CHAPTER I

INTRODUCTION

1.1. INFORMATION RETRIEVAL

The success of digital revolution, along with the growth of Internet technology, has made voluminous information available to users of World Wide Web (WWW). Recent years have envisaged a tremendous growth in the amount of this online information, which is created by various private, government, business and scientific sectors.

In the 21st century's information age, swift access to relevant information, in whatever form or medium, can dictate the success or failure of businesses or individuals. The timely provision of relevant information with minimal 'noise' is critical to modern society, and Information Retrieval (IR) is a field which is focused on this aspect. It is a dynamic subject, with current changes driven by the expansion of the World Wide Web, the advent of modern and inexpensive graphical user interfaces and the development of reliable and low-cost mass storage devices.

Information Retrieval is defined as finding text documents of an unstructured nature, that satisfies an information need (query), from within a collection that are usually stored in computers or servers (Manning *et al.*, 2009). An automated IR system responds to an IR request (termed as query) that contains desired characteristics of document components that have to be retrieved from a huge collection of documents.

An Information Retrieval system is defined as any system that matches a user request against a document collection, returning a list of documents considered relevant to the request. The user request is an expression of a user information need. Traditionally, users made such requests to a professional librarian, who would then suggest reading materials (relevant documents). The aim of IR system is to carry out a similar task automatically, using a software programme that stores and manages information on documents. The system assists users in finding the information they need. Unlike the so-called question answering systems (Voorhees, 2000), the system does not explicitly return information or answer questions. Instead, it informs

on the existence and location of documents that might contain the needed information. Some suggested documents will, hopefully, satisfy the user's information need. These documents are called relevant documents. Thus, IR is defined as the task to retrieve certain documents from a collection that satisfies a query or information needs.

A perfect retrieval system would retrieve only the relevant documents and not the irrelevant documents. However, perfect retrieval systems do not exist and will not exist, because, search statements are necessarily incomplete, and relevance depends on the subjective opinion of the user. Two users may pose the same query to an information retrieval system and give different relevance judgements on the retrieved documents.

Yates and Neto (1999) presents a formal model of IR as $[D,Q,F,R(q_i,d_j)]$, where D is a set of logical representations for the documents, Q is a set composed of logical representations of the user information needs called queries, F is a framework for modeling the documents, queries and their relationships and $R(q_i , d_j)$ is a ranking function. From this formal definition, it can be understood that the basic processes that should be supported by any IR system are,

- The representation of the content of the documents
- The representation of the user's information need
- The comparison of the two representations.

These IR processes are illustrated in Figure 1.1, where squared boxes represent data and rounded boxes represent processes (Croft, 1993; Hiemstra, 2001). Representation of the documents or indexing is an offline process and is used to obtain a formal representation of the document and does not involve the end user. The process of representing the information problem or need is often referred to as the query formulation process. The resulting formal representation is the query. Query formulation denotes the complete interactive dialogue between system and the end user, leading not only to a suitable query but possibly also to a better understanding by the user of his/her information need. In this process, the user and the system communicate the information need using queries from users, which is used by the system to retrieve a set of related documents.

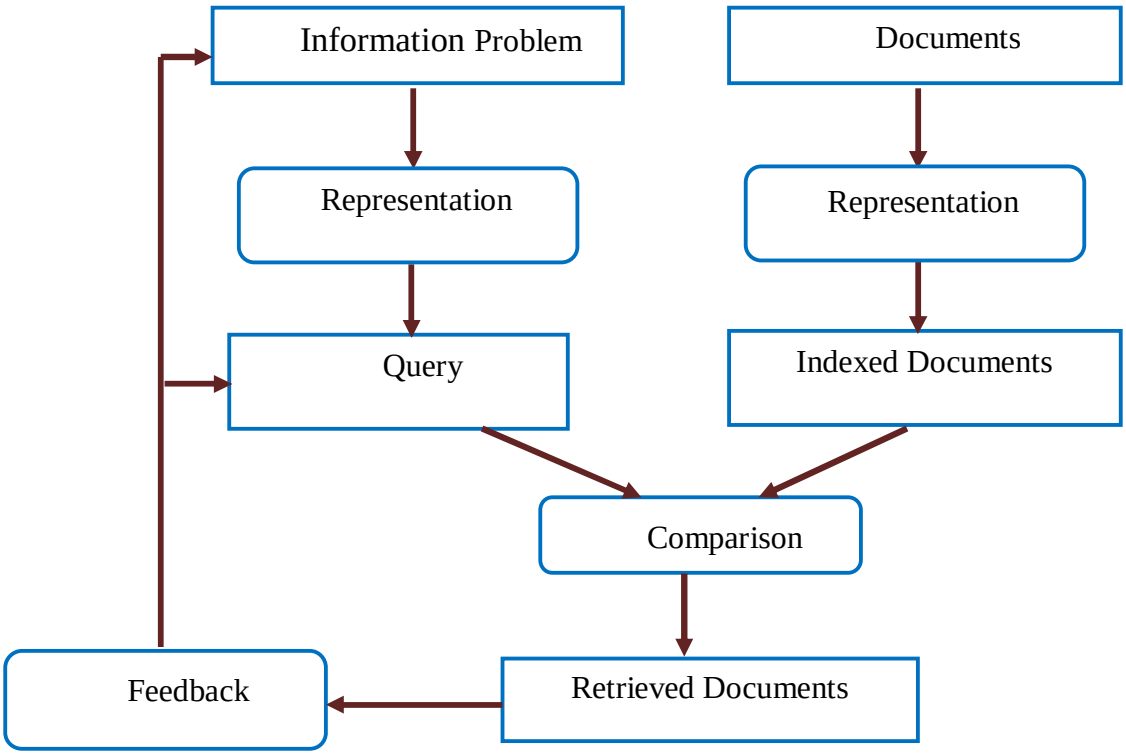

Figure 1.1: Processes of IR Systems

In this stage, humans would use natural language to communicate the information need, called as a request, and the resultant set of documents are treated as output of initial query. Then a relevance feedback is used to reformulate the query or request, to further refine the information need of the user and to improve the retrieval of relevant documents. Automatic query formulation inputs the request and outputs an initial query.

The next process is the comparison of the query against the document representations and is also called the matching process. The matching process results in a ranked list of relevant documents. Users can browse this list in search of the information they need. Ranked retrieval aims to place the high relevant documents at the top of the ranked list, thus minimizing the time the user has to invest on searching for relevant documents.

Traditional IR systems identify and retrieve relevant text documents in the same language as the query, generally termed as monolingual IR. Due to the multilingual environment of WWW, the field of IR is extended to include a subfield, called Cross-Language Information

Retrieval or Cross-Language Text Retrieval. The basics of CLTR are given in the following section.

1.2. CROSS-LANGUAGE INFORMATION RETRIEVAL

Cross–language information retrieval is a subfield of the traditional information retrieval. It provides users with access to information that is in a different language from their queries (Chen, 2006) and is defined as the process of retrieving information present in a language different from the language of the user's query (Nie, 2010). A typical CLIR scenario is shown in the Figure 1.2, where a user needs to retrieve documents from different languages using query in English.

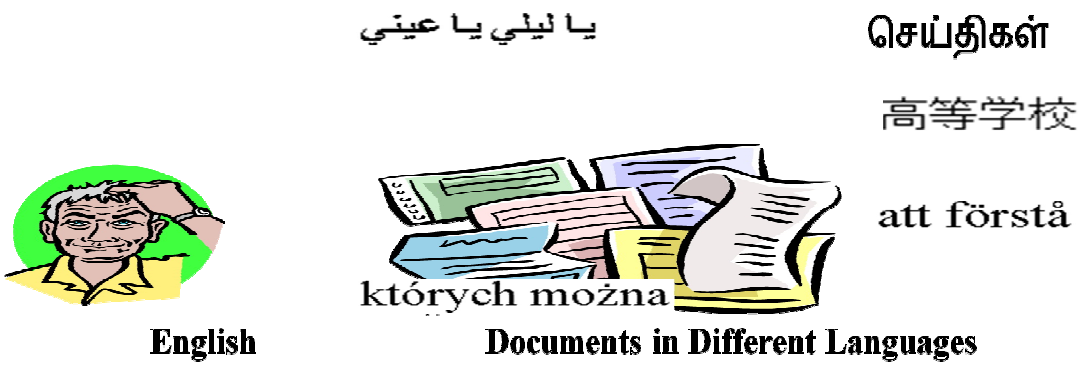

Figure 1.2: CLIR Scenario

These systems allow user to supply search queries in the form of text or speech in one's native language, which are then translated and used to retrieve relevant documents in other languages. In general, these systems attempts to solve term ambiguity between two different languages. This field is considered as a multi-disciplinary field that combines techniques from IR, Natural Language Processing (NLP), Machine Translation (MT), Speech Processing and human-computer interaction.

During Cross-Language search and retrieval, the CLIR system integrates MT and text retrieval technologies to provide the full function of finding information in languages different from users' queries. The basic processes involved in such a system are shown in Figure 1.3.

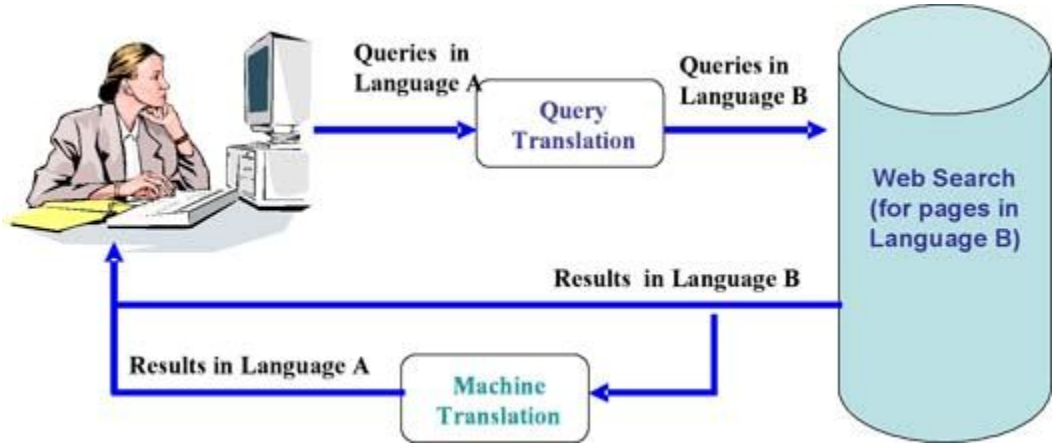

Figure 1.3 : Processes in CLIR System

In this research work, a user submits a query in one language and, the system after several steps, finds out the relevant documents written in a different language (CLIR process). As mentioned in the previous section, the basic strategy for information retrieval is to match documents to queries. A transformation on query side or document side or both is necessary if the queries and documents are not written in the same language, as in the case of CLIR, since the match cannot be directly conducted.

CLIR involves researchers from various fields like information retrieval, natural language processing, machine translation and summarization, speech processing, document image understanding and human-computer interaction. Due to increased global internet user population, the CLIR research is becoming more and more important for global information exchange and knowledge sharing. Examples include the following:

- Foreign Patent Information Access.
- Medical Information Access for Patients.
- Searching a monolingual collection in a language that the user cannot read.
- Retrieving information from a multilingual collection using a query in a single language.
- Selecting images from a collection indexed with free text captions in an unfamiliar language.
- Locating documents in a multilingual collection of scanned page images.

1.3. TYPES OF CLIR SYSTEMS

The CLIR systems can be broadly classified into three types, namely, monolingual system, bilingual systems and multilingual systems (Reddy and Hanumanthappa, 2013). Monolingual CLIR system is used to find relevant documents in the same language as the query was expressed. On the other hand, bi-lingual Systems allow querying in one language and finding documents in another language. Multilingual CLIR systems accepts query in one language and is capable of finding documents in multiple languages. Oard and Diekema (1999) identified three basic transformation approaches to CLIR, namely, query translation, document translation and inter-lingual techniques.

In Query Translation approaches, the query, given in one language (source language), is first converted to another language (target language), which is then used to search a document corpus having relevant documents in the target language. In other words, the query is translated into the language of the document.

In Document Translation approaches, the entire document corpus is translated from target to source language, and related documents are retrieved when given a query in second language. In other words, this approach translates the documents in foreign languages to the query language. As there are too many documents to be translated and each document is quite large as compared to a query, this approach suffers from scalability problem and hence is practically unsuitable.

The Inter-lingual Approach translates both documents and queries to a third representation. This approach generally requires huge resources, as the translation needs to be done online.

1.4. IMPORTANCE OF CLIR IN TAMIL-ENGLISH TEXT RETRIEVAL

According to Vanopstal *et al.* (2010), around 82% of information provided in WWW is in English and this statistics is increasing in a day-to-day fashion. Similarly, along with this growth, the number of languages used on the web is also diversifying. Non-English speakers are the fastest growing group of new web users and there is a growing interest in non-English sites as

the web becomes truly multi-lingual. According to Global Internet Statistics (2004), over 64% of the global web users are non-English speakers.

In India, for example, the number of Internet users has crossed the 300 million mark by December 2014 and is expected to reach 500 million users before end of 2016 (http://trak.in/tags/business/2014/11/19/india-300m-internet-users-2014). Moreover, the Global Reach statistics also shows that nearly 90% of the web users prefer to access the Internet in their native languages (http://global-reach.biz/globstats/index.php3). However, most of the users have good reading skills in their native language (large passive vocabulary) but have poor language productive skills (limited vocabulary) in another language (mostly English) and thus cannot express their information need in non-native language (Ogden *et al.*, 1999; Abdelali *et al.*, 2004). To solve this issue, the CLIR systems are used.

CLIR has the potential to allow people to find documents, irrespective of the language used in query or document. It is of growing importance because it can open up a whole world of information for the user, especially with the ease and convenience of access and delivery of foreign documents provided by the web.

The language barrier has become a major restriction on global information exchange and knowledge sharing and its impact is more significant in countries whose language is non-alphabetical, such as Tamil, Kannada, Hindi and Telugu. For example, most Tamil speaking web users have some knowledge of English, but due to limited English vocabulary, they may find it difficult to formulate effective English queries. Thus, the ability to retrieve relevant English documents using Tamil Language queries should be considered a necessary part of information access for Indian users, who desire to identify or monitor developments around the world. Moreover, there may be some instances where the user wants to retrieve documents in foreign languages, especially for instances where information available in languages other than the user's native language is more plentiful and detailed.

The importance of CLIR in the area of Tamil to English document retrieval has increased mainly due to the fact that most of the contents available online are written mainly in English. However, several users in WWW want to search for information using their own mother tongue. But while doing so in their own mother tongue, the search engines retrieve very little or

sometimes no documents. For example, the query 'குளிர் காய்ச்சல்' retrieved 7,120 documents, while the English query 'cold fever' retrieved 6,46,00,000 documents. As another example, the query "சிறப்பு குழந்தைகள் மருத்துவமனை" retrieved only 2 documents while its corresponding English query 'Children special care hospital' retrieved 10,90,00,000 documents. More scientifically, the query "பட செயலாக்க" retrieved only 275 documents, while its English version "image processing" retrieved 2,78,00,000 documents. These examples show that although there is a lot of useful information on the Internet, it may not be accessible while using Tamil queries.

The web phenomenon, coupled with increasing globalization of corporations and organizations, has led to a strong demand for tools that permit the users to find information regardless of language boundaries. The demand had also been stimulated by more corporations competing for a place in the world-wide marketplace driven by foreign language information. These factors have given rise to greater interest in Tamil to English CLIR.

1.5. CHALLENGES IN CLIR WITH TAMIL LANGUAGE

The major challenge faced by the CLIR system is the voluminous amount of multi-language information available, which is mostly unstructured and heterogeneous in nature. Moreover, the CLIR systems should also have to handle issues related to redundant data during document retrieval.

An effective CLIR system, for Tamil language, requires Tamil language focused algorithms to identify multiple characters and font encoding schemes. In today's environment, news articles in Tamil are published on web, using proprietary font encoding by the publishers. Most of them are non-standard and used by specific groups for individual purposes. To perform CLIR on these documents, the system has to decode such non-standard encoding and convert them into a standard form, so that it can be used for further processing.

Tamil is a highly agglutinative language, and analysis of such language is also a vital step in CLIR systems. These tools should include tokenizers, lemmatizers or stemmers, Part-of-Speech tagger, morphological analyzer and named entity recognizers, which are mandatory for normalizing and tagging the text before indexing.

The performance of the CLIR system is primarily dependent on the correctness of query translation. Query translation faces many challenges related to ambiguity, named entities, dictionary availability and machine translation technique used.

While using CLIR with Tamil language, usage of relevance feedback to improve the quality of retrieval might suffer; because, this process assumes that the user can rapidly make relevance judgements about retrieved documents, with little information like title and first few sentences of the document. Moreover, this information is provided in English language, which might lead inaccurate relevance feedback.

In summary, the following challenges are faced by CLIR systems with Tamil Language:-

- **Translation ambiguity:** While translating from Tamil language to English language, more than one translation may be possible. Selecting appropriate translation is a challenge. For example, the word “malai” has two meanings, viz., ‘garland’ and ‘evening’.

- **Phrase identification and translation:** Indentifying phrases in limited context and translating them as a whole entity rather than individual word translation is difficult.

- **Translate/Transliterate a term:** There are ambiguous names which need to be transliterated instead of translation. For example, sooriyan (Sun) in Tamil refers to a person’s name as well as the sun. Detecting these cases based on available context is a challenge.

- **Transliteration errors:** Errors during transliteration might end up fetching the wrong word in target language.

- **Dictionary coverage:** For translations using bi-lingual dictionary, the exhaustiveness of the dictionary is an important criteria for performance on system.

- **Font:** Many documents on web are not in Unicode format. These documents need to be converted in Unicode format for further processing and storage.

- **Morphological analysis:** The algorithm used is different for different languages, where as Tamil language is inflectional in nature, and the algorithm has to be specifically designed for this language.

- **Out-of-Vocabulary (OOV) problems:** Problems faced with words that do not tend to occur in the monolingual dictionary or thesaurus of a particular language. These new words get added to language which may not be recognized by the system.

Among the different challenges, the major factors which influence the performance of CLIR systems are given in detail below:

- **Limited size of Dictionary:** The limited size of dictionary contributes to translation errors. New words get added to the language quite frequently and maintaining the dictionary up-to-date with these new words is difficult. Also, compounds and phrases can be formed from existing words in the language. No dictionary can contain all possible compounds and phrases. A specific domain can generate a specific terminology which might not be present in a general dictionary. Inflected word forms are not included in a dictionary. Thus, normalization process like stemming becomes essential.

- **Query translation/transliteration performance**: The phenomenon of translation ambiguity is common in cross-lingual information retrieval and refers to increase of irrelevant search key senses due to lexical ambiguity in source and target languages. A search key may have more than one sense in source language, which may be expressed by the dictionary, by providing several alternatives. During the translation process, extraneous senses may be added to the query due to the fact that the translation alternatives may also have more than one sense. Thus lexical ambiguity appears in both source and target language.

1.6. OVERVIEW OF TECHNIQUES USED IN PROPOSED CLIR SYSTEM

The proposed CLIR system has three main components, namely, Speech Query and Recognition, Query Translation and Document Retrieval. This section presents an overview of these three components.

1.6.1. Speech Query and Recognition

The speech query component of the CLIR system accepts the query in the form of Tamil speech data and performs speech recognition to convert the voice data into text. Automatic speech recognition that performs Tamil speech recognition is a critical and challenging

component. It is a vital part of human-to-computer interaction system and is defined as the art of translating spoken words to machine readable form. It is one of the most popular and fast growing technology, as speech enables communication in a natural and effective manner. Research in Automatic Speech Recognition (ASR) has attracted a great deal of attention over the past few decades, and is still considered as a research area where accuracy of recognition remains as one of the most important research challenges (Chandra and Sujiya, 2014). Presently, speech recognition is achieved for many languages around the world, and now-a-days several researches have focused on developing ASR for Indian Languages.

The basic steps involved in an ASR system are given in Figure 1.4. In this research work, the user presents the speech data using microphone, which captures sound waves and generates electrical impulses. The sound card converts acoustic signal to digital signal. Then, the recognizers or speech recognition engine converts digital signal to phonemes and then to words. The recognizers are the software drivers that convert the acoustical signal to a digital signal and deliver recognized speech as text to the application.

ASR systems can be grouped into four different classes depending upon the usage mode. They are, isolated word recognition, connected word recognition, continuous word recognition and spontaneous speech recognition. Isolated word speech recognizers require each utterance to have silence (lack of an audio signal) on both sides of the sample window. It requires a single utterance at a time, and often these systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances. It is also referred as Isolated Utterance recognition. On the other hand, connected word speech recognition systems are similar to isolated word recognition, but allow separate utterances to be 'run-together' with a minimal pause between them.

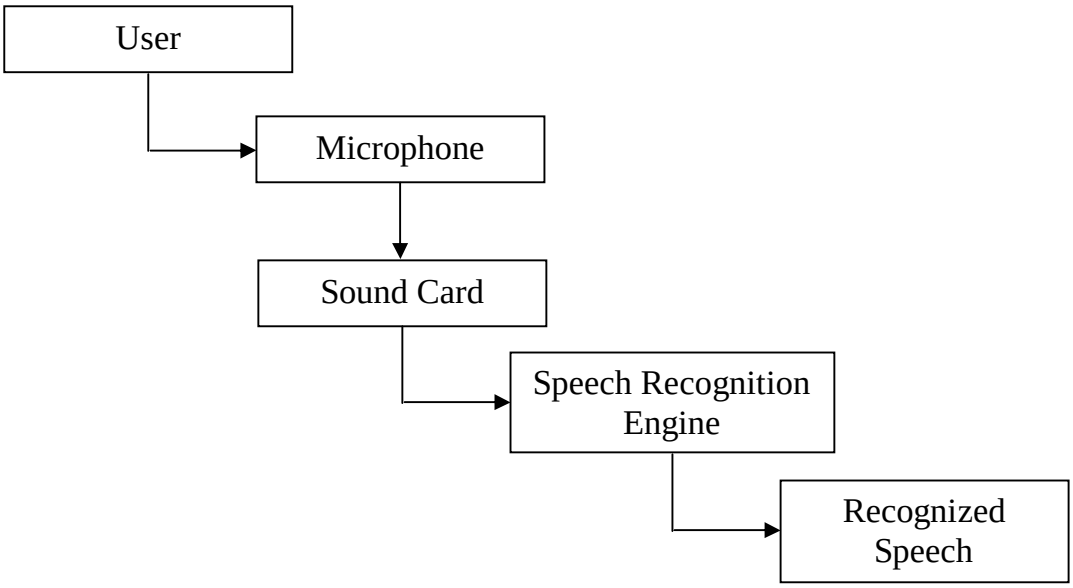

Figure 1.4: ASR System

In continuous speech recognition, users are allowed to speak almost naturally, while the computer determines the content. Spontaneous speech recognition is to recognition speech signal from its natural sounding and not rehearsed. An ASR system, with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

Speech recognition can also be speaker-dependant (in which case the templates have to be changed every time speaker changes) or speaker-independent (recognizes speech irrelevant of the speaker). The speaker- independent systems are more commercially attractive than dependent ones but are hard to implement. A speaker-dependent system is a system that must be trained on a specific speaker, asked to record predefined words or sentences that will be analyzed and whose analysis results will be stored. On the contrary, speaker-independent systems can be used by any speaker without any training procedure. Those systems are thus used in applications where it is not possible to have a training stage.

Depending on the algorithm used, speech recognition can also be text-dependent and text-independent (Reynolds and Rose, 1995). In the former case, the utterance presented to the recognizer is known beforehand. In the latter case, no assumption about the text being spoken is made. They can further be divided into two branches, namely, open set recognition and closed

set recognition. If the speech signal is presented only by the registered speakers, the recognition task is a closed-set problem; else the task is called an open-set problem. This research work proposes a speaker independent, text independent, and open-set ASR system that works with isolated words.

1.6.2. Query Translation

The query translation is the generally accepted approach that has been applied by most CLIR systems because of its simplicity and effectiveness (Zhou *et al.*, 2012; Sokolov *et al.*, 2014). Query translation for CLIR is one of the most widely used techniques to access documents in a different language from the query. Query translation based CLIR systems use various knowledge resources, such as bilingual dictionaries, Machine Translation systems, parallel texts or a combination of these, to translate queries from one language into another language, and then conduct monolingual search to retrieve relevant documents. This approach uses several algorithms like parsing, segmentation, indexing using morphological analyzer, Part of Speech Tagging, stemming and stop word removal.

This research work is focused on the first approach (query translation) where the query is given as a speech signal in Tamil (native language), which is recognized and translated into English. This translated query is then used to retrieve documents from English document corpuses.

The query and document can be translated using controlled vocabulary or free text-based method. The controlled vocabulary is the traditional technique widely used in libraries and documentation centres. In this process, documents are indexed manually using fixed terms which are also used for queries. However, this approach remains limited to application whose vocabulary is still manageable. The efficiency and effectiveness degrade radically when size of vocabulary increases. The alternate approach to controlled vocabulary is to use the words which appear in the documents themselves as the vocabulary, and such systems are referred as free text retrieval systems.

In the free text-based method, the techniques can be classified as machine translation, dictionary (or knowledge based), ontology-based and corpus-based CLIR systems (https://en.wikipedia.org/wiki/Cross-language_information_ retrieval).

Machine translation is another linguistic and knowledge-based approach available for query translation. Machine Translation is the process that utilizes computer software to convert free text from one language to another; the output seeks to be accurate and fluent for human consumption. An MT system can be used to translate the query, the document or both into the same language, and the retrieval process could then be treated with a general IR system. However, machine translation systems are able to produce high quality translations only in limited domains (Oard and Dorr, 1996). They need information about context and are based on syntactic analysis. Syntactic analysis is not possible for the translation of bag-of-word queries, lacking grammatical structure. However, machine translation has been used as a method in several research reports on cross-language retrieval (Ture *et al.*, 2014; Goto *et al.*, 2015).

The dictionary based approach (Lehtokangas *et al.*, 2006; Kumar and Das, 2013) uses a lexical resource to translate words from source language to target document language. The lexical resources are based on knowledge structures, which are in the form of multi or bi-lingual dictionary or thesaurus applied for free text retrieval or in the form of sophisticated ontology. This translation can be done at word level or phrase level. The main assumption in this approach is that user can read and understand documents in target language. In case the user is not conversant with the target language, he/she needs to use some external tools to translate the document in foreign language to their native language. Such tools need not be available for all language pairs. Dictionaries are used to translate each word of the source language query to the desired target language. In the translation process, words can be translated by not one unique term but a set of terms appearing as equivalent translations in the dictionary. This approach offers a relatively cheap and easily applicable solution for large-scale document collections. The major problems of dictionary based approach are translation ambiguity, out-of-vocabulary terms, word inflection and phrase identification.

A corpus is a repository of a collection of natural language material. It analyzes large collections of existing texts (corpora) and automatically extracts the information needed, on

which the translation will be based. The corpus-based approaches start from text analysis. Document text collections in different languages form the text corpora are needed for this approach. The aim is to extract the information needed for the translation from the existing texts. The text collections can include exactly the same texts in several languages (parallel corpora), or the texts can include documents belonging to the same subject category (comparable corpora) (Oard, 1997). Relevant documents in the source language are retrieved and words are extracted from parallel or related documents in the target language. Corpora-based systems are hard to maintain and tends to be domain/application dependent to achieve effective performance.

Ontology defines the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary. Ontology is an explicit specification of a conceptualization Ontology can be implemented in translation systems to extract conceptual relations for monolingual and CLIR (Saraswathi *et al.*, 2010).

A summary of the different query translation techniques is presented in Figure 1.5.

1.6.3. Document Retrieval

Document retrieval is a process of finding relevant documents against user translated queries. Initially, document is preprocessed using various techniques such as tokenization, stop word removal, handling morphological variants, case folding, normalization and bag of words. The three main steps involved in the design of a document retrieval system are query expansion, text classification and ranking. Query expansion is an approach to boost the performance of document retrieval. It consists of expanding a query with the addition of terms that are semantically correlated with the original terms of the query. The text classification algorithm uses machine learning classifiers to identify the category which is close or similar to the user query and retrieves all the documents in that category as retrieved documents. In order to improve the retrieval process, the second step called ranking is used, to arrange the retrieved documents based on its relevance to the query words.

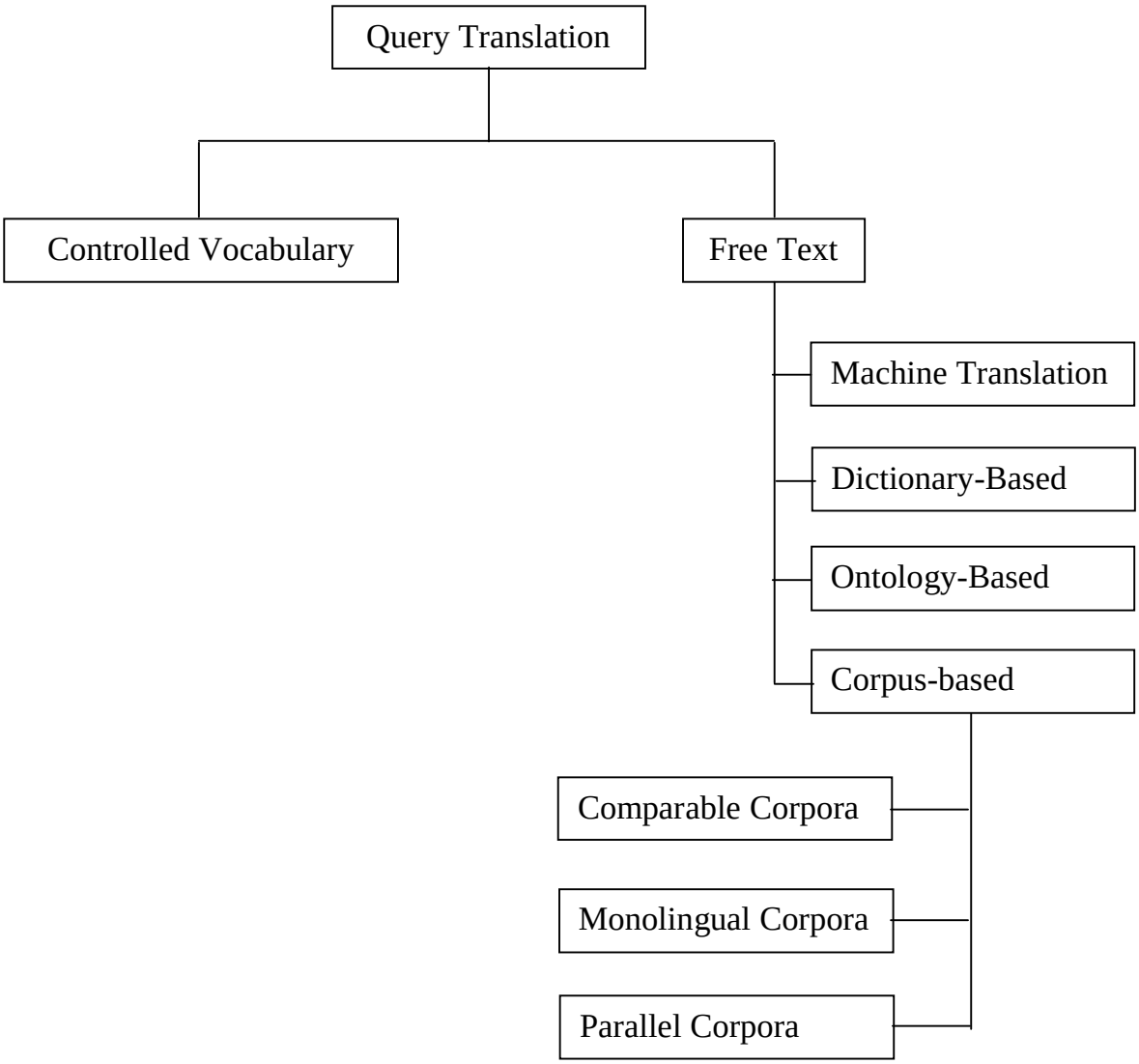

Figure 1.5: Query Translation Techniques

1.7. MOTIVATION AND OBJECTIVES

The development of information repositories is creating many opportunities and also new challenges in information retrieval. The availability of online documents in many languages makes it possible for users around the world to directly access previously unimagined sources of information. However, in conventional information retrieval systems, the user must enter a search query in the language of the documents in order to retrieve it. This requires that users can express their queries in those languages in which the information is available and can understand the documents returned by the retrieval process. This restriction clearly limits the amount and type of information that an individual user really has access to.

Moreover, most of the existing systems perform a search for the solution in a standard set of languages. For example, Jagarlamudi and Kumaran (2008) searches for the solution only in English documents. This does not turn out to be successful in all cases, because, the domain of the query may be related to a particular region where a particular language is spoken. Thus, information systems, such as digital libraries or corporate information management systems, would better serve their users if language support services were integrated into their systems.

India is a multilingual country where the spoken language changes after every 50 miles. There are 22 official languages and approximately 2000 dialects spoken by different communities in India. English and Hindi are used for official work in most states of India. The Tamil Nadu state government in India predominantly carries out their official work in their respective regional language. Many newspapers are also published in Tamil languages. Examples include Thina Thanthi, Thina Malar and Malai Malar. Sarching and translating these documents manually is very time consuming and costly. Hence there is need to develop good translation and information retrieval systems to address all these issues, in order to establish a better communication between states and Union governments and exchange of information amongst the people of different states with different regional languages.

Currently, several works have been carried out in CLTR, but only a limited have focused on Tamil-English CLTR, and very few of these have focused on retrieving English documents using Tamil speech queries. Developing a complete and well specified CLTR models for any language with limited electronic resources is always a challenging and demanding task, where several issues involving speech recognition accuracy, translation accuracy and retrieval accuracy are still in the research stage. The performance of the CLTR system depends on the individual performance of each of these steps, and this research work proposes algorithms that aim to improve the working of each of these steps, so as to increase the overall performance of CLTR.

The research problem is given as follows:

> "To design and develop Tamil-English CLTR systems based on text or speech query by integrating key technologies like speech analysis, translation, document retrieval and ranking along with ensemble machine learning for accurate relevant English document retrieval."

To design and develop such a system the following specific objectives were formulated.

- To design and develop a Tamil Speech Query Recognition system that uses wavelet-based methods for noise removal and enhanced feature extraction, optimized through the use of Self-Organizing Map and ensemble Support Vector Machine (SVM) classifier for speech recognition.
- To design and develop optimizing methods to convert the recognized Tamil text into English text.
- To design and develop Heterogeneous Ensemble Model that uses hybrid associative and improved KNN algorithm for document retrieval.

1.8. LAYOUT OF THE CHAPTERS

The underlying objective of this research work is to develop an effective and efficient CLIR system for Tamil and English Languages. This chapter (Chapter 1, Introduction) presented the introductory materials covering the various concepts related to information retrieval, CLIR and speech recognition. The rest of the thesis is organized as follows.

The literature review is a critical look at the existing research that is significant to the work carried out. Several researchers have addressed the problem of speech recognition, CLIR, translation and text retrieval. A critical look at the various available literatures related to the present research work is given in **Chapter 2, Review of Literature.**

Chapter 3, Methodology, presents the research methodology and identifies the different steps of the proposed CLIR systems. The various methods and techniques used are briefly explained in this chapter. **Chapter 4, Design of Tamil Speech Query Recognition System,** presents a detailed description of the techniques used for Tamil speech Recognition.

Chapter 5, Design of Query Translation System, presents the various techniques used to convert the recognized Tamil words to its corresponding English Words. **Chapter 6, Design of Text Retrieval System**, presents details regarding the techniques used for improving the process of document retrieval.

Chapter 7, Results and Discussion, tabulates and discusses the various results obtained while testing the proposed algorithms. The findings of the study are summarized along with future research directions in **Chapter 8, Summary and Conclusion**.

The work of several researchers are quoted and used as evidence to support the concepts explained in this thesis. All such evidences used are listed in the **Bibliography** of the thesis.

1.9. CHAPTER SUMMARY

This chapter provided an overview of information retrieval with particular emphasis on Cross-Language text retrieval. The chapter also presented the formulated research objectives. Various researchers have contributed to the improvement of these technologies and the studies related to the present work are reviewed in the next chapter, **Review of Literature**.