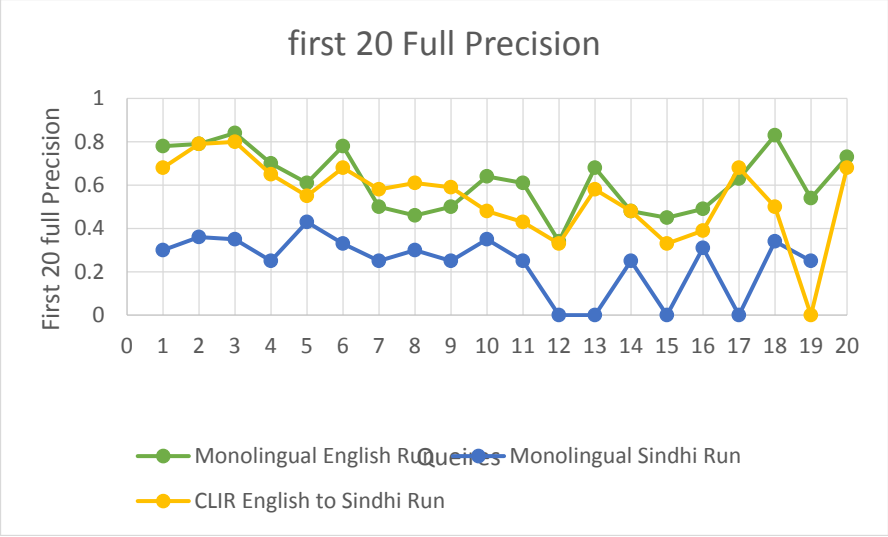

Fig 13: First 20 full precision of three scenarios

The fig. 14 summarizes a graphic representation of search length-2 for Google of all the 20 documents that were analysed. The data clearly indicate that for Sindhi to English CLIR retrieval user effort search length-2 is higher than the English search query retrieval search length-2, but lesser than monolingual Sindhi query retrieval .Which also shows improvement with proposed model.

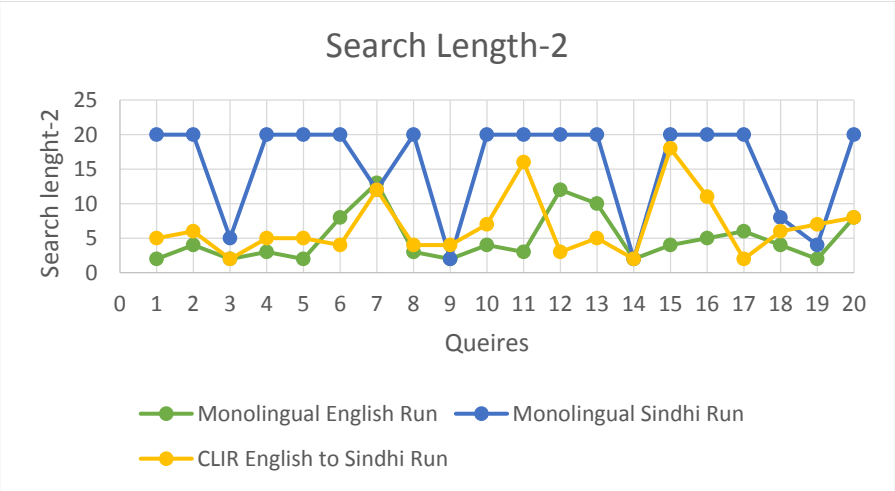

Fig 14: Search length-2 of three scenarios

The results of first 20 full precision and Search length-2 of all three scenarios for comparison is given in Table: 4 and Table: 5 respectively.

Query number	Monolingual English Run	Monolingual Sindhi Run	CLIR English to Sindhi Run
1	0.78	0.04	0.68
2	0.79	0	0.79
3	0.84	0.09	0.80
4	0.70	0.25	0.65
5	0.61	0.43	0.55
6	0.78	0.33	0.68
7	0.50	0.15	0.58
8	0.46	0.30	0.61
9	0.50	0.18	0.59
10	0.64	0.25	0.48
11	0.61	0.03	0.43
12	0.34	00	0.33
13	0.68	00	0.58
14	0.48	0.03	0.48
15	0.45	00	0.33
16	0.49	0.31	0.39
17	0.63	00	0.68
18	0.83	0.14	0.50
19	0.54	0.05	O.53
20	0.73	00	0.68

Table 4: First 20 full precision of all three scenarios

Query number	Monolingual English Run	Monolingual Sindhi Run	CLIR English to Sindhi Run
1	2	20	5
2	4	20	6
3	2	5	2
4	3	20	5
5	2	20	5
6	8	20	4
7	13	12	12
8	3	20	4
9	2	2	4
10	4	20	7
11	3	20	16
12	12	20	3
13	10	20	5
14	2	2	2
15	4	20	18
16	5	20	11
17	6	20	2
18	4	8	6
19	2	4	7
20	8	20	8

Table 5: Search length-2 of all three scenarios

DISCUSSION AND CONCLUSION

The Sindhi-English CLIR System plays an important role in IR/NLP based applications for Pakistani languages. The Sindhi –English CLIR System aims at resolving the issues encountered in the dictionary based query translation retrieval system. Review of literature revealed, there was very limited knowledge on information retrieval in Sindhi. In fact, there had been no any cross-language information retrieval research in Sindhi. We address one of the most primary issue in CLIR, i.e., the question of how to retrieve what the searcher means with what the document author meant. This naturally led us to the two solutions of either translating the query in document language i.e., translate the whole document into user's query language. Translating query is more convenient than translating whole document in user's query language. We choose query translation followed by tokenization, stop word removal, stemming, and translating Sindhi query keywords into English keywords, where only query translation knowledge is used. The English keywords are then usedto retrieve English documents in the database. Despite the big differences between the two language pairs, our experiments on Sindhi-English CLIR consistently confirmed these findings, showing that proposed cross-language tools and technique is not only effective, but also robust. The importance of query processing in dictionary is relies on the number of words available in the dictionary and resources available in source language. Process was particularly successful for the Sindhi-English Cross Language Information Retrieval, where the Sindhi words usually appear in inflected form, query translation was done with the help of Sindhi–English dictionary. The performance evaluation results show that developed model reduces the incorrectness of result. It is found that the proposed model successfully retrieve better results for the CLIR Sindhi to English than the existing monolingual search engines.

REFRENCES

HASAN, M.M. and MATSUMOTO, Y. *(*2000*)* Japanese-Chinese Cross-Language Information Retrieval. *Journal of Computational Linguistics and Chinese Language Processing,* Vol. 5, pp. 59-86*.*

PINGALI, P., TUNE, K.K. and VARMA, V. (2007) Hindi and Telugu to English Cross Language Information Retrieval. In: C. Peters, P. Clough, F.C. Gey, J. Karlgren, B. Magnini, D.W. Oard, M. D. Rijke, M. Stempfhube *(editors) Evaluation of Multilingual and Multi-modal Information Retrieval, Hyderabad, India* pp. 35-42.

SARASWATHI, S., SIDDHIQAA.A.M. AND KALAIYARASI.M. (2010) Bilingual Information Retrieval System for English and Tamil. *Journal of Computing*, Vol. 2, pp. 85-89.

SHERIDAN, P., SCHAUBLE, P. (1997) Cross-Language Multi-Media Information Retrieval. *Proceedings of the 3rd DELOS workshop; Cross-Language Information Retrieval*, ERCIM Workshop Proceedings No. 97-W003, (ISBN: 2-912335-02-7).
(Accessed: 2013, September 11)

WANG, J. and OARD, D.W. (2006) Combining bidirectional translation and synonymy for cross-language information retrieval. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, pp. 202-209.

ZHOU, D., TRURAN, M., BRAILSFORD, T. and ASHMAN, H. (2008) A Hybrid Technique for English-Chinese Cross Language Information Retrieval. *Journal of Asian Language Information Processing*, Vol. 7, pp. 1-35*.*