Journal of Computer Science

Original Research Paper

English-Hindi Cross Language Information Retrieval System: Query Perspective

[1]Pratibha Bajpai, [2]Parul Verma and [3]Syed Q. Abbas

[1]*Department of Engineering and Technology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow, India*
[2]*Department of Information Technology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow, India*
[3]*Department of Computer Science, Ambalika Institute of Management and Technology, India*

Article history
Received: 31-01-2018
Revised: 22-03-2018
Accepted: 02-05-2018

Corresponding Author:
Pratibha Bajpai
Department of Engineering and Technology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow, India
Email: pratibhabajpai@gmail.com

Abstract: The abundance of multilingual content on internet other than English gives an urge to develop information retrieval system that can cross language boundaries. Such cross lingual information retrieval systems will bridge this language gap and allow user to ask a query in regional language and retrieve relevant documents in a different language. The problem of finding relevant document in language different from source language is the most challenging application of any cross lingual information retrieval. This paper discusses the development process of complete English to Hindi cross language information retrieval system along with the contribution of individual components to the system. The main focus of this paper is to discuss how optimization is done to our disambiguation approach, which we named as 'Two level Disambiguation method'. The experimental results obtained affirm that the addition of a component 'Analyzer' to our CLIR architecture increases the efficiency of our proposed disambiguation algorithm.

Keywords: Cross Lingual Information Retrieval System, Translation Analyzer, Disambiguation, Translation Ambiguity

Introduction

The English content on web has shrunk from 39 to 27% in last decade (Narasimha Raju and Bhadri Raju, 2015). On other side web content for languages like Chinese, Japanese, Hindi, Arabic etc. is showing gradual growth. The increasing number of users on internet who desire to access information expressed in languages other than their own has established cross lingual information retrieval as a major issue in information retrieval. The retrieval is bilingual if one source language (e.g., English) and one document language (e.g., Hindi) is used. The multilingual retrieval system accepts user query in one language while outputs documents in multiple languages. Sometimes an intermediate language is used as a means of translation, thereby making process transitive (Gollins and Sanderson, 2001).

The basic solution to CLIR is to translate the query into target language and consequently compute document scores using retrieval model like vector space or probabilistic model. This is one of the solutions. Other strategies can be: Direct matching of terms in different languages without translation, translating each document into query language or translating query and document into some common representation (Oard, 1998). Over the years query translation has evolved as the most well-liked strategy by researchers. But simple cross language query translation is less effective as compared to monolingual retrieval when typical measures like mean average precision and recall are used. Researchers suggest that by adopting simple linguistic techniques as translating phrases over individual words or limiting translation alternatives for query terms as provided by bilingual dictionary can raise the performance of CLIR to 75% of monolingual effectiveness (Oard and Diekema, 1998; Davis and Ogden, 1997; Hull and Grefenstette, 1996). In this study, we propose an effective method for limiting the size of translation candidates set for query words for optimization of our proposed query translation and disambiguation model.

The constitution recognizes Hindi and English as the only official languages of India (Chakrawarti and Bansal, 2017). In this study, we have tried to bring together a body of work that completely describes English to Hindi cross lingual information retrieval system. This

Science Publications

has many key folds. Our goal is not merely to describe the state of the art but to demonstrate the effect of the techniques involved in our framework on retrieval effectiveness for the two languages (English and Hindi) differing in their characteristics. While developing the process, the major concern has been on the following issues: (i) Restructuring of source query (ii) analyzing translation candidates and (iii) ambiguity removal.

Indian language internet user base has reached 234 million users at the end of 2016 surpassing the English internet users. This growth is likely to reach 536 million by 2021 compared to English internet user base. In particular Hindi internet user base is likely to outgrow English user base by 2021 (KGMP, 2017). This impressive growth in Indian language internet users motivates us to design and develop an English-Hindi Cross Language Information Retrieval (CLIR) System.

The paper commences by the related work in section 2 and contrastive analysis for the language pair English and Hindi in section 3. The contribution of the components of the processes is discussed in section 4. Section 4 also talks about the algorithm framed for short listing the translation candidates obtained for query terms from bilingual dictionary along with the demonstration through an example. Section 5 evaluates cross lingual retrieval system. Our results indicate that retrieval effectiveness is positively correlated with translation candidates set size and hence validate the utility of 'Analyzer' component in our CLIR framework and in increasing the effectiveness of our disambiguation algorithm. Finally we conclude in section 6.

Related Work

Query translation can be done by using any of the three resources namely Machine translation, Machine readable dictionaries or Parallel corpus. The dictionary translation is more preferred by researchers as this approach is simple and practical. But the method suffers from the problem of translation ambiguity as there is often one-to-many translation in bilingual dictionaries for source query words. To eradicate this problem researchers have tried measuring co-occurrence frequency of query terms. The method relies on the hypothesis that words appearing in the same document tend to share related senses and thereby represent a coherent content.

Croft and Ballesteros select the translation with the highest coherence score for Spanish-English language pair and reveal that the method is very successful for language pairs with scarce resources (Ballesteros and Croft, 1998).

Adrani approached the similar problem and used maximum similarity score between translation candidates for different query terms (Adriani, 2000). Later Gao *et al.* claimed that increase in distance between two terms weakens the association between them. They refined the disambiguation algorithm by incorporating decaying factor with the mutual information statistics. Liu *et al.* (2005) published an algorithm on maximum coherence model. They maximized the overall coherence of the query to estimate the translation probabilities of query terms using an iterative machine learning approach based on expectation maximization. Zhou *et al.* (2007) viewed the co-occurrence of possible translation terms within a given corpus as a graph and determines the importance of a translation using global information recursively drawn from the entire graph.

Giang *et al.* (2013) used mutual summary score based on word distribution in document collection to outperform basic model. Duque *et al.* (2015) Technique combines both the dictionary and co-occurrence graph to select the most suitable translation from the dictionary.

Contrastive Analysis of English and Hindi Language

Before we start discussing the proposed CLIR system for English-Hindi language pair, we need to see English from Hindi viewpoint, to make our system capable of performing contrastive analysis of the two languages. Both languages differ in morphological richness. Hindi is morphologically rich language whereas English has relatively simple morphology (Bhattacharyya, 2012).

Language topologists categorize English as an Subject-Verb-Object (SVO) and Hindi as Subject-Object-Verb (SOV) language. This classification is merely encoding of grammatical relations between Subject, Verb and Object between the two languages. In English a verb is preceded by the subject and followed by an object, while in Hindi the subject is followed by an object which is then followed by a verb. But in Hindi, the constituents of a sentence can be relatively moved freely around in the sentence without affecting the core meaning. E.g., the following sentence pair conveys the same meaning with different word order:

- **राम ने सीता को देखा** Ram ne Sita ko dekha
- **सीता को राम ने देखा** Sita ko Ram ne dekhaa

The identity of Ram as the subject and Sita as the object in both sentences comes from the case markers **ने** (ne – nominative) and **को** (ko –accusative) whereas, the two English sentences have exactly opposite meanings with similar change in the order of words.

Rats kill cats	Cats kill rats

This is because English does not have a morpheme for an accusative marker. The missing accusative marker is compensated by the subject position in English. This increases the structural differences between the two languages in the following way:

- In English, prepositions precede the words to which they relate. In Hindi, such words are called postpositions because they follow the words they govern

  On the table (English) **मेज पर** (mej par) (Hindi)

- Verb gets different meanings by using articles in English

  look at, look for, look after etc.

  whereas there are no articles in Hindi. Definiteness of a noun is indicated through pronoun, context or word order.
- The order between main verb and auxiliary verb is reversed

  खा रहा है (kha raha hai)(Hindi) is eating (English)

- English does not mark gender on the verb whereas Hindi does

  I go(English) while in Hindi **वो जाता है** **वो जाती है**

- Hindi gets advantage over English as it does not have a subject sharing rule

  रवि ने फल खरीदा और शाम तक खा भी गया
  (Ravi ne phal khareede aura sham taka kha bhii gaya).
  Here karma(phal) in first sentence is same as kartaa in the second sentence.
  While the English sentence

  Shyam dropped the melon and burst

  Is interpreted by native Hindi speaker as

  श्याम ने तरबूजा गिराया और तरबूजा फूटा (Shyam ne tarabuuja giraayaa aura tarabuuja phutaa)

  To give the above sentence correct interpretation, English constructs the sentence as

  Shyam dropped the melon and it burst
  श्याम ने तरबूजा गिराया और वो रो पड़ा
  (Shyam ne tarabuuja giraayaa aura wo ro padaa)

- In English, subject position can't be empty. This forces English to bear an extra overload of dummy 'it' and existential 'there'.

  It is Monday today (English) **आज सोमवार है** (Hindi) aaj somvar hai
  There are kites in the sky (English) **आसमान में पतंग है** (aasmaan me patang hai(Hindi)

  Again look at the following two sentences in English:

  E: Sita is eating mango
  E: Is Sita eating mango?

The above two sentences differ only in the word order. This change in word order makes first sentence as declarative while second as interrogative sentence. There is no explicit morpheme to mark the interrogativeness in English.
From the Hindi translations of these sentences:

H: **सीता आम खा रही है** (Sita aam khaa rahi hai)
H: **क्या सीता आम खा रही है**? (kyaa Sita aam khaa rahi hai?)

It is clear that Hindi has an explicit word 'kyaa' to mark the 'yesno' question while English codes this information in word order. The missing marker corresponding to yes-no question is compensated by the 'subject auxiliary verb inversion' in English. This weakens the proximity between main verb and auxiliary verb.
Consequences of Missing yes-no interrogative marker:

- Subject Position can't be empty as it indicates declarativeness or interrogativeness of the sentence.
- Insertion of auxiliary do in interrogatives:

  If a verb form does not involve an auxiliary verb, then a dummy 'do' is inserted, as shown below.

  She eats mango.
  Does she eats mango?

Thus, we conclude that English structurally differs from Hindi because of the absence of accusative marker and yes-no marker in English. To recompense for this shortcoming, English depends on its word order which in turn increases the differences between the language pair (Bhattacharyya, 2012; Bharati and Kulkarni, 2005; Bharati and Vineet, 2000).

English to Hindi Cross Lingual Information Retrieval System

Ideally, any CLIR system should retrieve all the relevant documents, ranked in decreasing order of relevancy for any user query. However, search results omit many relevant documents and often include many documents which are irrelevant. The primary reasons to this inconsistency can be attributed to few facts like morphological analysis of search keys, translation of search keys, selection of search keys translations and search key ambiguity.

Keeping in mind the grammatical complexities between the two languages and the primary reasons stated above, we have proposed the following CLIR system whose data flow has been shown in Fig. 1.

Figure 1 illustrates the data flow between the key components in our reference architecture. Before we initiate the preprocessing of query terms, the query needs to be tokenized. Here we are lucky enough as both English and Hindi languages are written with space-delimited words and thereby extracting terms from an English query or indexing terms from Hindi documents becomes too simple.

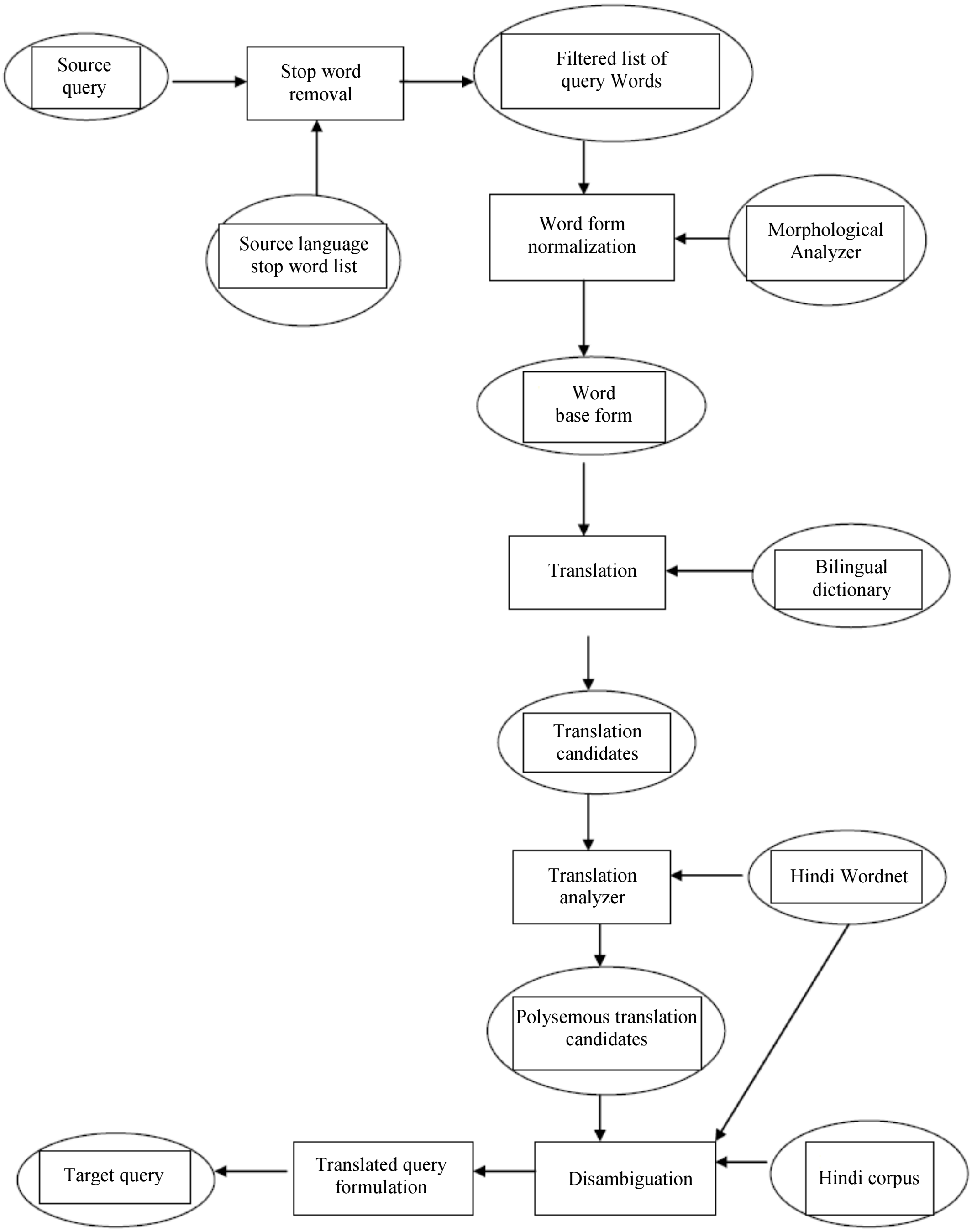

Fig. 1: Data flow in proposed CLIR System

The process is as follows:

Stop Word Removal

We use an English stop word list of 507 English words to remove stop words from the queries formulated for evaluation.

Word Form Normalization

Normalization is quiet simple for morphologically simple languages, such as English. Porter stemming algorithm is used to reduce inflected query words to base form in our system (Porter stemmer).

Translation

The most crucial step in performing Cross-Lingual Word Sense Disambiguation is the choice of a good bilingual dictionary (Andres *et al*., 2015). We use publicly available online bilingual English to Hindi dictionary Shabdanjali developed in IIIT, Hyderabad and containing 28K Hindi words to translate English queries to Hindi language queries (Shabdanjali English-Hindi Dictionary). The dictionary required conversion from ISCII to UTF-8 encoding and some basic normalization.

Analyzer

Dictionary translation leads to spurious equivalent translations in target language. All the translations are not desirable as many being synonyms of each other. The proposed model thereby concentrates only on the translation candidates of a query term having different meanings dropping the synonyms. Previous researches whereas treats all translation candidates equally and give undue advantage to query terms with more number of translations. We use Hindi WordNet, a lexical database for Hindi which is provided by the Linguistic Data Consortium and developed by IIT Bombay for filtering undesired translations (Pande *et al*., 2001). It contains 103438 unique Hindi words and 39271 number of synset.

To remove the synonyms, we suggest an easy algorithm as outlined below. This step in our CLIR system aims to optimize our proposed disambiguation model termed 'Two Level Disambiguation model'. It will also improve the relevancy of documents retrieved against the user queries.

Algorithm

```
Input: Source query $Q = \{q_1,q_2, \ldots\ldots,q_n\}$.

1. For each $q_i$ ($i$ = 1 to $n$), retrieve a set of translation
   candidates $S_i$ from bilingual dictionary.
2. For each translation candidate $h_j$ ($j$ = 1 to $|S_i|$), do
   steps 2.1 and 2.2
    2.1 Retrieve all synonyms from Hindi Wordnet.
        Call it set $P_k$.
    2.2 Remove sense $h_k(k = 1$ to $|S_i|$ and $k \neq j)$ from $S_i$
        if it occurs in set $P_k$.

Output: For each $q_i$, the set of senses $S_i$ contains
only those translation candidates which have different
sense.
```

To demonstrate the above algorithm, let us consider a query 'Renewable power'. From the bilingual dictionary 'Shabdanjali' we retrieve a list of translation candidates of 'power' as:

S_i = {**सामर्थ्य , हुकूमत , इख़्तियार, इख्तियार, ईश्वरत्व , कुदरत, बूता, अधिकार, अधिपति , विक्रम, ऊर्जा, विद्युत्, ओझा, क्षमता , ज़ोर, जोर, विद्युत् शक्ति , ताकत, अपरिमित परिमाण, ताक़त , आसुरी ताकत , अतिशक्तिशाली राष्ट्र, पराक्रम, प्रतिभा, प्रभाव, अगणित संख्या, प्राधिकार, राजकीय सत्ता, बल, बिजली, बिसात, राज, राज्य , वश, विभव, विभूति, शक्ति , शासन, सकता, सत्ता**}.

Next we retrieve the synsets of first term '**<u>सामर्थ्य</u>** ' from Hindi Wordnet:

P_k= { **औकात, औक़ात, बिसात, हैसियत, सामर्थ, सामर्थ्य शक्ति, निष्क्रय, इख्तियार, इख़्तियार , योग्यता, क़ाबिलियत, काबिलियत, क़ाबिलीयत, काबिलीयत, उपयुक्तता, लियाकत, लियाक़त, हुनर, सलीका, सलीक़ा, माद्दा, क्षमता, अर्हता, इल्मीयत, इस्तेदाद, ताक़त, ताकत, समर्थता, क्षमतापूर्णता, शक्तिपूर्णता**}

The common terms in the above two sets are {**इख्तियार, इख़्तियार, क्षमता**}

Removing common terms, set S_i = {**सामर्थ्य , हुकूमत, ईश्वरत्व , कुदरत, बूता, अधिकार, अधिपति , विक्रम ऊर्जा, विद्युत्, ओझा, ज़ोर, जोर, विद्युत् शक्ति , ताकत, अपरिमित परिमाण, ताक़त आसुरी ताकत , अतिशक्तिशाली राष्ट्र, पराक्रम प्रतिभा, प्रभाव, अगणित संख्या, प्राधिकार, राजकीय**

सत्ता, बल, बिजली, बिसात, राज, राज्य , वश, विभव, विभूति, शक्ति , शासन, सकता, सत्ता}.

Next we retrieve the synsets of second term '**हुकूभत**'from Hindi Wordnet

P_k = { **शासन, प्रशासन, शासन-प्रबंध, शासन-प्रबंधन, अनुशासन, सियासत, राज्य, राज, राजशाही, राज्यव्यवस्था, राज्य-व्यवस्था, राज्य व्यवस्था, अधिशासन, अभिशासन, अमल, अमीरी, एडमिनिस्ट्रेशन, एडमिनिस्ट्रैशन सत्ता, प्रभुत्व, स्वामित्व, आधिपत्य, अधिकार, शासनाधिकार, सत्ता, प्रभुत्व, स्वामित्व, प्रभुता, अधिकारिता, अधिकारित्व, प्रभुसत्ता, संप्रभुत्ता, संप्रभुत्व, मिल्कियत, मिलकियत, अमलदारी, इख्तियार, इख़्तियार**}

The common terms in the above two sets are {**शासन, राज्य, राज, अधिकार, सत्ता**}

Removing common terms, set S_i = {**सामर्थ्य , हुकूमत, ईश्वरत्व , कुदरत, बूता, अधिपति , विक्रम ऊर्जा, विद्युत्, ओझा, ज़ोर, जोर, विद्युत् शक्ति , ताकत, अपरिमित परिमाण, ताक़त, आसुरी ताकत , अतिशक्तिशाली राष्ट्र, पराक्रम प्रतिभा, प्रभाव, अगणित संख्या, प्राधिकार, बल, बिजली, बिसात, वश, विभव, विभूति, शक्ति , सकता**}.

There is no change in set S_i for the next two terms **ईश्वरत्व** and **कुदरत**. Next we retrieve the synsets of term '**बूता** 'from Hindi Wordnet.

P_k = { **शक्ति, बल, क्षमता, ताक़त, ताकत, दम, दमखम, कुव्वत, कूवत, बूता, हीर, दम-खम, दमख़म, दम-ख़म, दाप, ज़ोर, जोर, वृजन, वयोधा, वाज, अवदान, पावर, सत्व, सत्त्व, वीर्या, क्षत्र** }

The common terms in the above two sets are { **शक्ति, बल, ताक़त, ताकत, ज़ोर, जोर** }

Removing common terms, set S_i = {**सामर्थ्य , हुकूमत, ईश्वरत्व , कुदरत, बूता, अधिपति , विक्रम, ऊर्जा, विद्युत्, ओझा, विद्युत् शक्ति , अपरिमित परिमाण, आसुरी ताकत , अतिशक्तिशाली राष्ट्र, पराक्रम, प्रतिभा, प्रभाव, अगणित संख्या, प्राधिकार, बिजली, बिसात, , वश, विभव, विभूति, सकता**}

Similarly continuing for other terms in set S_i, the final set contains only the translation candidates which have different meanings:

S_i = {**सामर्थ्य , हुकूमत, ईश्वरत्व , कुदरत, बूता, अधिपति , विक्रम, ऊर्जा, विद्युत्, ओझा, विद्युत् शक्ति , अपरिमित परिमाण, अतिशक्तिशाली राष्ट्र, प्रतिभा, प्रभाव, अगणित संख्या, प्राधिकार, बिजली, विभव, विभूति, सकता**}

In this way the cardinality of the set *Si* is reduced from 40 to 21 and thereby leaving behind only the translation candidates having different meanings.

Disambiguation

Cross Lingual word sense disambiguation performs disambiguation of source language words while translating them to target language (Rekabsaz *et al*., 2017). We have proposed a disambiguation algorithm termed as 'Two level disambiguation model' which performs disambiguation at two levels. At first level we deal with the translation candidates in pairs only. This is done with the aim to obtain partial data for the likelihood of a translation in the perspective of other query terms. For a given query word, instead of taking binary decision for its translation alternatives, we measure the importance of each of the candidates in the context of given query. A translation candidate is assigned a high importance factor if it is rational with the semantic meaning of the user query. At second level we aim to find the most suitable translation for the given query. We compute the coherence between all possible combinations of translation candidates of query terms. This resolves the problem of translations being selected independently from selected and unselected translations of remaining query terms. Select the combination with highest score as the target language query.

Algorithm

Input: Source query $Q = \{q_1, q_2, \ldots\ldots, q_n\}$.

```
1. For each $q_i (i = 1 \text{ to } n)$, retrieve a set of translation candidates Si from bilingual dictionary.
2. For each translation candidate $h_j (j = 1 \text{ to } |S_i|)$, do steps 2.1 and 2.2
   2.1 Retrieve all synonyms from Hindi Wordnet. Call it set $P_k$.
   2.2 Remove sense $h_k (k = 1 \text{ to } |S_i| \text{ and } k \neq j)$ from $S_i$ if it occurs in set $Pk$.
3. For each $q_i (i = 1 \text{ to } n)$, do step 3.1
   3.1 For each $h_j (j = 1 \text{ to } |S_i|)$, do steps 3.1.1 to 3.1.5
       3.1.1 Retrieve all example sentences for its synset, hypernyms and homonyms from Hindi WordNet.
       3.1.2 Count the usage of a translation candidate $h_j$ in example sentences of translation candidates $t_p$ of other query terms $q_k$, where $1 <= k <= n, k \neq i$ and $p = 1$ to $|S_k|$.
       3.1.3 Find the sum of usage of $h_j$ to obtain $UC_i$, the Usage Count of a particular translation candidate with respect to translation candidates of other query terms.
       3.1.4 Normalize $UC_i$ to obtain $IF_i$, Importance Factor of translation candidate $h_j$.
4. For $i = 1$ to $n$, do step 4.1
   4.1 For every combinations $C = \{h_1, h_2, \ldots\ldots, h_n\}$ where $h_i$ is a translation candidate of $q_i$, do step 4.1.1
       4.1.1 For $j = 1$ to n and $i \neq j$
       4.1.1.1 Compute WSDC as:
```

$$WSDC(C) = \sum_{h_i, h_j \in C} \left(DC(h_i, h_j) * IF(h_i) * IF(h_j) \right), where$$

$$DiceCoefficient, DC(h_i, h_j) = \frac{2 * freq(h_i, h_j)}{freq(h_i) + freq(h_j)}$$

$freq(h_i)$	= The number of occurrences of term h_i in training corpus
$freq(h_j)$	= The number of occurrences of term h_j in training corpus
$freq(h_i,h_j)$	= Co-occurrence frequency of terms h_i and h_j in a sentence in documents.

```
5. Select the combination with highest WSDC score as the target language query $Q^t$ of the source query $Q$:
```

$$Q^t = \arg\max_C WSDC(C)$$

Output: Disambiguated Hindi query Q^t for English query Q.

Experiment

In this section we will discuss how the addition of the component 'Analyzer' to our CLIR architecture increases the efficiency of our proposed disambiguation algorithm.

Evaluation Environment

An evaluation environment consists of a set of 50 topics which are designed as web user queries; and web documents which are searched to find documents relevant to the topics. The web documents are fetched from Google (http://www.google.com/) and Bing (http://www.bing.com/) indexed database. The relevance judgments for the Hindi documents obtained with respect to English queries is established with the help of three Hindi speaking volunteers from Indian Institute of Technology (BHU). Document which is judged as relevant by all the three volunteers is marked as relevant else treated as irrelevant. Evaluation is done by computing Mean Average Precision (MAP) for first 50 documents retrieved on two different search engines Google and Bing. For our Cross-Language Information Retrieval evaluation, we also measure how well the cross-language IR performs with respect to monolingual information retrieval on the same set of web documents.

Result Analysis

The following methods are compared to investigate the effectiveness of our model for query translation and disambiguation:

- *Monolingual*: Retrieval using the Hindi queries translated manually by Hindi language expert. Monolingual run provides unreachable performance ceiling for any cross lingual information system as translation process is inherently noisy
- *Proposed model*: Retrieval using the proposed two level disambiguation model
- *Proposed model with analyzer*: Retrieval using two level disambiguation model using polysemous translation candidates only

Table 1 describes our experimental results. For each method, we give average values of P@k with k= 10, 20 and 50 using Google search engine.

Table 2 compares the MAP value of two level disambiguation method with analyzer with baseline method i.e., monolingual run and proposed disambiguation method for English queries. The performance of disambiguation method is 79.53% while using analyzer it increases to 87.45% of monolingual run.

Table 3 gives average values of P@k with k = 10, 20 and 50 with Bing search engine.

Table 1: Run statistics for English queries with Google search engine

Experimental run	P@10	P@20	P@50
Monolingual	0.483	0.420	0.309
Two level disambiguation	0.383	0.336	0.240
Two level disambiguation with analyzer	0.421	0.387	0.272

Table 2: Mean average precision of experimental runs for queries with Google

Experimental run	Mean Average Precision (MAP)	Percentage monolingual
Monolingual	0.518	--
Two level disambiguation	0.412	79.53%
Two level disambiguation with analyzer	0.453	87.45%

Table 3: Run statistics for English queries with Bing search engine

Experimental run	P@10	P@20	P@50
Monolingual	0.412	0.358	0.263
Two level disambiguation	0.310	0.271	0.199
Two level disambiguation with analyzer	0.334	0.291	0.162

Table 4: Average retrieval precision of experimental runs for queries with Bing

Experimental run	Mean Average Precision (MAP)	Percentage monolingual
Monolingual	0.441	--
Two level disambiguation	0.333	75.5%
Two level disambiguation with analyzer	0.358	81.1%

Table 4 compares the MAP value of two level disambiguation method with analyzer with baseline method i.e., monolingual run and proposed disambiguation method for English queries using Bing search engine. The performance of disambiguation method is 75.5% while using analyzer it increases to 81.1% of monolingual run.

We have used same set of English test queries (designed on the lines of TREC and CLEF guidelines) and Hindi document collection, which is used to evaluate our disambiguation algorithm. Here we have evaluated our algorithm on Bing search engine along with Google to check whether the proposed algorithm is favored by a particular search engine. The MAP of two level disambiguation algorithm which is more than 75% of monolingual search with both search engines proves the effectiveness of our algorithm and no favourism of search engine.

Adding the component analyzer to our disambiguation algorithm increases the effectiveness of the disambiguation algorithm. All the synonyms obtained from bilingual dictionary during translation phase for a query word are removed keeping behind a single word before the query is disambiguated (same has been explained by an example above). After this process only polysemous translations of query words are left. These synonyms are replaced by the same word in Hindi documents too. This will increase the co-occurrence of correct translations in Hindi documents, thereby increasing the probability of correct translation to be selected as final translation of English query word. This in turn increases the number of relevant Hindi documents retrieved on both search engines for given English query. This is in accordance with the test result shown in Table 2 and 4.

Conclusion

In earlier works using machine readable dictionaries, user queries were formed including all translations for all query terms. Due to this some retrieval methods which treat term contribution as independent can give undue advantage to query terms having more number of translations. This is in general an objectionable trait for any retrieval system.

In this study we have tried to optimize our proposed query translation and disambiguation model by addition of a new valuable component Analyzer in the basic Cross Language Information Retrieval (CLIR) system. Our effort has been able to resolve the objectionable trait of any retrieval system and provides precise and quality target language translations. Hence we have been able to propose an inexpensive and easy to be implemented CLIR system.

Acknowledgment

We would like to thank the department of Computer Science, Indian Institute of Technology (BHU) who provided insight and expertise that greatly assisted the research.

Author’s Contributions

Pratibha Bajpai: Development, experimentation, validation of algorithms and writing the manuscript for the journal.

Parul Verma: Concept development and proof reading.

Syed Q. Abbas: Concept development and proof reading.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that there are no ethical issues involved.

References

Adriani, M., 2000. Using statistical term similarity for sense disambiguation in cross-language information Retrieval. Inf. Retr., 2: 71-82. DOI: 10.1023/A:1009989801965

Andres, D.F., M.R. Juan and A. Lourdes, 2015. Choosing the best dictionary for Cross-Lingual Word Sense Disambiguation.

Ballesteros, L. and W.B. Croft, 1998. Resolving ambiguity for cross-language retrieval. Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 24-28, ACM, Melbourne, Australia, pp:64-71. DOI: 10.1145/290941.290958

Bharati, A. and A. Kulkarni, 2005. English from hindi viewpoint: A paaninian perspective. Linguistic Society of India, held at CALTS, University of Hyderabad, Hyderabad.

Bharati, A. and C. Vineet, 2000. Dipti misra sharma, amba kulkarni modern technology for language access: An aid to read English in Indian context. Osmania Papers Ling., 26-27: 111-126.

Bhattacharyya, P., 2012. Natural language processing: A perspective from computation in presence of ambiguity, resource constraint and multilinguality. CSI J. Comput.**,** 1: 3-13.

Chakrawarti, R.K. and P. Bansal, 2017. Approaches for improving Hindi to English machine translation system. Indian J. Sci. Technol., 10: 1-8. DOI: 10.17485/ijst/2017/v10i16/111895

Davis, M.W. and W.C. Ogden, 1997. Implementing a large-scale cross-language text retrieval system. Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 27-31, ACM, Philadelphia, Pennsylvania, pp: 92-98. DOI: 10.1145/258525.258542

Duque, A., L. Araujo and M.R. Juan, 2015. CO-graph: A new graph-based technique for cross-lingual word sense disambiguation. Natural Lang. Eng., 21: 743-772. DOI: 10.1017/S1351324915000091

Giang, L., V.T. Hung and H.C. Phap, 2013. Experiments with query translation and re-ranking methods in Vietnamese-English Bilingual Information Retrieval.

Gollins, T. and M. Sanderson, 2001. Improving cross language retrieval with triangulated translation. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development In Information Retrieval, Sept. 9-12, New Orleans, pp: 90-95. DOI: 10.1145/383952.383965

http://www.bing.com/

http://www.google.com/

Hull, D.A. and G.Q. Grefenstette, 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 18-22, ACM Zurich, Switzerland, pp: 49-57. DOI: 10.1145/243199.243212

KGMP, 2017. Indian languages- defining India's Internet, a study by KGMP in India and Google.

Liu, Y., R. Jin and J.Y. Chai, 2005. A maximum coherence model for dictionary-based cross-language information retrieval. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, ACM Salvador, Brazil, pp: 536-543. DOI: 10.1145/1076034.1076125

Narasimha Raju, B.N.V. and M.S.V.S. Bhadri Raju, 2015. Dictionary based translation approaches in cross language information retrieval: State of the Art. Int. J. Scient. Eng. Res., 6: 324-330.

Oard, D.A., 1998. A comparative study of query and document translation for cross-language information retrieval. Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup, Oct. 28-31, Springer-Verlag London, pp: 472-483.

Oard, D.W. and A.R. Diekema, 1998. Cross-language information retrieval. Annual Rev. Inform. Sci. Technol., 33: 223-256.

Pande, P., P. Bhattacharya, S. Jha and D.A. Narayan, 2001. Wordnet for hindi.

Porter stemmer. https://www.drupal.org/project/porterstemmer

Rekabsaz, N., M. Lupu, A. Hanbury and A. Duque, 2017. Addressing cross-lingual word sense disambiguation on low-density languages: Application to Persian. Comput. Sci. Comput. Lang.

Shab danjali English-Hindi Dictionary from IIIT Hyderabad

Zhou, D., T. Mark and B. Tim, 2007. Disambiguation and unknown term translation in cross language information retrieval. Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum for European Languages, Sept. 19-21, Springer Berlin, Heidelberg, pp: 64-71. DOI: 10.1007/978-3-540-85760-0_8b