Chapter 3
Literature Review

3.1 Introduction

This study fits into the larger theoretical context of research in information retrieval (IR) both multilingual and monolingual. This literature review points to the need to acknowledge the importance of information seeking processes. This chapter reviews the literature from various sources including books, articles, journals, magazines etc. Focus of the review is to study various IR, mining, NLP tools used worldwide for information access, developments in the area of cross and monolingual IR particularly in India. Key lessons learned from the review are summarized in the end of the chapter.

3.2 Existing Smart Web Tools

There are a number of tools used by common man, researchers and academicians etc for searching information in any of their native language/s. Some of the tools provide more relevant information where as some do not. Some of the search tools do deal with multiple languages but the efforts required by the users for getting desired information are much more. A brief description of the most popular search tools is given in the proceeding section.

3.2.1 Search Engines

Search engines [GOOGLE, MSN, YAHOO, BING] focus on indexing the WWW providing a method of finding web documents based on a given query. A common problem with search engines is that they frequently return documents that are irrelevant to what the user was expecting. This can be due to two reasons. Firstly, the keywords used to form the query are a poor choice with regard to expected document set returned. A lack of experience is the primary cause for this
weakness, where users are unfamiliar with the interface and the options available when forming an accurate query. Secondly, the search engine itself provides a poor set of indexes for the collection of document. Even the perfect query can be limited by the accuracy of the set of indexes it determined and currently holds.

### 3.2.2 WebWatcher

WebWatcher [Arm97] is a goal-oriented browsing assistant that makes link suggestions by highlighting links that may lead to the pre-defined goal. WebWatcher requests an initial explicit goal from the user, and the e-mail address to keep track of the user’s interests. The agent tracks user behavior and constructs new training examples from the encountered hyperlinks. Each hyperlink is evaluated using a utility function based on the Page, the Goal, the User and the Link. The function assigns a value from 0 to 1 to the links, suggesting the best ones. The system uses TF-IDF vectors, cosine similarity, statistical prediction and Boolean functions.

### 3.2.3 Letizia

Letizia [Lie95, Lie97] is an autonomous interface agent that assists a user browsing the WWW. Letizia doesn’t require the user to provide an explicit goal like WebWatcher. Instead, it automates a browsing strategy consisting of a breadth-first search, within the boundary of linked pages, augmented by a set of content and context based heuristics to infer the goal. For example, when a user saves a bookmark for a document, this is interpreted as interest in the content of the document. Similarly, if a document link is not clicked, this is interpreted as disinterest in the subject. Upon request, Letizia will make recommendations on which links on the current page the user should follow based on the interest profile inferred by the agent. The recommended links for the user to follow are presented in a preference ordered list. There is little detail given by Lieberman as to how Letizia represents a user profile and the exact learning techniques employed. What is apparent is that Letizia treats documents as a set of
keywords/topics. Letizia is implemented in Macintosh Common Lisp with an adapted Netscape web browser.

3.2.4 PowerScout

PowerScout [Lie01] is another tool from Lieberman adopting a similar user interest profile as Letizia. PowerScout’s approach to recommendations differs from Letizia’s forward scouting by using search engines to determine new pages of interest that match a given user’s interests. Like Letizia, PowerScout is implemented with an adapted Netscape web browser.

3.2.5 WebMate

WebMate [Che98] is an agent that helps users to effectively browse and search the web by monitoring user’s browsing behavior based on a user selecting a page as “I like”. The profile constructed consists of a collection of TF-IDF document vectors under the Vector Space Model [Sal83] using cosine similarity function to determine vector similarity. WebMate compiles a personal newspaper based on the user’s profile by either querying popular search engines or by automatically trawling a list of user-defined URLs.

3.2.6 Lira

Lira [Bal95] is an agent that autonomously searches the WWW for interesting web pages. Documents are represented as vectors under the Vector Space Model [Sal83]. Lira incorporates stemming [Por80] so that the terms in the document vectors are represented by their stems only and not whole words. Weighting of the document terms are performed using a sophisticated TF-IDF scheme which normalizes for a given document length:

3.2.7 Almathaea

Almathaea [Mou97] is a co-evolution model of information filtering agents that adapt to the various users’ interest and information discovery agents that monitor and adapt to the various on-line information sources. Moukas states
that users should rely on search engines and use the result of their queries as starting points for their agents. Therefore, Almathaea dispatches several of these filtering agents to query search engines and filter responses using an adapted TF-IDF measure.

### 3.2.8 Footprints

Footprints is a prototype system created to help people browse complex web sites by visualizing the paths taken by users who have been to the site before [Wex97]. The system comprises of two main pieces: a front end and a back end. The back end runs in a batch mode at a given time (say once per night), processing web logs of a single web site. The front end reads the data file output by the backend and is used to create a customized site map for a user as they enter the new web site. Footprints presents the results of statistical analysis performed on data representing the history of interaction between users and a given web site. It provides a variety of tools such as maps, path views, annotations comments or signposts by other users. These paths are shown as a graph of linked document nodes, with the links colour-coded to visualize the frequency of use of the different paths. [2]

### 3.3 Web Data Extraction Tools

- Text Mining Tools and NLP Tools
- IR and CLIR Tools

#### 3.3.1 Text Mining and NLP TOOLS

A brief description of various text mining and NLP tools is presented in the proceeding paragraph.
3.3.1.1 NetOwl Extractor

http://www.textmining.com/

Initially developed for the most demanding government intelligence applications, NetOwl Extractor is based on advanced computational linguistic and natural language processing technology. By intelligently analyzing structure and context within text, NetOwl accurately identifies key information. NetOwl Extractor is an automatic indexing system that finds and classifies key phrases in text, such as personal names, corporate names, place names, dates, and monetary expressions. NetOwl Extractor finds all mentions of a name and links names that refer to the same entity together. NetOwl Extractor combines dynamic recognition with static look-up to achieve high accuracy and coverage at very high speed.

3.3.1.2 Text Analyst: natural language text analysis software


The new text mining system, Text Analyst, implements a variety of important analysis functions based on utilizing an automatically created semantic network of the investigated text. This system is built on the results of twenty years of research and development of a new paradigm by a team of mathematical linguists. The key advantage of TextAnalyst against other text analysis and information retrieval systems is that it can distill the semantic network of a text completely autonomously, without prior development of a subject-specific dictionary by a human expert. The user does not have to provide TextAnalyst with any background knowledge of the subject - the system acquires this knowledge automatically.

3.3.1.3 Mega Search

http://www.megaputer.com./html/megasearch.html

MegaSearch is natural language query based document retrieval tool. It runs fast and easy semantic search and retrieves relevant documents from your PC.
Chapter 3: Literature Review

or an entire local network. MegaSearch turns documents stored on your machine or network into a personal electronic encyclopedia without any effort on your part.

3.3.1.4 Intelligent Miner for Text


Provides a comprehensive suite of text analysis and text search tools:

- The Language Identification tool: This tool automatically discovers the language in which a document is written. You can also train the tool to cover additional languages;

- The Feature Extraction tool: This tool recognizes significant vocabulary items in text, automatically, and without requiring you to predefine a domain-dependent vocabulary;

- The Summarizer tool: Analyzes the words and sentences in a document to produce a summary of the document;

- The Topic Categorization tool: This tool automatically assigns documents to categories, topics, or themes that you have previously defined;

- The Clustering tools: These tools divide up a set of documents into groups, or clusters. The members of each cluster are similar to each other because they share common features. The clusters are not predefined; they are derived from the document collection automatically. Turns unstructured information extracted from workgroup applications and large corporate solutions into business knowledge. Includes components for building scalable knowledge management, text mining and text search applications.
3.3.1.5 ICrossReader

http://www.insight.com.ru

- Hunt only highly relevant documents across WWW.
- Screen texts within an unstructured database and perform semantic clusterization of information.
- Digest the documents to sift paragraphs or sentences relevant to your subject.
- Compose original text surveys.

3.3.1.6 Yahoo Planet

http://www-ai.ijs.si/DunjaMladenic/yplanet.html

Yahoo Planet is a project where Yahoo hierarchy of Web documents was used as a base for automatic document categorization. Several top categories are taken as separate problems, and for each an automatic document classifier is generated.

3.3.1.7 Dataset

http://www.ds-dataset.com/default.htm

DataSet combines Relational-Database (RDB) paradigm with Focused Information Retrieval paradigm. RDB technology is supplemented with DataSet's unique capabilities to manage text. DataSet provides comprehensive search and retrieval tools that can locate Items almost instantly, by words, phrases and much more; Interrelationships between stored items are identified, providing tools that allow navigation through text, with unprecedented ease and accuracy. [28]. Other Related Tools are listed in the table:
## Chapter 3: Literature Review

<table>
<thead>
<tr>
<th>S.NO</th>
<th>Text And NLP Tools</th>
<th>Web Address</th>
<th>Description</th>
<th>Flags</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Concept Map User Interface</td>
<td><a href="http://www.cdimension.com/products.html">http://www.cdimension.com/products.html</a></td>
<td>This patented technology automatically organizes and groups the results list of documents returned by a text search engine, allowing the user to quickly find useful documents and to view only those documents that meet their needs.</td>
<td>Paid</td>
</tr>
<tr>
<td>2</td>
<td>ThemeScape</td>
<td><a href="http://www.cartia.com/products/index.html">http://www.cartia.com/products/index.html</a></td>
<td>creates a visual landscape of information - a topographical map - that actually shows you what's Inside large collections of documents and web pages.</td>
<td>Paid</td>
</tr>
<tr>
<td>3</td>
<td>twURL</td>
<td><a href="http://www.twurl.com">http://www.twurl.com</a></td>
<td>A power tool for the WWW Information Professional who needs to process web content for high quality, relevance, and rapid insight.</td>
<td>Demo version</td>
</tr>
<tr>
<td>4</td>
<td>Taxonomy engine</td>
<td><a href="http://www.semiocom">http://www.semiocom</a></td>
<td>Semio's Taxonomy engine creates a multi-level directory structure that includes thesaurus-like links.</td>
<td>Paid</td>
</tr>
<tr>
<td>5</td>
<td>(SMIsC)</td>
<td><a href="http://www.hintech.ru/smisce/index.htm">http://www.hintech.ru/smisce/index.htm</a></td>
<td>SMIsC is a text-mining computer software tool enabling the user to integrate large unstructured masses of text into browsable networks of local coherence.</td>
<td>Paid</td>
</tr>
<tr>
<td>6</td>
<td>TextFinder</td>
<td><a href="http://www.paracel.com/html/textfinder.html">http://www.paracel.com/html/textfinder.html</a></td>
<td>TextFinder is the fastest, most accurate, adaptive information-filtering system in the world.</td>
<td>Paid</td>
</tr>
<tr>
<td>7</td>
<td>ZyIMAGE</td>
<td><a href="http://www.zylab.nl/zylab1/Products/products.htm">http://www.zylab.nl/zylab1/Products/products.htm</a></td>
<td>ZyIMAGE is a robust imaging system that lets you retrieve documents based on their content with great flexibility and power.</td>
<td>Paid</td>
</tr>
<tr>
<td>8</td>
<td>Knowledge Server</td>
<td><a href="http://www.autonomy.com/knowledge/ksfeatures.html">http://www.autonomy.com/knowledge/ksfeatures.html</a></td>
<td>Automates the accurate categorization and tagging of large volumes of both internal and external information.</td>
<td>Paid</td>
</tr>
<tr>
<td>9</td>
<td>Verity Developer Kit</td>
<td><a href="http://www.verity.com/products/devokit/index.html">http://www.verity.com/products/devokit/index.html</a></td>
<td>Provides unparalleled high-speed search precision. Queries are expanded automatically using thesaurus, word stemming and linguistic analysis.</td>
<td>Paid</td>
</tr>
<tr>
<td>10</td>
<td>InQuery</td>
<td><a href="http://www.sovereign-hill.com">http://www.sovereign-hill.com</a></td>
<td>The most advanced knowledge discovery tool in the industry. InQuery couples advanced intelligent search technology with intuitive concept mining to provide transparent and extensive access to information in over 200 file folders.</td>
<td>Demo version</td>
</tr>
</tbody>
</table>
# Chapter 3: Literature Review

<table>
<thead>
<tr>
<th>Tool</th>
<th>URL</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Dragon Toolkit</td>
<td><a href="http://dragon.ischool.dr.exel.edu">http://dragon.ischool.dr.exel.edu</a></td>
<td>Java development package for Text Mining • Includes tools for text retrieval, classification, clustering, summarization, and topic modeling • Integrates a set of NLP tools</td>
<td>Open Source</td>
</tr>
</tbody>
</table>
Chapter 3: Literature Review


Table 3.1 A List of existing text mining and NLP Tools

3.3.2 Cross Language Information Retrieval Tools

The design of a multilingual system faces specific challenges regarding the best way of handling multiple languages, best query translations and presentational requirements for the user. In the past, most research has been focused on the retrieval effectiveness of cross language systems through IR test collection approaches [29]; whereas few researchers focused on the user interface requirements with respect to the multilingual retrieval task [30]. Despite the clear effort which has been directed toward retrieval functionality and effectiveness, only little attention was paid to developing multilingual interaction tools, where users are really considered as an integral part of the retrieval process. One potential interpretation of this problem is that users of CLIR might not have sufficient knowledge of the target languages and therefore they are usually not involved in multilingual processes [31]. However, the involvement of the user in CLIR systems, by reviewing and amending the query, have been studied, e.g., the New Mexico State University Keizai system [30], the German Research Center for Artificial Intelligence’s (DFKI) MULINEX system [32], a Multilingual Information Retrieval Tool
Chapter 3: Literature Review

UCLIR [33], the Maryland Interactive Retrieval Advanced Cross-Language Engine MIRACLE system [34], MultiLexExplorer [35] and In the following we describe these CLIR tools in detail.

3.3.2.1 Mulinex

Mulinex supports cross-lingual search by giving the users possibilities to formulate, expand and disambiguate queries. Furthermore, the users are able to filter the search results and read the retrieved documents by using only their native language. Mulinex performs the multilingual functionality based on a dictionary-based query translation. Besides the cross lingual functionality, where the query is submitted in one language and the retrieved documents are presented in another language, Mulinex provides the automatic translation of documents and their summaries. In Mulinex, three languages are supported, French, German, and English. In Mulinex, the CLIR process is fully supported by the translation of the queries, documents and their summaries. Hereby, users do not need to have any knowledge about the target language. Mulinex provides a lot of functionality to support the retrieving of the documents in multilingual collections. Examples of these functionalities are translation of the user’s query, interactive disambiguation of the query translation, interactive query expansion, on-demand translation of summaries and search results, etc., The Mulinex interface is available in three languages English, German, and French. Since the search engine queries are usually between 2.4 and 2.7 in length [35] which typically does not provide enough context for automatic disambiguation, Mulinex using ”query assistant” provides an opportunity for interactive query translation disambiguation. This task is performed by the “query assistant” by performing the back translation. The translated query terms are translated back into the original query language. However, this approach has some clear limitations. When no synonyms can be found in the dictionary, the technique is not helpful; and significant homonymy in the target language can result in confusing back translations [34]. In Mulinex, the back translation concept is used for expanding the original query with potentially
relevant terms. The query term translation is translated back to the original query language; the result of this step is having a list of possible translation in the query’s original language. The user, in this case, can select some of these translation alternatives, in order to expand the user query. For example, the user submits the query, “fair”, in English. The system provides the user with alternative translations in French and German. In order to expand the query, the system translates back the translated user query terms. The result of this step is having translation alternatives in the user’s original query language [36].

Mulinex Search Functionality

- Your search queries are translated to the other languages
- With the "Query Assistant" you can look at the translations and refine them
- Search results are presented with their language, categories and a summary
- You can filter the results by language and category
- Automatic translations of summaries and web pages will be added soon

Personal Preferences

- MULINEX is available in German, French and English (you can switch at any time!)
- You can become a registered user and store your personal preferences
- Registered users are (optionally) informed about system updates
- A personal agent for registered users will be available in March 1999

3.3.2.2 Keizai

The goal of the Keizai project is to provide a Web based cross-language text retrieval system that accepts the query in English and searches Japanese and Korean web data. Furthermore, the system displays English summaries of the top ranking retrieved documents. In Keizai the query terms are translated into Japanese or Korean languages along with their English definitions and thus this feature allows the user to disambiguate the translations based on the English definitions of the translated query terms, the user who does not understand the
Japanese or Korean language can select the appropriate translation, out of several possible translations. Once the user selects those translations whose definitions are consistent with the information needed, the search can be performed. Only documents that are relevant to the selected translations will be retrieved. For each retrieved document in Japanese or Korean, an English summary along with a target document language summary will be displayed in the Keizai interface. Keizai investigates the effectiveness of representing the retrieved documents together with small images, which they call “Document Thumbnail Visualizations”. Using this document representation, the retrieved documents are retained with a familiar shape and format and thus the user can see how the query terms are distributed in the retrieved documents [37].

3.3.2.3 UCLIR

In UCLIR, the Arabic language was included. The system performs its task in any of the following three different modes: the first mode, using a multilingual query (query can consist of terms of different languages), the second mode using English query without user involvement in the multilingual query formulation, the third mode using an English query with user involvement in the formulation of the multilingual queries. The first system mode: Multilingual query, in this mode the system accepts a query which consists of terms of various languages. The system will retrieve the relevant documents regardless of the query term language. The documents in the entire multilingual collection, those relevant to one of the query terms, will be retrieved. The second system mode: English query: non-interactive approach, this mode is based on the use of a set of bilingual dictionaries for translating an English query into the different target languages. First, for the English query term a set of possible translations will be obtained from the bilingual dictionaries. Second, the set of possible translations will be compared with an index word list (obtained from the system’s entire multilingual resource); the translations which are not in the index word list will be eliminated from the query. The filtered query then can be used to retrieve the
Chapter 3: Literature Review

relevant documents from the system’s entire resource and these retrieved
documents are then displayed to the user in the system interface. The third system
mode: English query: interactive approach, in this system mode, the user is
involved in the selection of appropriate translations. The same as in the second
mode, a set of possible translations will be obtained from the bilingual
dictionaries and compared with the index word list; the translations which are not
in the index word list will be eliminated from the query. The rest will be kept and
presented to the user in the system interface along with their English translation
beside other information e.g., part of speech. At the end, the user selects the
appropriate translation out of the filtered translation list. The selected multilingual
terms then can be used to form the multilingual query which is then submitted to
retrieve the relevant documents from the system’s entire multilingual resource.
After the retrieval process is performed, the relevant retrieved documents can be
then translated into English. To perform the document translation, two approaches
are used. The first approach is word-level translation, where the user can click on
the selected word and this word will be translated using the dictionary and
displayed as a popup view to the user with its lexical information. The second
approach is a document-level translation, where the whole retrieved document,
using a translation system, is translated into English. Similar to Keizai, UCLIR
uses “Document Thumbnail Visualizations”. The retrieved documents are
retained with familiar shape and format which make it possible for the user to see
how the query terms are distributed in the retrieved documents. Although the
system in the second mode automates the process of the appropriate translation
selection by comparing a set of possible translations with an index word list (the
translations which are not in the index word list will be eliminated from the
query). However, this can include an irrelevant translation to the user query since
it is possible that not all translations can be relevant to the original query terms
[38].
3.3.2.4 MIRACLE

In order to support the interactive CLIR, the system uses the user-assisted query translation. The user assisted-query translation feature supports the user to select the correct translation. However, there might be a case when the user might delete a correct translation. The system reacts, in that the searcher can see the effect of the choice and have possibilities to learn better control of the system. This is done by providing the following features, the meaning of the translation (loan word or proper name), using back translation, a list of possible synonyms are provided. Translation examples of usage are obtained from translated or topically-related text.

In MIRACLE, there are two types of query translations, fully automatic query translation (using machine translation) and user-assisted query translation. In fully automatic translation the user can be involved only once. After the system translates the query and retrieves the search results, the user can refine the query if he/she isn’t satisfied after examining the search results. In the user-assisted query translation, four possible refinement steps give the user an opportunity to be involved in the translation process. First, based on evidence about the meanings of the proposed translations by the system, the user has an opportunity to deselect some of the proposed translations before the search can be performed. Second the user can reform the query based on evidence about the meanings of the proposed translations. Third, the user can reform the query based on examining the search results. Fourth, in case the search result doesn’t satisfy the user’s needs, the user has a possibility to deselect/reselect the translations. In other words, the user submits his query; the system provides him/her with translation alternatives. Before the search can be performed, the user has an opportunity to deselect some of the proposed translations. The user has an opportunity to refine his/her query based on evidence about the meanings of the proposed translations by the system. After the search is performed, the system provides the user with the search results. If the user is satisfied with the search result then there will be no further actions by the system. In contrast, based on examining the search result, the user has two
opportunities: refine his/her query and perform a new search or deselect/reselect a translation out of the translation alternatives proposed by the system. The interaction between the system and the user, gives the user possibilities to see the effect of his/her decision (selection, de-selection of the translation or query refinement) in that the user can cycle the search till it satisfies his/her needs. A very important aspect in MIRACLE is that the system provides the user with immediate feedback in response to any action, which gives the user an important opportunity to refine his/her search. The rapid integration of new languages was taken into account in the design of the MIRACLE system. The query language is always English, in MIRACLE. However, language resources that are available for English can be leveraged, regardless of the document language. Currently, MIRACLE works with a simple bilingual term list. However, it is designed to readily leverage additional resources when they are available

3.3.2.5 MultiLex Explorer

The goal of the MultiLexExplorer tool is to support multilingual users in performing their web search. Furthermore, the MultiLexExplorer supports the user in disambiguating word meanings by providing the user with information about the distribution of words in the web. The tool allows users to explore combinations of query term translations by visualizing EuroWordNet1 relations together with search results and search statistics obtained from web search engines. Based on the EuroWordNet, the tool supports the user with the following functionality:

- Exploring the context of a given word in the general hierarchy, searching in different languages, e.g., by translating word senses using the interlingual index of EuroWordNet,
- Disambiguating word sense for combinations of words
- provide the user with the possibility to interact with the system i.e., changing the search word and the number of retrieved documents,
Chapter 3: Literature Review

- Expanding the original query with extra relevant terms, and in automatically categorizing the retrieved web documents [36].

3.4 Natural language in IR and CLIR

Natural languages are constantly evolving, quite vague means of communication. Rather than being a demerit, vagueness tells about adaptability and power of expression of natural languages (Karlsson 1994, 3). Because of adaptability of natural languages, introducing new concepts, as well as using old concepts in new contexts is possible. Naturally, due to vagueness, meanings of natural language expressions are not stable. For example, the meaning of words may broaden, narrow or change according to human needs (Akmajian & al. 1995, 85-87).

Natural languages can be seen as structured systems. There are five sub-systems which construct a language: semantic, phonologic, lexical, morphological and syntactic sub-systems. Semantics deals with the meaning, while phonology deals with phonetic forms of languages. The lexical system is constructed of the words of a language. Morphology deals with the internal structure of words, and syntax with the structure and the rules of sentences. Of these sub-systems, phonologic, lexical, morphological and syntactic subsystems are tangible or linguistic systems, while semantics is intangible dealing with human meaning construction processes. (Karlsson 1994, 14-15.) All natural languages are constructed of words and sentences. A word can be 1) a lexical word, that is the word in its basic form (a cat), 2) a word in its inflected form (cats), 3) a derived word (catty), or 4) a word consisting of two or more single words (a catfish). Words 1-3 are all irreducible words: they contain only one “autonomous” word (Karlsson 1994, 75-76). Number 4) is sometimes called a compound, but there are also broader definitions for a compound. Compounds can be divided in three types: 1) the closed form, where the parts of a compound are written together (makeup), 2) the hyphenated form, where the parts are attached by a hyphen (daughter-in-law), and 3) the open form (post office). In the present research, the
closed form and the hyphenated form are called compounds, while the open form is called a phrase. Compounds of this definition are quite rare in English, while for example Finnish, Swedish and German have a lot of them. Phrases are used in English instead of compounds. The present thesis concentrates on morphological aspects of natural languages. The most crucial morphological concepts for this research are a morpheme, a stem and a root (also called a linguistic root). A morpheme is the smallest linguistic unit which has an independent meaning or at least a linguistic function. For example the word unbelievable has three morphemes: un, believ and able. A stem is the base part of a word not including inflectional morphemes. The stem for the word cats is cat, and the stem of the word unbelievable is unbeliev. A root is the minimal unit of the stem representing the semantic content of the word - thus, it cannot be split into smaller parts. (Karlsson 1994, 100-101.) In English, the root and the stem are often the same. The root of the word cats is cat, and the root of the word unbelievable is believ.

There are two schools of thought in the IR field with respect to the choice of the language level and language processing in indexing and retrieval. One prefers the semantic level approach. This approach could be called conceptual information retrieval. Semantic processing requires large amounts of pre-coded knowledge. The other school of thought bases its approach on the morphologic level. It states that it is not important to understand the concepts, but locate the relevant documents. This approach is sometimes called natural language information retrieval, while it operates only at the morphological level. (Sheridan & Smeaton 1992.)

Morphological properties of natural languages are very diverse. Pirkola has presented a morphological classification of languages for IR purposes. It is based on two variables: the index of synthesis and the index of fusion. The index of synthesis tells the number of affixes in a language, while the index of fusion describes the ease with which affixes can be segmented in words in a language. These indicators can be used for developing and evaluation of IR systems. (Pirkola 2001.)
3.4.1 Problems caused by natural language features for IR and CLIR

The vagueness of natural languages is a problem from the IR point of view. Natural languages are flexible and constantly evolving systems: for example, a word may have various senses depending on the context and new words are created constantly. (Ingwersen & Järvelin 2005, 151.)

The meaning of a single natural language word is not necessarily precise. A word may even have several quite varied meanings. On the other hand, there might be several separate words to express the same concept. The well-known phenomenon where two or more words have an equal sense or denotation is called synonymy (Karlsson 1994, 217-218). Exact synonymy is quite rare, however, because words are often synonymous only in some context(s). Sometimes the word quasisynonymy is used instead of synonymy to refer to words which have the same meaning in some context. (Pirkola 1999, 23.)

Synonymy causes difficulties for IR because various (synonymous) words may be used in documents to refer to the same concept. The user should supply all those words in the query in order to retrieve all the documents discussing the subject. In CLIR, the situation is more complicated. If the source language query includes all the synonyms and quasisynonyms for a word, the translated query may become too broad. On the other hand, sometimes translation acts like query expansion in a good sense, because translation dictionaries often include synonyms for a given word. Homonymy and polysemy are the opposites of synonymy. Homonymy and polysemy are related concepts and it is hard to make a clear distinction between them. The traditional distinction is that in polysemy, one word has several senses, and in homonymy, two different words happen to have the same form. An etymological criterion for making the distinction may be utilized: if a word has differing sound or spelling variants for its senses in the history, it is not a polysemy, but a homonymic word. The etymology of a word may be unknown, however. (Kilgarriff 1992.) In addition, this might not hold in Finnish, for example, where sound and spelling variants are rare. In Finnish, the following criterion may be used: if the inflection rules for various
Chapter 3: Literature Review

senses of the word are equal, the word is polysemic (Karlsson 1994, 213-214). For example a Finnish word kieli is a polysemic word. It means a tongue, a language, speech, a flap, a clapper and a string of an instrument. The Finnish word kuusi is a homonymous word. It has two meanings: six and a spruce, and their inflection rules differ. In addition, a special case for homonymy may be defined: inflectional homonymy. It means that some inflected forms of words happen to be identical while the lemmas are different. For example the Finnish inflected word hauista has three meanings: from retrievals, from pikes, biceps (the partitive form). Homonymy and polysemy may cause problems for IR and CLIR: if the query includes a word with many senses, documents with any of those senses will be retrieved, even if the user would be interested in one sense only. For example if the user inputs a one word query kuusi, searching for information on spruces, he would retrieve documents including number six as well. The problem is often solved by supplying more query words, which together disambiguate the query in a natural way. (Pirkola 1999, 14.)

Compounds and phrases are problematic for IR and CLIR. The headword of a compound may be inaccessible in retrieval, which might cause loss of relevant documents. In addition, the meaning of a compound or a phrase is often more than the meaning of their constituents alone. Word inflection causes difficulties for IR and CLIR, especially for languages with strong morphology. Spelling errors as well as spelling variation across languages are problematic as well. (Ingwersen & Järvelin 2005, 151.) [22]

3.5 Developments in Indian Language IR system: CLIR and MONOLINGUAL Information Retrieval.

3.5.1 Bengali and Hindi to English CLIR

Debasis Mandal, Mayank Gupta, Sandipan Dandapat, Pratyush Banerjee, and Sudeshna Sarkar Department of Computer Science and Engineering IIT Kharagpur, India presented a cross-language retrieval system for the retrieval of
Chapter 3: Literature Review

English documents in response to queries in Bengali and Hindi, as part of their participation in CLEF 2007 Ad-hoc bilingual track. They followed the dictionary-based Machine Translation approach to generate the equivalent English query out of Indian language topics. Their main challenge was to work with a limited coverage dictionary (of coverage \( \leq 20\% \)) that was available for Hindi-English, and virtually non-existent dictionary for Bengali-English. So they depended mostly on a phonetic transliteration system to overcome this. The CLEF results point to the need for a rich bilingual lexicon, a translation disambiguator, Named Entity Recognizer and a better transliterator for CLIR involving Indian languages [39].

3.5.2 Hindi and Marathi to English Cross Language Information Retrieval

Manoj Kumar Chinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om P. Damani Department of CSE IIT Bombay presented Hindi -> English and Marathi->English CLIR systems developed as part of their participation in the CLEF 2007 Ad-Hoc Bilingual task. They took a query translation based approach using bi-lingual dictionaries. Query words not found in the dictionary are transliterated using a simple rule based approach which utilizes the corpus to return the ‘k’ closest English transliterations of the given Hindi/Marathi word. The resulting multiple translation/transliteration choices for each query word are disambiguated using an iterative page-rank style algorithm which, based on term-term co-occurrence statistics, produces the final translated query. Using the above approach, for Hindi, they achieve a Mean Average Precision (MAP) of 0.2366 in title which is 61.36\% of monolingual performance and a MAP of 0.2952 in title and description which is 67.06\% of monolingual performance. For Marathi, they achieve a MAP of 0.2163 in title which is 56.09\% of monolingual performance [40].
3.5.3 Hindi and Telugu to English Cross Language Information Retrieval

Prasad Pingali and Vasudeva Varma Language Technologies Research Centre IIIT, Hyderabad presented the experiments of Language Technologies Research Centre (LTRC) as part of their participation in CLEF 2006 ad-hoc document retrieval task. They focused on Afaan Oromo, Hindi and Telugu as query languages for retrieval from English document collection and contributed to Hindi and Telugu to English CLIR system with the experiments at CLEF [41]

3.5.4 FIRE-2008 at Maryland: English-Hindi CLIR

Tan Xu and Douglas W.Oard College of Information Studies and CLIP Lab, Institute for Advanced Computer Studies, University of Maryland. Forum for Information Retrieval Evaluation (FIRE), the University of Maryland participated in the Ad-hoc task cross-language document retrieval task, with English queries and Hindi documents. Their experiments focused on evaluating the effectiveness of a “meaning matching” approach based on translation probabilities. The FIRE Hindi test collection provides the first opportunity to carefully assess some of the resources and techniques developed for the Translingual Information Detection, Extraction And Summarization (TIDES) program’s “Surprise Language” exercise in 2003, in which a broad range of language engineering tools were constructed for Hindi in a comparatively short period. The results reported appear to confirm that some of the language resources developed for the Surprise Language exercise are indeed reusable, and that meaning matching yields reasonably good results with less carefully constructed language resources than had previously been demonstrated [42].

3.5.5 English to Kannada / Telugu Name Transliteration in CLIR

Mallamma v reddy, Hanumanthappa Department of Computer Science and Applications, Bangalore University. They present a method for automatically learning a transliteration model from a sample of name pairs in two languages. Transliteration is mapping of pronunciation and articulation of words written in
Chapter 3: Literature Review

one script into another script. Transliteration should not be confused with translation, which involves a change in language while preserving meaning. CLIR is the acronym of a great variety of techniques, systems and technologies that associate information retrieval (normally from texts) in multilingual environments. Dictionaries have often been used for query translation in cross language information retrieval (CLIR). However, they are faced with the problem of translating Names and Technical Terms from English to Kannada/Telugu. The most important query words in information retrieval are often proper names. [43].

3.5.6 Kannada and Telugu Native Languages to English Cross Language Information Retrieval

Mallamma v reddy, Hanumanthappa Department of Computer Science and Applications, Bangalore University conducted experiments on translated queries. One of the crucial challenges in cross lingual information retrieval is the retrieval of relevant information for a query expressed in as native language. While retrieval of relevant documents is slightly easier, analyzing the relevance of the retrieved documents and the presentation of the results to the users are non-trivial tasks. To accomplish the above task, they present their Kannada English and Telugu English CLIR systems as part of Ad-Hoc Bilingual task by translation based approach using bi-lingual dictionaries. When a query words not found in the dictionary then the words are transliterated using a simple rule based approach which utilizes the corpus to return the ‘k’ closest English transliterations of the given Kannada/Telugu word. The resulting multiple translation/transliteration choices for each query word are disambiguated using an iterative page-rank style algorithm which, based on term-term co-occurrence statistics, produces the final translated query. Finally they conduct experiments on these translated queries using a Kannada/Telugu document collection and a set of English queries to report the improvements, performance achieved for each task [44].
Chapter 3: Literature Review

3.5.7 Bilingual Information Retrieval System for English and Tamil

Dr. S. Saraswathi, Asma Siddhiqaa.M, Kalaimagal.K, Kalaiyarasi.M addresses the design and implementation of BiLingual Information Retrieval system on the domain, Festivals. A generic platform is built for BiLingual Information retrieval which can be extended to any foreign or Indian language working with the same efficiency. Search for the solution of the query is not done in a specific predefined set of standard languages but is chosen dynamically on processing the user’s query. Their research deals with Indian language Tamil apart from English. The task is to retrieve the solution for the user given query in the same language as that of the query. In this process, an Ontological tree is built for the domain in such a way that there are entries in the above listed two languages in every node of the tree. A Part-Of-Speech (POS) Tagger is used to determine the keywords from the given query. Based on the context, the keywords are translated to appropriate languages using the Ontological tree. A search is performed and documents are retrieved based on the keywords. With the use of the Ontological tree, Information Extraction is done. Finally, the solution for the query is translated back to the query language (if necessary) and produced to the user [45].

3.5.8 Recall Oriented Approaches for improved Indian Language Information Access

Pingali V.V. Prasad Rao Language Technologies Research Centre International Institute of Information Technology Hyderabad:

Their research is an investigation into Indian language information access. The investigation shows that Indian language information access technologies face severe recall problem when using conventional IR techniques (used for English-like languages). During this investigation they crawled the web extensively for Indian languages, characterized the Indian language web and in the process came up with some solutions for the low recall problem. They focused their
investigation on the loss of recall in monolingual and cross-lingual based IR and text summarization. The following are some of their major contributions.

- They showed that Indian language information access technologies that use state-of-the-art technologies used by English like languages, face low recall. They observed the recall loss to be relatively higher when the target language corpus is English.
- They came up with a unified information access framework which can address the problems of monolingual and Cross-lingual Information Retrieval and Text Summarization.
- They showed that, word spelling normalization is an essential component of Indian language information access systems and proposed a linguistically motivated rule based approach and showed that this approach works better than the various approximate string matching algorithms.
- They modeled the problem of Dictionary based query translation as an IR problem [46].

### 3.5.9 A high recall error identification tool for Hindi Treebank Validation

Bharat Ram Ambati, Mridul Gupta, Samar Husain, Dipti Misra Sharma
Language Technologies Research Centre, International Institute of Information Technology Hyderabad.

Their proposed tool has been used for validating the dependency representation of a multi-layered and multi representational tree bank for Hindi (Bhatt et al., 2009). The tool identifies errors in the Hindi annotated data at POS, chunk and dependency levels. Additionally, the identification of errors can help resolve ambiguous cases and thus improve the guidelines for annotation. Improved guidelines will directly make the task of annotation more consistent.

They proposed a new tool which uses both rule-based and hybrid systems to detect errors during the process of treebank annotation. They tested it on Hindi dependency treebank and were able to detect 75%, 62.5% and 40.33% of errors in
Chapter 3: Literature Review

POS, chunk and dependency annotation respectively. For detecting POS and chunk errors, they used the rule-based system. For dependency errors, they used the combination of both rule-based and hybrid systems. The proposed approach works reasonably well for relatively smaller annotated datasets [47].

3.5.10 English Bengali Ad-hoc Monolingual Information Retrieval Task Result at FIRE 2008

Sivaji Bandhyopadhyay, Amitava Das, Pinaki Bhaskar Department of Computer Science and Engineering Jadavpur University, Kolkata.

Their experiments suggest that simple TFIDF based ranking algorithms with positional information may not result in effective ad-hoc mono-lingual IR systems for Indian language queries. Any additional information added from corpora either resulting in query expansion could help. Application of certain machine learning approaches for query expansion through theme detection or event tracking may increase performance. Document-level scoring entailment technique also could be a new direction to be explored. Application of word sense disambiguation methods on the query words as well as corpus would have a positive effect on the result. A robust stemmer is required for the highly inflective Indian languages [48].

3.5.11 Using Morphology to Improve Marathi Monolingual Information Retrieval

Ashish Almeida, Pushpak Bhattacharyya IIT Bombay. They study the effects of lexical analysis on Marathi monolingual search over the news domain corpus (obtained through FIRE-2008) and observe the effect of processes such as lemmatization, inclusion of suffixes in indexing and stop-words elimination on the retrieval performance. Their results show that lemmatization significantly improves the retrieval performance of language like Marathi which is agglutinative in nature. Also, it is observed that indexing of suffix terms, which show spacio-temporal properties, further improve the precision. Along with these,
the effects of elimination of stop-words are also observed. With all three methods combined they are able to get mean average precision (MAP) of 0.4433 for 25 queries [49].

3.5.12 A Query Answering System for E-Learning Hindi Documents

Praveen Kumar, Shrikant Kashyap, Ankush Mittal Indian Institute of Technology, Roorkee, India developed a Question Answering (QA) System for Hindi documents that would be relevant for masses using Hindi as primary language of education. The user should be able to access information from E-learning documents in a user friendly way, that is by questioning the system in their native language Hindi and the system will return the intended answer (also in Hindi) by searching in context from the repository of Hindi documents. The language constructs, query structure, common words, etc. are completely different in Hindi as compared to English. A novel strategy, in addition to conventional search and NLP techniques, was used to construct the Hindi QA system. The focus is on context based retrieval of information. For this purpose they implemented a Hindi search engine that works on locality-based similarity heuristics to retrieve relevant passages from the collection. It also incorporates language analysis modules like stemmer and morphological analyzer as well as self constructed lexical database of synonyms. The experimental results over corpus of two important domains of agriculture and science show effectiveness of their approach. [50]

3.5.13 Om: One tool for many (Indian) languages

Ganpathiraju Madhavi, Balakrishnan Mini, Balakrishnan N., Reddy Raj (Language Technologies Institute, Carnegie Mellon University, Pittsburgh) (Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India)

Om transliteration and integrated editor have been developed by a large number of people at the
Chapter 3: Literature Review

Multimedia Systems Lab at the Supercomputer Education and Research Centre, Indian Institute of Science and at the ISRI, Carnegie Mellon University. Of particular mention are the names of Sravan Kumar, Jiju Verghese, Sheik, Tina Joseph and Umi who contributed towards specific Indian languages. Jiju Verghese developed the Web interface and the Java standalone version was developed by Atul Kumar.

They describe the development of a transliteration scheme Om which exploits this phonetic nature of the alphabet. Om uses ASCII characters to represent Indian language alphabets, and thus can be read directly in English, by a large number of users who cannot read script in other Indian languages than their mother tongue. It is also useful in computer applications where local language tools such as email and chat are not yet available. Another significant contribution presented in their research is the development of a text editor for Indian languages that integrates the Om input for many Indian languages into a word processor such as Microsoft WinWord. The text editor is also developed on Java platform that can run on UNIX machines as well. They propose this transliteration scheme as a possible standard for Indian language transliteration and keyboard entry [51].

3.5.14 A multimodal Indian language interface to the computer

Hema A Murthy, C Chandra Sekhar Dept. of Computer Science & Engineering IIT Madras, Chennai developed a multimodal interface to the computer that is relevant for India. Although India’s average literacy level is about 65%, less than 5% of India’s population can use English for communication. And even though the world-wide web and computer communication has given us access to information at the click of a mouse, 95% of our population is excluded from this revolution due to dominance of English. To overcome this problem they propose to set up an Indian Language Systems Laboratory at IIT Madras. Their initial goal will be to develop a multimodal interface to the computer that is relevant for India, i.e., one that enables Indic computing [52]. The components of this Indian language interface will be:

1. Keyboard and display interface
2. Speech interface
3. Handwriting interface

3.5.15 Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey

Dinesh Kumar, Gurpreet Singh Josan Department of Information Technology DAV Institute of Engineering & Technology Jalandhar, Punjab, INDIA

The problem of tagging in natural language processing is to find a way to tag every word in a text as a particular part of speech, e.g., proper pronoun. POS tagging is a very important preprocessing task for language processing activities. Their paper reports about the Part of Speech (POS) taggers proposed for various Indian Languages like Hindi, Punjabi, Malayalam, Bengali and Telugu. Various part of speech tagging approaches like Hidden Markov Model (HMM), Support Vector Model (SVM), and Rule based approaches, Maximum Entropy (ME) and Conditional Random Field (CRF) have been used for POS tagging. Accuracy is the prime factor in evaluating any POS tagger so the accuracy of every proposed tagger is also discussed in this paper [53].

3.5.16 Post Translation Query Expansion using Hindi Word-Net for English-Hindi CLIR System

Sujoy Das, Anurag Seetha, M. Kumar, J.L. Rana have investigated impact of query expansion using Hindi WordNet in the context of English-Hindi CLIR system. The WordNet is a lexical database, machine readable thesaurus Hindi language. They have translated English query using Shabdanjali dictionary. The translated queries have been expanded using Hindi WordNet and nine query expansion strategies have been formulated. In these runs title field of topic was used for query formulation and expansion and in one run title + description field was used for query formulation and expansion. The queries are translated, then expanded and are submitted to the retrieval system to retrieve documents from the
Chapter 3: Literature Review

Fire Hindi Test collection. Their observations suggest that simple query expansion using Hindi WordNet is not effective for English- Hindi CLIR system [54].

3.5.17 Automated Evaluation of Search Engine Performance via Implicit User Feedback

Himanshu Sharma Department of Industrial and Manufacturing Engineering. The Pennsylvania State University
Bernard J. Jansen School of Information Sciences and Technology.
Measuring the information retrieval effectiveness of Web search engines can be expensive if human relevance judgments are required to evaluate search results. Using implicit user feedback for search engine evaluation provides a cost and time effective manner of addressing this problem. Web search engines can use human evaluation of search results without the expense of human evaluators. An additional advantage of this approach is the availability of real time data regarding system performance. They capture user relevance judgments actions such as print, save and bookmark, sending these actions and the corresponding document identifiers to a central server via a client application. They use this implicit feedback to calculate performance metrics, such as precision. They can calculate an overall system performance metric based on a collection of weighted metrics [55].

3.5.18 Learning User Interaction Models for Predicting Web Search Result Preferences

Eugene Agichtein Eric Brill Susan Dumais Robert Ragno Microsoft Research
Evaluating user preferences of web search results is crucial for search engine development, deployment, and maintenance. They present a real-world study of modeling the behavior of web search users to predict web search result preferences. Accurate modeling and interpretation of user behavior has important
applications to ranking, click spam detection, web search personalization, and other tasks. Their key insight to improving robustness of interpreting implicit feedback is to model query-dependent deviations from the expected “noisy” user behavior. They show that their model of clickthrough interpretation improves prediction accuracy over state-of-the-art clickthrough methods. They generalize their approach to model user behavior beyond clickthrough, which results in higher preference prediction accuracy than models based on clickthrough information alone. They report results of a large-scale experimental evaluation that show substantial improvements over published implicit feedback interpretation methods [56].

3.5.19 Incorporating user search behavior into relevance feedback

Ian Ruthven, Mounia Lalmas and Keith van Rijsbergen in their paper present five user experiments on incorporating behavioral information into the relevance feedback process. In particular they concentrate on ranking terms for query expansion and selecting new terms to add to the user’s query. Their experiments are an attempt to widen the evidence used for relevance feedback from simply the relevant documents to include information on how users are searching. We show that this information can lead to more successful relevance feedback techniques. They also show that the presentation of relevance feedback to the user is important in the success of relevance feedback [57].

3.6 Conclusion

As far as development in IR with respect to Indian languages is concerned, a lot work has been done and still going on particularly in the field of cross language information retrieval. Research is also going on in other related areas as well such as NLP machine translation etc. Various regional languages have been taken into consideration by researchers for CLIR. Even government organization like TDIL (Technology Development for Indian Languages) has made significant contributions for standardization of Indian Languages on web. But at the same
time a very less work has been done in the field of monolingual IR for Hindi language in particular. It seems that Hindi language which is the national language of India and widely used worldwide has not been given much importance. Indian Search engines like Guruji, Raftaar etc. are now present for Hindi IR but the monolingual issues are not well addressed by any of them. The objective of the research work is to highlight various issues involved in monolingual IR and suggest ways and means to solve those issues through the design and development of a specialized tool which will take care of various highlighted issues.