Using Monolingual Clickthrough Data to Build Cross-lingual Search Systems

Vamshi Ambati
Institute for Software Research International
Carnegie Mellon University
Pittsburgh, PA
vamshi@cmu.edu

Rohini U
Language Technologies Research Center
International Institute of Information Technology
Hyderabad, India
rohini@research.iiit.ac.in

ABSTRACT

A major portion of the World Wide Web(WWW) is still dominated by a few languages, with English being on the top. Monolingual information retrieval systems have been setup for such languages and are widely in use. To cater to a wider and diverse language speaking web users, we need Cross Lingual Information Retrieval (CLIR) systems that are capable of receiving a query in one language and returning results from a different language. To our knowledge not much work has been done in creating CLIR systems on the WWW. This is partly due to the unavailability of bilingual resources required for a major portion of the languages that are still a minority language in terms of the documents present on the WWW. Another important reason being the time and effort required to create a practical and useful CLIR system. In this paper, we address the problem of creating CLIR systems for language pairs in which the source language is a minority language and the target language is a majority language with existing search engines. We use clickthrough data from a monolingual search engine to learn translation models that could be used to perform cross lingual search. This approach has enabled us to generate practical CLIR systems on a large scale with less effort and with bilingual resources. We experiment and report the evaluation of our approach by creating CLIR systems for an Indian language and a few other European Languages.

1. INTRODUCTION

The task of Cross Lingual Information Retrieval (CLIR) addresses a situation when a query is posed in one language but the system is expected to return the documents written in another language. Once a user obtains a set of relevant documents in a foreign language, he can use automatic machine translation software to get a sense of the content. What remains is the problem of retrieving that set of documents, starting with a query in the user's native language. In recent years, the problem of Cross Lingual Information Retrieval has enjoyed significant interest from the research community, and a number of techniques were proposed to solve the problem [12],[15],[21],[5]. Most of these techniques center around a common idea of translating the query from the user's language to the language of the documents. In most cases, the translation is done in a word-by-word fashion using a dictionary, a machine translation system, or a similar resource. These are broadly called MT based approaches. Although MT based approaches have proven to be useful, some well known problems of these approaches are "missing dictionary translations" and "polysemy"[1]. Measures have been proposed to overcome these problems, such as the use of parallel corpora for query translation in CLIR [15] [8]

Another significant approach for CLIR that has been explored recently is the language modeling and statistical translation based approaches which treat the query translation and retrieval as an integrated process. Such approaches usually depend upon a probabilistic translation model either constructed from a dictionary by assigning uniform probabilities to the translations, from a parallel corpus or from a parallel corpus mined from texts on the web [15]. One potential advantage of such approaches is that they provide multiple translations for the same meaning. The translation of a query would then contain not only words that are true translations of the query, but also related words. Experiments have shown that these multiple translations have indeed helped in improving the performance of CLIR and in fact outperforming the traditional "MT followed by IR" approaches [12]. The performance of these systems however depends to a large extent on the effectiveness of the translation models.

Parallel corpus which has been a backbone of such translation models, is a difficult resource to acquire. Although projects like EuroParl [10] and others were successful in creating a large parallel corpus for European languages, many other languages like Indian and other less densely spoken languages do not enjoy such availability of parallel corpus and so these approaches are hindered from generating successful cross lingual information systems in these languages. Though a CLIR system could be built by collecting parallel corpus and other lexical resources for a given source-target language pair, performing the same to build CLIR systems for a large number of languages is a difficult task, if not impossible.

Most search engines today receive query requests that correspond to the information need of a large number of users. Our assumption is that, for a CLIR system the information need of the users using it is similar to that of the users using

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Workshop SIGIR '06 Seattle, USA
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.

any monolingual search system. In section 4 we also discuss the experiment conducted on a real world search engine query log to justify this assumption. This simple assumption, expose ways to exploit monolingual query logs which to our knowledge have not been exploited to a large extent.

In this paper, we identify query logs from a monolingual search engine as a huge resource that could be exploited to build cross lingual information retrieval systems at large for different language pairs. Given query logs from a monolingual search engine for language T, our approach builds rapid CLIR systems for any language pair (x, T), , provided there exists a bilingual dictionary from T to x and optional monolingual corpus in x.

We first create a cross lingual query log from the monolingual query log using a bilingual dictionary for translation and an optional monolingual corpus for resolving translation disambiguation. All the queries in the monolingual clickthrough data are translated to the source language intended for the cross lingual retrieval system in consideration. We then build a translation model from the resulting synthesized cross lingual query log and use it in the later retrieval model of the CLIR system. Our usage of monolingual query logs to create cross lingual probabilistic translation models enables us to generate CLIR systems with minimal language resources and effort. Our process of learning a translation model has the following advantages

- since the clickthrough data and the documents contained act as a good sample of the WWW, the translation model and the words in it are as close as possible to the distribution of the documents on the WWW.
- due to the volume of the existing clickthrough data, the translation models learnt have better coverage
- since queries are short without an associated context, disambiguation of information need is difficult. Usage of query logs has an implicit advantage of being able to disambiguate the word to the popular usage of the sense in terms of the documents clicked.
- since monolingual clickthrough data is readily available, existing search engines can set up with rapid CLIR systems for different languages with a minimal bilingual dictionary that are readily available over the web or which could be built from parallel texts [15]

The rest of the paper is structured as follows. Section 2 briefly surveys approaches in CLIR and recent developments in the field that motivate our approach. Section 3 describes in detail a query log based statistical translation model, that is used in our approach of creating CLIR systems. Section 4 describe our process of synthesizing a cross lingual query log from monolingual query logs in order to build the translation models. Section 5 discusses a CLIR system that uses the translation model. Section 6 discusses the experiments, evaluation and concludes with a discussion of the test data and results.

2. RELATED WORK

The area of Cross Lingual Information Retrieval has been well explored in the past few decades. The task was approached in two popular schools of thought. One was a translation of the query followed by a retrieval in monolingual domain, where as the second was translating the documents into the query language and performing retrieval[5]. Broadly, it can be said that the task has been seen as a translation followed by retrieval approach. For purposes of translation, existing bilingual dictionaries were used, parallel corpus was mined for extracting dictionaries which were then used in the translation. Methods have been proposed for disambiguation of words using statistics from a large corpus [21] [15].

Recently, a new approach to IR based on statistical language models has gained wide acceptance[17] [4] [6] [13]. These methods consider the information retrieval process as a generative process i.e. the documents generates the query and compute the relevance of a document for a given query by computing the probability that the document generates the query. These approaches have successfully been applied to CLIR setting [12] [11] [20], which treat query translation and retrieval as an integrated process. There are theoretical motivations for embedding translation into the retrieval model [11]. Xu et al [20] showed that combining statistics from various lexical resources can help correct problems of coverage and lead to significant improvements. Berger and Lafferty [4] view information retrieval as statistical translation which could readily be extended to perform cross lingual information retrieval. However they used a small data set consisting of documents and queries synthesized from a small set of documents. for learning translation models. In our approach we discuss the learning of these translation models from an abundantly and readily available data resource, clickthrough data of monolingual search engine.

While the prior approaches depend on dictionaries, parallel corpus or the web for creating these translation models, the lack of such resources for a number of languages on the WWW can be a drawback. In this paper we discuss how such effective translation models can be computed using a fairly untapped resource of query logs from monolingual search engine and a readily available resource, the bilingual dictionary for the languages.

3. QUERY LOG BASED PROBABILISTIC TRANSLATION MODEL

One of the essential parts of a CLIR system is query translation. In an MT based approach for CLIR , query translation it is an explicit process. In language modeling based approaches it is implicit and is guided by effective translation models. Learning the translation model is often done from either bi lingual dictionaries by assigning equal probabilities to all the translations of given word or from parallel texts[15]. The parallel texts are either manually made or generated using automatic Machine Translation systems or mining from the web[11]. Both parallel texts and MT systems are unavailable for many existing languages, which makes generation of a CLIR system a effortful endeavor.

In this section, we describe query log based probabilistic model which is learnt from a cross lingual query log, consisting of queries in the source language and their corresponding clicked documents. The model learnt from such a query log contains a word in source language, a word in target language and the probability of translation of the word in source language to the word in target language similar to the translation model learnt using parallel texts and bi

lingual corpora.

Our model is motivated by the translation model of Berger and Lafferty et al [4] which computes the translation model from queries and their respective relevant documents. Our process of learning a probabilistic model from a cross lingual query log is as follows. We first pick the cross lingual query log for a pair of languages. The query log consists of queries in source language and the urls of the documents in target language which are clicked for the respective query. We then align each of the query in the log with their respective click documents from the logs. This is done by considering the query as an entry in source language and the content in the click documents as an entry in target language. From these alignments, we used Statistical machine translation methods like IBM Model 1 motivated by Berger and Lafferty [4] etc to create the translation models. The resulting translation models consists of the source and target language words along with the probability of the translation of the source to the target word. Similar to Berger and Lafferty et al, we ignore the subtle aspects of language translation like word order etc.

Consider typical query log $Q_S D_T$ consists of queries and their clicked documents. Let $(Q,_s D_1, D_2...D_n)$ be a typical entry in the query log. consisting of the query Q_s and the clicked document set $D_1, D_2...D_n$ and $Q_s = q_1, ..., q_m$. Let $D_i = w_1, ...w_n$ A typical query log based translation model consists of the query word, the document word and the probability of their translation $(q_i \ w_j \ t(q_i|w_j))$. Our model captures the relation between the query words and document words in a probabilistic interpretation.

4. BUILDING CROSS LINGUAL QUERY LOGS

Although probabilistic translation models could be built from cross lingual clickthrough data, not many search engines exist today that perform CLIR on a large scale in real world. Hence, cross lingual query logs are rare and even if existing, are less in size. On the other hand, monolingual search engines that specialize in crawling and indexing of documents in a specific dominant language operate on a large scale generating large volumes of clickthrough data. In the following subsections we first motivate the construction of a cross lingual query log from a monolingual log and then discuss the creation of the same using a bilingual dictionary for translation and a corpus for translation disambiguation.

4.1 Motivation

Search engines usually store interactions of a user during a search session in the form of query logs or clickthrough data. Clickthrough data related to a user contains the query posed by the user along with the documents that he clicked among the number of search results returned for his query. The clicking of a document by the user is an indication of the relevance of the document to the query posed. Query logs have been used for various tasks in Information Retrieval research like personalization of retrieval results [18] among others. Although such intentions are planned for a CLIR[1], they can only be achieved after we have basic systems for as many languages possible. A sample of clickthrough data released by the search engine Alltheweb.com can be seen in table 1.

Query requests received by the search engines correspond to the information need of a large set of web users. Our assumption is that, for a CLIR system the information need of the users using it is similar to that of the users using any monolingual search system. This simple assumption, motivates us and exposes ways to exploit query logs from monolingual search engines as a valuable resource for the task of CLIR. To our knowledge such a resource has not been exploited earlier for building of CLIR systems.

ip address	time	Query	Url
4.16.103.153	14:47:19	mp3 to wav	mp3towave.com
4.16.103.153	15:00:23	cd to mp3	zy2000.com
4.16.103.153	15:06:29	cd to mp3	birdcagesoft.com
4.16.116.98	22:19:03	free pics	piczone.com
4.16.194.253	21:06:07	travel agent	travel.yahoo.com
4.16.194.253	21:06:37	travel agent	expedia.msn.com
4.16.195.144	21:49:01	voyeurweb	voyeurweb.com

Table 1: Sample clickthrough data from Alltheweb.com 2001

Lang	Total	Translated	Match with Eng	Overlap
Dutch	2994	703	515	0.732
Russian	3125	474	369	0.778
Italian	5131	1671	1284	0.768
French	69723	33419	27334	0.817
German	38311	4660	3079	0.660

Table 2: Queries overlap of Search Engine users of different languages with a monolingual English query log

In order to understand and arrive at this assumption, we conducted a sample experiment with query log data released by Alltheweb.com in the year 2002 [9]. The queries were posed by mostly European users in a particular month of the year. The data provided contains queries belonging to different European languages apart from English. The queries in the data also contained information of the language they belonged to. We picked the top 5 languages according to the number of queries posed in that language. They were French, German, Italian, Dutch and Russian. Using all these queries and also the monolingual English queries, which were quite large in number we tried to calculate the overlap in information need of the web users by comparing the similarity of queries belonging to these different languages with the English queries. We first translated the non-english language queries into English using the respective bilingual dictionaries. Queries related to proper nouns were not considered in this experiment. These translated queries are now matched with the actual English language queries from the query log. A partial match is also considered as a match. The number of overlaps are thus calculated for each language pair.

As can be seen from results displayed in table 2, there was an average of 75.1% overlap of queries in other languages with those of English. The remaining non-overlapping words could be attributed to the issues of coverage problems of the dictionary and polysemy. This experiment on such a limited query log data from a monolingual search engine proves of a possible information need overlap. In real world, the rate at which search engines like Google and MSN receive query requests and the volumes of clickthrough data formed, if a similar experiment is performed, we expect the results to only be even better.

Alltheweb.com being a search engine with a large share

of English documents, can primarily be treated as a monolingual search engine with English as its primary language. The number of overlaps indicates the overlap of information need of the users of a CLIR system with that of the users of monolingual search system. With such an overlap of information needs, as seen in the results of the experiment, we justify the construction of a cross lingual query log from a monolingual query log for purposes of CLIR. The one problem that now remains is the effective construction of a CLIR query log for an effective translation model.

4.2 Synthesizing cross lingual query logs

In this section we describe how a cross lingual query log could be synthesized from monolingual query logs with the help of a bilingual dictionary and an optional corresponding source language corpus. We first formalize our proposition and then discuss the creation of the cross lingual query log in detail.

Given a monolingual query log, $Q_T D_T$ with queries in language T and the documents in language T, we propose to create a cross lingual query log, $Q_S D_T$ with queries in language S and documents in language T. The lexical resources used are a dictionary from T to S and monolingual corpus in S.

Bilingual dictionaries are becoming a readily available resource over the WWW. With initiatives like the Universal Dictionary Project at Carnegie Mellon University, increasingly many dictionaries are available with reliable accuracy of translation. However, any dictionary based translation needs to address the problems of "missing dictionary entries" and "polysemy". If one can address such issues while creating a cross lingual query log across different languages, we can readily create an indispensable resource for CLIR.

Many words or phrases in one language can be translated into another language in a number of ways. For instance, the English word "free" can be translated into Hindi to an equivalent of free as in "freedom" and also free as in "free of cost". The choice of the translation depends on the context in which the word occurs. Therefore the translation of the word "free" in the queries "free pics" and "free bird" is ambiguous. Translation ambiguity is very common and needs to be addressed for better results in CLIR.

One way to address the translation disambiguation problems is to apply word sense disambiguation on the source language query and then use only those translation candidates that are associated with the appropriate sense. Unfortunately, word sense disambiguation is a non-trivial task and for most languages the appropriate resources, e.g., ontologies like WordNet [14], do not exist. Most important of all, the query posed in search engines is usually very small, with an average size of 2 words [9] and hence context of the query cannot be inferred for disambiguation.

Some approaches have been proposed to tackle the problem of query translation disambiguation [3] [8], but most of them assume existence of corpus and or a lexical resource that is tough to acquire for a lot of languages. We follow the approach of modeling context for the problem of translation selection using co-occurrences between translation terms. For instance, the simultaneous occurrence of the terms w1 and w2 count as a co-occurrence if they appear within a certain window, where a window can be a particular number of words, a sentence, a paragraph, or a document. Co-occurrences are more flexible than linear n-gram based approaches as they do not put any constraints on adjacency or word order. For example, given a query $Q = \{s1, s2, s3\}$, where the set of possible translations of s1 is $\{t_{1.1}, t_{1.2}, t_{1.3}\}$, s2 is $\{t_{2.1}, t_{2.2}\}$, and s3 is $\{t_{3.1}\}$, one compares all possible triples and selects the pair of terms that co-occur most frequently as the most likely translation. In this case, s3 has only one possible translation in the dictionary and so is not ambiguous. We calculate the co-occurrence statistics in calculating statistics for the translation of s1 and s2:

$freq(t_{1.1}, t_{2.1}, t_{3.1}) = n1$
$freq(t_{1.2}, t_{2.1}, t_{3.1}) = n2$
$freq(t_{1.3}, t_{2.1}, t_{3.1}) = n3$
$freq(t_{1.1}, t_{2.2}, t_{3.1}) = n4$
$freq(t_{1.2}, t_{2.2}, t_{3.1}) = n5$
$freq(t_{1.3}, t_{2.2}, t_{3.1}) = n6$

We also use a simple back-off technique and measure the co-occurrence frequencies of $n-1$ terms whenever the co-occurrences of n terms does not exist in the corpus.

Another problem as mentioned earlier is the issue with missing dictionary entries. This to a large extent depends upon the coverage that the bilingual dictionary provides. Although the coverage could be improved by application of learning algorithms to extract dictionaries from parallel corpus, we have not experimented such algorithms. The focus of this paper is to be able to produce a search system with minimal bilingual resources. We resort to make use of only a bilingual dictionary and have not assumed the existence of parallel corpus in the framework.

Proper nouns and domain specific terms are special cases of unknown words. Problems with their translation can not be treated as a problem with dictionary coverage. An effective approach proposed to deal with proper names is transliteration [19]. Domain specific lexicon acquisition has been proposed in [7] and phrasal translations have been explored [2] for improving CLIR systems. Although we have not experimented such methods in the CLIR system, the usage of such modules greatly boost the performance of the system. In this paper, we only focus on a methodology for generating practical CLIR systems from query logs of monolingual search systems. Approaches for phrasal translations [2] , transliteration [19] and other improvements in translation models, although not discussed, still can be applied in the framework to improve the generated CLIR systems.

5. A CLIR SYSTEM

5.1 Translation Model

We use a query log based translation model as described in Section 3. As mentioned earlier, we compute the probabilities similar to the statistical machine translation methods used by [4], We use GIZA++ [16] to compute the probabilities in the query log based translation model by aligning the queries and documents. GIZA++ is an extension of the program GIZA, which is part of the SMT toolkit EGYPT. GIZA++ was designed for word alignment of sentence aligned parallel corpora.

The queries are usually very short with an average of around 2 words [9] and the documents contain a large content in comparison to the query. This huge difference in the size of the document and the query can be a disadvantage to be trained with the statistical models for training like IBM model which depend a lot on such length variations. Also, documents typically contain certain noise words which we

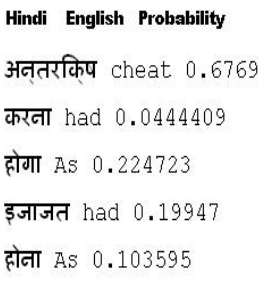

Figure 1: Sample from a Translation Model learnt from a synthesized Hindi-Eng query log

do not want to consider. Therefore while training our query based translation model using the query log, we do not consider the entire document, but a bag of words surrounding the query match in the document. We obtain the bag of words information not from the synthesized cross lingual query log, but from the monolingual query log. From each document we only pick 'k' words to the left and right of every query match in the document. Results reported here are using k=5 and an upper limit on the size of the bag of words as 25.

We further enhance the coverage of our translation model by augmenting it with a bi lingual dictionary based translation model. The bi lingual dictionary translation model is computed from a source language to target language bi lingual dictionary assuming equal probabilities for all the different translations of a given word in source language. Each entry in our enhanced translation model is a linear mixture of the qlog based translation model proposed in Section 3 and the bi lingual dictionary dictionary translation model.

Let $t_{dict}(q|w)$ be the probability of the word q given the word w obtained from the bi lingual dictionary based translation model. It is computed as $t_{dict}(q|w) = \frac{1}{n}$ where n is the number of different translations for the word q. Let $t(q|w)$ be the probability obtained from the qlog based translation model Then probability of q being translated as w obtained from the enhanced translation model $t_{enhanced}(q|w)$ is computed as

$$t_{enhanced}(q|w) = \beta\ t_{dict}(q|w) + (1-\beta)\ t(q|w)$$

A sample translation model we obtained is shown in Figure 1. We set the value of β to be 0.4 based on simple experimentation.

5.2 Retrieval Model

We use a probabilistic cross lingual retrieval model for performing retrieval. Motivated by earlier works[20][4][12] , we use a generative model to estimate the probability that a document in one language is relevant, given a query in another language. The probability that a given query is relevant to the a document is computed by probability that query is generated given document. This has also been used in cross lingual retrieval setting in several approaches [12] [20].

In the Cross lingual setting, consider a query Q_s in the source language and document D in target language. The probability that the document D is relevant to the query Q_s is computed as the probability that the document D in target language generates the query in source language as follows.

$$P(Q_s|D) = \Pi_{q_s in Q_s}[\alpha P(q_s|GSC) + (1-\alpha)\sum_{w\in Di} P(w|D)P(q_s|w)] \quad (1)$$

Similar to Xu et al, we fixed the parameter α to 0.3 in this study based on prior experience. We estimate the probability of a source language word in source language ($P(q_s|GSC)$, using a General Source Language corpus (GSC).

$$P(q_s|GSC) = freq(q_s, GSC)/|GSC|$$

where $freq(q_s, GSC)$ is the is the frequency of a word in source language in GSC and $|GSC|$ is the size of the General source language corpus. We use the part of the synthetic query log and a monolingual corpus to estimate this value in our experiments. $P(w|D)$ is the probability of observing the word w in the document D. It is computed as

$$P(w|D) = freq(w, D)/|D|$$

where $freq(w, D)$ is the frequency of word w in D and $|D|$ is the length of the document. $P(q_s|w)$ is the probability of translation of a query word in source language q_s given a word w in a document in target language. $P(q_s|w)$, depends on q_s and w only. This is obtained from the translation model described in the earlier section. Through out these calculations, we assume that the translation of a term is independent of the document and independent of the query in order to deal with data sparseness.

6. EXPERIMENTAL EVALUATION

6.1 Data and Experimental Setup

Query log data used in the experiments consist of the query, the clicked URLs for the query and the user identifier (ip addresses) and the time of click of the document. Such information though invaluable for research on information retrieval, is not released by major search engines. Recently, Alltheweb.com[1] has made available its search logs for research purposes. The data was collected from queries mainly submitted by European users on 6 February 2001. The data set contains approximately a million queries submitted by over 200,000 users and 977,891 unique click URLs. Further information on the data can be found in [9].

We use the query log data released by alltheweb.com to perform our experiments on learning the translation model and evaluating the cross lingual information retrieval system. We use the query, ip address and the click urls from the query logs for the purpose of experiments in this paper. We first create synthesized query logs from the alltheweb.com query logs as described in Section 4. We used a free hindi to english dictionary for synthesizing query logs. We then divided the data into two parts. One part containing 3/4 of the data is used for learning the translation models as described in Section 3. The remaining data is used for evaluating the retrieval effectiveness.

We first obtain all the documents corresponding to the queries in the testing data by crawling the click URLs and storing them as a repository. We were only successful in retrieving about 40% of the actual click URLs due to broken

[1]http://alltheweb.com

English Monolingual	0.392	
Hindi CLIR	Qlog based	Dictionary + Qlog
small Queries	0.183	0.223
long Queries	0.154	0.19

Table 3: Precision @ 10 calculation using method A for Hindi - Eng CLIR

English Monolingual	0.29	
Hindi CLIR	Qlog based	Dictionary + Qlog
small Queries	0.142	0.201
long Queries	0.114	0.14

Table 4: Precision @ 10 calculation using method B for Hindi - Eng CLIR

CLIR	Dictionary + Qlog
Dutch-Eng	0.172
Russian-Eng	0.121
Italian-Eng	0.14
French-Eng	0.18
German-Eng	0.19

Table 5: Precision @ 10 calculation using method B for Rapid CLIR systems built for 5 languages

links, network problems and restructuring of the WWW. These retrieved documents constitute the document repository used in current test experiments. With the volume of query log data we are working with, this repository could be considered as an analog to the WWW that corresponds to the query logs in discussion. For the purposes of these experiments, we name this repository as the mini-WWW, consisting of about 35,000 documents. We also pick queries from the query log data and pose it to Google to fetch and randomly pick and download a few of the top 100 documents. These documents are added to the mini-WWW. This prevents any kind of bias that may have been introduced in the construction of mini-WWW from click URLs in the query log data. With availability of every day query log data we expect the proposed approaches to scale and be useful in the WWW scenario.

6.2 Evaluation

Our approach discusses a novel practical method for generating quick CLIR systems using monolingual clickthrough data. To our knowledge no other approach has been proposed to use query log data in CLIR system creation and hence no standard data set or evaluation technique exists. Hence it is difficult to evaluate the effectiveness of a CLIR system generated using our approach and compare it with other existing methodologies. Hence we evaluate our approach by calculating the precision and recall of the retrieval in the system generated using the following two methods - a) human judgments of search results b) comparing the search results against the existing clickthrough data.

In method A, we have done a complete human judge evaluation. We first pick a set of queries in English to represent the sample information needs and display the same to the human. The judge (a speaker of Hindi) first translates the information need into the source language which in our case is Hindi. He then poses the same to the generated CLIR system which in turn returns results. He then provides his judgement of relevancy among the top 10 documents returned. This process is repeated for both the systems - a system that uses a translation model built just from query logs and a system with an enhanced translation model as discussed in our approach that is built using a dictionary and the query logs. Due to limited resources and as human effort is expensive, we only perform method A on one language for a set of 50 queries consisting of 30 short and 20 long queries. A small query is a single word query and a long query consists of more than 1 word in the query. Results are shown in Table 3. We also show the monolingual retrieval effectiveness to provide an estimate of the relevant documents in the data set considered for evaluation.

In method B, we automate the process of evaluation using the synthesized cross lingual query logs as described in Section 4. We first separate one fourth of the portion of query logs as the test data. We only train to learn the translation probabilistic model on the remaining 3/4th of the portion of the clickthrough data. This trained model is then evaluated on the test data. The process of evaluation is straight forward. Each query from the test data is posed to the CLIR system and the results are matched with the clicked documents portion in the clickthrough data. This kind of evaluation on query logs is the next best solution when human judgments is expensive to obtain. We used a set of 1000 queries from the test data for evaluation according to method B. Results of evaluation according to method B are shown in Table 4.

6.3 Building Rapid CLIR Systems

Using just the bilingual dictionaries, and no monolingual corpus for translation disambiguation, we have generated some CLIR systems and also evaluated the same according to method B as discussed in the earlier section. We have used freely available dictionaries on the WWW for the languages of Russian, Italian, Dutch, French and German. The CLIR systems were generated by synthesizing the corresponding cross lingual query logs from the monolingual corpus released by Alltheweb.com. Results from the automatic evaluation of these systems is shown in table 5. The English monolingual retrieval precision at 10 is the same as the above experiment "0.29".

6.4 Discussion of test data

The queries used in the evaluation of our approach were extracted from the test portion of the query log. According to [9], about 22.5% of the queries in the data released by Alltheweb.com contain names of places, people and things. Currently the systems generated by our approach can not handle proper names. This is because, although, approaches like transliteration have been proposed to deal with proper names[19], we haven't incorporated such techniques while synthesizing the cross lingual query logs. Also incorporating such modules are quite specific to the language and can not be generalized to all the languages. Although such modules could complement the system created by our approach, we have not included the same in our current experiments. Therefore queries used in the evaluation did not contain proper names. One other issue is the existence of a lot of domain specific words in the query like 'mp3', 'sat' etc. [7] proposes a system that extracts domain specific lexicons.

Appending such lexicons to the bilingual dictionaries used in the synthesis part of our approach, will greatly enhance the translation models learnt. However we currently did not include such queries with domain specific words in our evaluation.

The queries were primarily trying to exercise the disambiguation capability of the system, which has been the strength of the statistical translation model based approaches in CLIR. Some of the example queries that constituted the test set in evaluation were of the type "free bird", "free download","travel island" etc, which were useful in testing this capability.

7. CONCLUSION

In this paper we attempted to generate practical and useful CLIR systems with minimal effort and language resources. We discussed the usage of clickthrough data from a monolingual search engine for the learning of statistical translation models between the source language query and the target language documents. We first synthesized a cross lingual query log from the monolingual query log and used it in the process of learning translation models. A bilingual dictionary and optional monolingual corpus was used effectively in the creation of the cross lingual query log. We used the query log released by Alltheweb.com to build a CLIR system for Hindi English language pair and reported the results. We also create CLIR systems for 5 other European languages based on the same query log and respective bilingual dictionaries for the languages and report the results.

8. REFERENCES

[1] James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J. Harper, Djoerd Hiemstra, Thomas Hofmann, Eduard Hovy, Wessel Kraaij, John Lafferty, Victor Lavrenko, David Lewis, Liz Liddy, R. Manmatha, Andrew McCallum, Jay Ponte, John Prager, Dragomir Radev, Philip Resnik, Stephen Robertson, Roni Rosenfeld, Salim Roukos, Mark Sanderson, Rich Schwartz, Amit Singhal, Alan Smeaton, Howard Turtle, Ellen Voorhees, Ralph Weischedel, Jinxi Xu, and ChengXiang Zhai. Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst, september 2002. *SIGIR Forum*, 37(1):31–47, 2003.

[2] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91, 1997.

[3] Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71, New York, NY, USA, 1998. ACM Press.

[4] Adam Berger and John D. Lafferty. Information retrieval as statistical translation. In *Research and Development in Information Retrieval*, pages 222–229, 1999.

[5] Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. Translingual information retrieval: A comparative evaluation. In *IJCAI (1)*, pages 708–715, 1997.

[6] D. Hiemstra. *Using language models for information retrieval*. PhD thesis, University of Twente, 2001.

[7] D. Hiemstra, F. de Jong, and W. Kraaij. A domain specific lexicon acquisition tool for cross-language information retrieval, 1997.

[8] D. Hiemstra and Franciska de Jong. Disambiguation strategies for cross-language information retrieval. In *European Conference on Digital Libraries*, pages 274–293, 1999.

[9] Bernard J. Jansen and Amanda Spink. An analysis of web searching by european alltheweb.com users. In *Information Processing and Management*, volume 41, pages 361–381, 2005.

[10] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*, 2005.

[11] Wessel Kraaij, Jian-Yun Nie, and Michel Simard. Embedding web-based statistical translation models in cross-language information retrieval. *Comput. Linguist.*, 29(3):381–419, 2003.

[12] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. Cross-lingual relevance models. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 175–182, New York, NY, USA, 2002. ACM Press.

[13] D. Miller, T. Leek, and R. Schwartz. A hidden markov model information retrieval system. In *Proceedings on the 22nd annual international ACM SIGIR conference*, page 214221, 1999.

[14] G. Miller. Wordnet: an on-line lexical database. *International Journal of Lexicography*, 4(3), 1990.

[15] Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–81, New York, NY, USA, 1999. ACM Press.

[16] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[17] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.

[18] Rohini Uppuluri and Vamshi Ambati. Improving re-ranking of search results using clickthrough data from a search engine. In *Proceedings of AAAI 2006 Workshop on Intelligent Techniques for Web Personalization*, 2006.

[19] Paola Virga and Sanjeev Khudanpur. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition*, pages 57–64, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[20] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of SIGIR'01*, pages pages 105–110, 2001.

[21] Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, and Robert E. Frederking. Translingual information retrieval: Learning from bilingual corpora. *Artificial Intelligence*, 103(1-2):323–345, 1998.