CHAPTER III

METHODOLOGY

Information retrieval is the process of finding any kind of relevant information from WWW in an easily accessible and understandable form. However, due to globalization, content storage and retrieval must be possible in all languages. Diversity of languages is becoming a great barrier to understand and enjoy the benefits in digital world. These benefits can be attained through the use of Cross-language Information Retrieval system.

In this research work, in order to meet the objectives outlined in Chapter 1 (Introduction), an amalgamation of tools are used and enhanced to improve the overall performance of CLIR. The main scope of this research work is to retrieve English text documents, using either a text-based query or a speech-based query, obtained in Tamil language. This chapter introduces the research methodology, along with the main steps involved in the proposed CLIR system. The techniques enhanced in each step of the proposed system are also presented. Detailed descriptions of these techniques are provided in the subsequent chapters of this thesis.

3.1. PROPOSED METHODOLOGY

As mentioned in Chapter I, Introduction, the primary focal point of this research work is to develop Tamil to English Cross-language text retrieval system, using two types of queries, namely, text and speech. These systems differ in the manner of accepting the query input and are termed as TECLTR-T (Tamil-English Cross-Language Text Retrieval for Text query) and TECLTR-S (Tamil-English Cross-Language Text Retrieval for Speech query) in this research work respectively.

The steps involved in the design and development of the proposed TECLTR-T system that uses a text-based query in Tamil to retrieve related documents in English are listed below.

(i) Query Translation and Transliteration
(ii) Query Expansion
(iii) Document Retrieval
(iv) Document Ranking

The TECLTR-S is speech driven text retrieval system that extends TECLTR-T to include a speech module that performs two extra steps. The two steps are,

(i) Acquiring Tamil Speech Query

(ii) Speech Recognition

The speech module of TECLTR-S, thus integrates two powerful technologies, namely, Speech Recognition and CLIR. It first accepts a pronounced query, identifies the actual speech data and then recognizes them into textual form. The output of the Speech interface is then used by TECLTR-T to retrieve documents. Except for the inclusion of speech module, the rest of the steps are the same as that of TECLTR-T. Thus, the TECLTR-S allows an average user to search for information without the need to learn special computer skills or training. The system allows the user to use common devices that they are familiar with to view their documents.

Figure 3.1 presents the steps involved in the proposed TECLTR-T and TECLTR-S. Initially, a user supplies search query in the form of Tamil spoken signal, which is recognized using a speech recognizer. The recognized Tamil words are then translated into English query, which is then enhanced or reformulated using query expansion technique. These words are then used to search English document corpora to retrieve relevant documents, which are ranked and displayed according to the relevance to query words. In this research work, to increase the performance of each of these steps, techniques based on ensemble machine learning models are applied and studied.

Ensemble classification, a method used to improve the classification accuracy, uses multiple classifiers and allows solutions that would be difficult to reach with a system that uses single model for classification / clustering (Bramer, 2013). The usage of ensemble models in various fields like applied statistics (Hastie *et al.*, 2001), machine learning (Dietterich, 2000) and pattern recognition (Kuncheva, 2004) has grown tremendously during the past few decades. Several studies have proved the advantages of using ensemble classifiers over its traditional single classifier systems (Bauer and Kohavi, 1999).

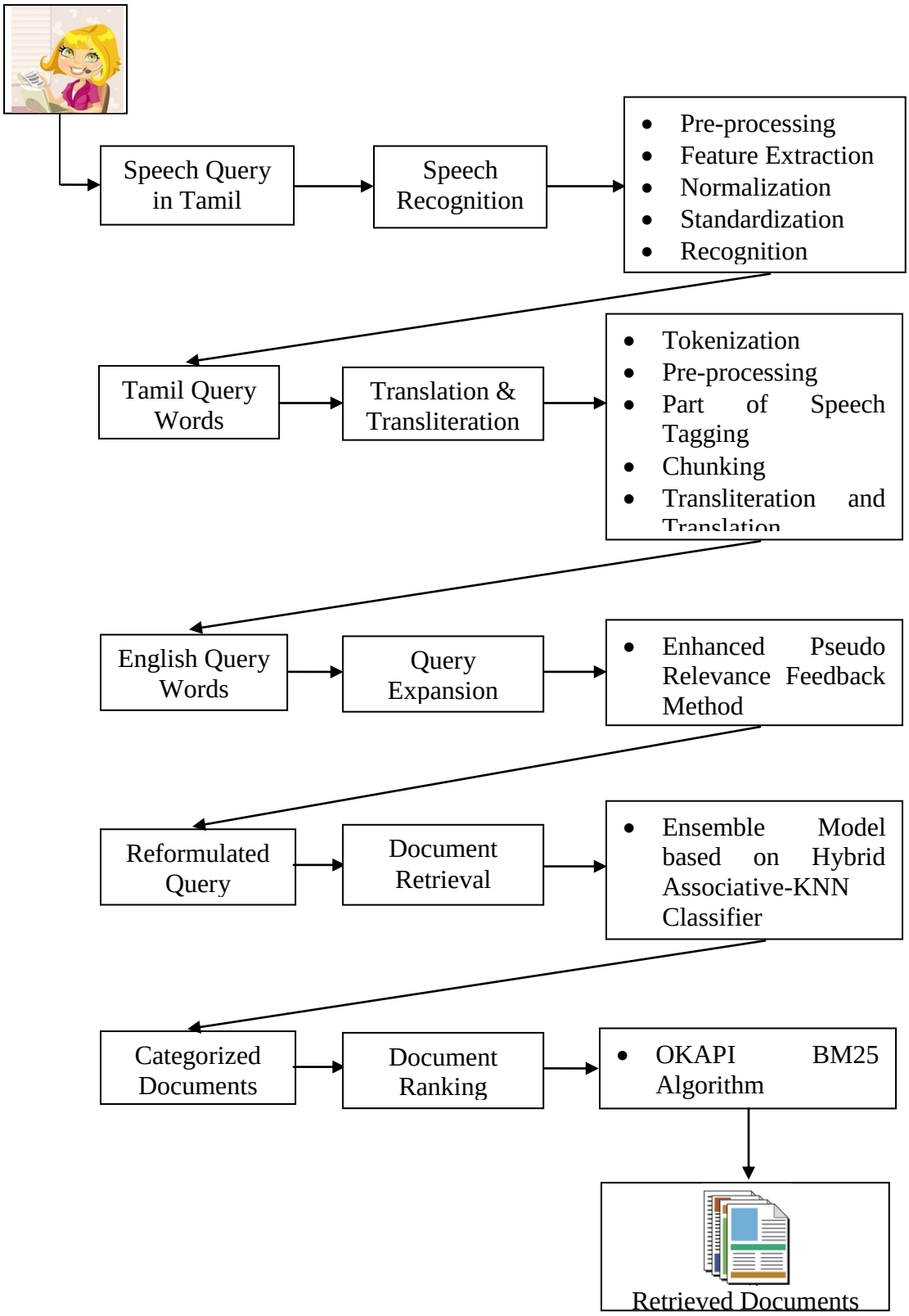

Figure 3.1: Steps in Proposed CLTR System

In this research work, each of the steps presented in Figure 3.1 is designed to find optimized solutions to meet the research objectives and the steps are grouped into three phases (Figure 3.2) and a short description of the same is provided in the following sections.

Phase I : Speech Recognition

Phase II : Query Translation

Phase III : Text Retrieval

3.2. PHASE I: SPEECH RECOGNITION

In this research work, ASR is used as an interface between user and CLTR system, and is used to convert the query supplied as Tamil speech signal to text. In general, the existing ASR systems, after acquisition of speech signal, perform two important tasks, namely, feature extraction and classification. Feature extraction is one of the key dimensions of ASR systems, and the most widely adopted approaches are MFCC and LPCC. After feature extraction, several of the recent proposals use machine learning classifiers to recognize the words (Hinton *et al.*, 2012; Deng and Li, 2013).

Daqrouq *et al.* (2011) proposed a speech recognition system that used Discrete Wavelet Transformation (DWT) based LPCC features with Probabilistic Neural Network for Arabic speech recognition. Various aspects of this system are modified in this research work to improve the process of speech recognition. The algorithm is enhanced by using pre-processing, enhanced features and ensemble classification model.

The pre-processing step performs two tasks, namely, noise removal and silence removal. In CLTR systems, the quality of speech query signal is very important for successful and accurate translation. In the first step of preprocessing, a wavelet-based noise removal algorithm that uses a combined, soft and hard thresholding method is proposed to improve the quality of speech.

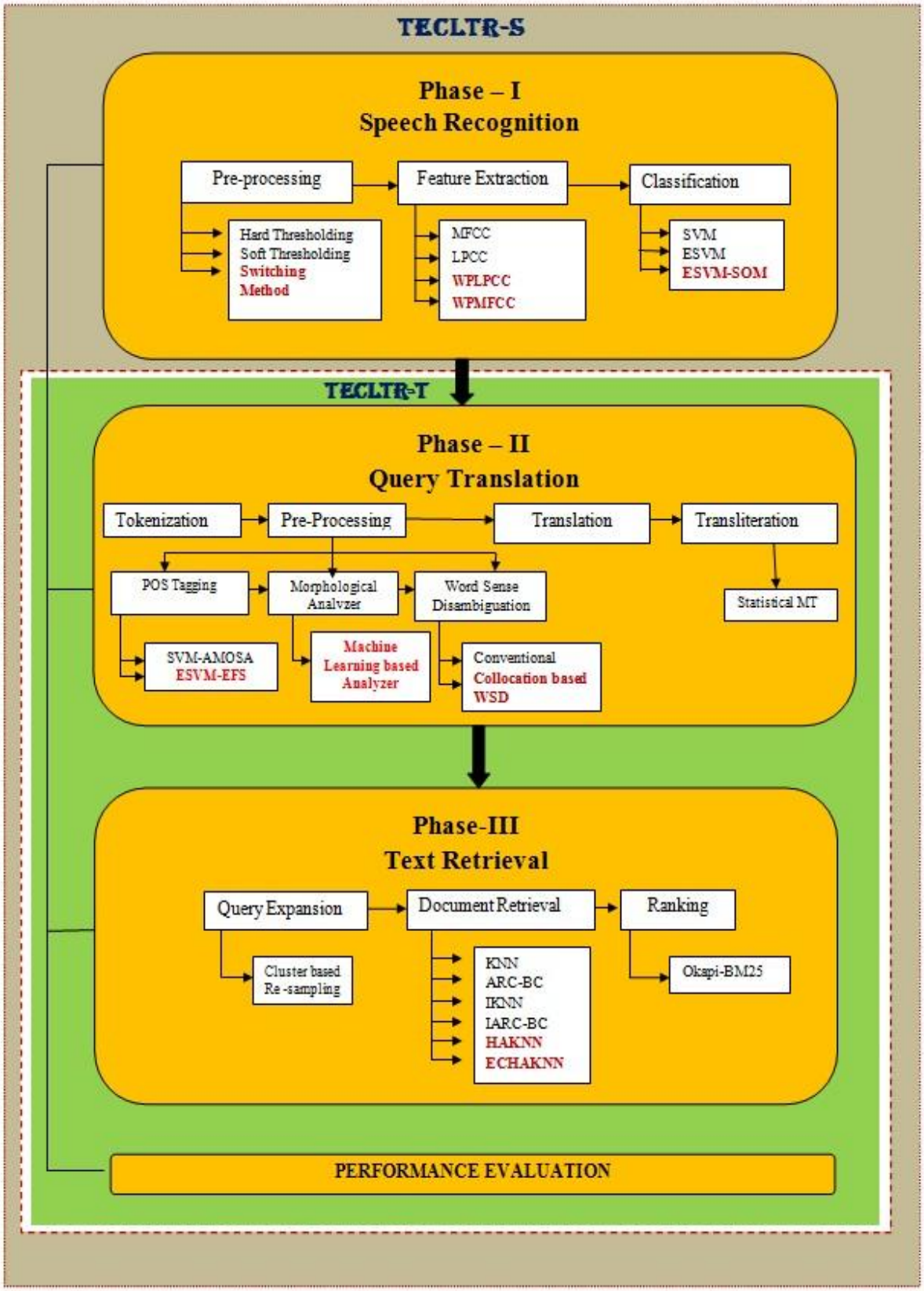

Figure 3.2: Research Methodology

The proposed algorithm performs D4 wavelet decomposition to obtain coefficients, which are ordered in increasing frequency. The Median Absolute Deviation (MAD) is estimated on the largest coefficient. In this research work, a Switching Thresholding Method, that combines the advantages of both soft and hard thresholding, is used during denoising. All coefficients below this threshold are considered as noise and are removed. Finally, an Inverse DWT (IDWT) is applied to obtain a quality enhanced signal.

The silence removal algorithm aims to identify the presence of speech signal. Two criteria were used to identify the presence of a spoken word. Initially, the total energy is measured, from which the numbers of zero crossings are counted. Both of these are necessary for speech signal identification, as voiced sounds tend to have a high volume (and thus a high total energy), but a low overall frequency (and thus a low number of zero crossings), while unvoiced sounds were found to have a high frequency, but a low volume. It was also discovered that only background noise have both low energy and low frequency. Experiments proved that this method can successfully detect the beginning and end of the several speech queries tested.

As mentioned earlier, both LPCC and MFCC acoustic features are used extensively during speech recognition. In this research work, during the extraction of LPCC and MFCC, a Discrete Wavelet Packet Transformation (DWPT) based method was used as an alternative method to perform filter-bank energy-based feature extraction. The rich coverage of time-frequency properties of DWPT, can obtain more discriminative features that can improve the performance of speech recognition, when compared to the conventional usage of MFCC and LPCC. These features are then normalized to have zero mean and unit variance. Further, the research work also proposes the fused DWPT-based LPCC and MFCC features to improve the speech recognition accuracy. In the next step, the fused variable sized vector is converted into constant size using Self Organizing Map (SOM). This step is included again to improve the recognition performance.

Finally, in the last step of Phase I, the fusion feature vector was used to train and test the homogeneous ensemble classifier for speech recognition. The ensemble classifier is created using 25 different versions of Support Vector Machine (SVM) classifier, created using Bagging subspace sampling method (Schapire *et al.,* 1998). Each classifier in the ensemble model is

trained separately in an independent fashion. Majority voting scheme is used as the aggregation algorithm.

Detailed description of the above methodology used for Tamil Speech Recognition is presented in Chapter 4, Design of Tamil Speech Query Recognition System and the results of performance evaluation are presented in Chapter 7, Results and Discussion.

3.3. PHASE II: QUERY TRANSLATION

The second phase of the research work is query translation and transliteration, which is an area of research work that focuses on converting the recognized text in Tamil language into English language. It is a process that takes a character string in a Source Language (Tamil) and generates equivalent mapped character string in the Target Language (English).

The proposed hybrid machine translation system is a combination of Rule-based Machine Translation and statistical machine translation approaches. Rule-based MT consists of several steps such as tokenization, pre-processing, and translation. The transliteration mining and error correction are done using statistical MT.

The first step in rule-based MT is tokenization, which is the process of breaking up the given text into units called tokens. This research work uses blank space as word separator. In general, the pre-processing step consists of procedures which reduce the complexity of the translation process and increase the accuracy of translator. The pre-processing consists of three steps namely POS tagging and chunking, morphological logical analyzer and named entity recognition, Word Sense Disambiguation (WSD). The next step after pre-processing is translation and transliteration.

After tokenization the first step in processing any language sentence is POS tagging, and is defined as a process of assigning the part-of-speech label to words in a given text. The Tamil language has a very rich morphological structure which is agglutinative, and Tamil words are made up of lexical roots followed by one or more affixes. This makes tagging a Tamil word complex.

Ekbal and Saha (2013) proposed a system that used 11 features with SVM based ensemble method along with an enhanced Simulated Annealing (SA) based Majority Voting

Algorithm, Multi-Objective Optimization (MOO) Algorithm – AMOSA (Archived Multi-Objective Simulated Annealing) and an objective function to increase the accuracies of all the individual POS classes for tagging Bengali and Hindi language words. The accuracy is reduced by more than 12.6% when applied to Tamil language.

To increase the accuracy, the present research work introduces a feature selection algorithm and enhances the ensemble classifier used during POS. An ensemble feature selection that combines three algorithms, namely, Split decision tree approach, Discriminate function approach and F-score approach, is initially used to obtain an optimal set of features. The ensemble classifier is improved by using a Wavelet Neural Network (WNN)-SVM hybrid approach. The SVM classifier is used as preprocess, to reduce the training set to a subset version, that contains more critical details in deciding classification boundary. This is accomplished by first extracting support vectors. The actual classification process starts only after this, where the set of support vectors form the training set. WNN is trained using this set, after which the test data is applied to WNN to analyze its efficiency.

Followed by POS tagging the Chunking is performed, which organizes information obtained from the previous step into familiar groupings. It is a NLP that separates and segments sentences into their sub constituents such as noun, verb and prepositional phrases. Examples of chunks include noun phrases, prepositional phrases and verb phrases. Chunking works on POS tagged text, and hence its accuracy depends upon the accuracy of POS tagger. The tool provided by http://tdil.mit.gov.in (Technology Development for Indian Languages) is used to perform chunking.

. The next step in pre-processing is Morphological Analyzer (MA) which accepts POS tagged sentences as input. The morphological analyzer consists of five components, and each component analyzes each POS tag forms. The purpose of an MA is to return root word and grammatical information about all the possible word classes for a given word. MA also includes extraction of the grammatical information including number, gender and tense information for all the tokens. As Indian languages have a rich inflectional morphology, morphological analyzer is an essential tool for such languages. Conventionally, translation step uses a rule-based approach, where every rule depends on the previous rule. Hence, if one rule fails, it will affect the quality of the entire rules that follow.

To solve this issue, this research work proposes the use of machine learning classifier. In machine learning, all the rules, including complex spelling rules, are also handled by the classification task. Machine learning approaches do not require any hand coded morphological rules and require only corpora with linguistic information. These morphological or linguistic rules are automatically extracted from the annotated corpora. Here, input is a word and output is root and inflections. Input word is denoted as 'W' root and inflections are denoted by 'R' and 'I' respectively ([W] Noun/Verb = [R] Noun/Verb + [I] Noun/Verb). The machine learning classifier used in this research work is an SVM-based ensemble classifier.

Followed by MA, the Named entity recognition is performed using the tool provided by http://tdil.mit.gov.in (Technology Development for Indian Languages).

The third step in pre-processing is Word Sense Disambiguation (WSD) algorithm, which is used to handle ambiguity problem based on collocations. Collocations are defined as nearby words, that strongly suggests the sense of the ambiguous word, in a given occurrence. WSD is an important and challenging task during translation. In general, a WSD algorithm initially uses a manual process to extract collocations. It then identifies sense-collocation words related to the identified collocation using either a dictionary or a thesaurus. This existing process is time consuming, and the manual process may introduce errors. To solve this issue, in this research work, the manual collocation extraction process is replaced using an automatic extraction procedure that uses an enhanced K-Means clustering algorithm.

The main concern of K-Means algorithm is the optimal selection of 'K' parameter, which is solved using an ensemble approach. An ensemble of clustering algorithms is build with different K values. This ensemble generates a set of clusters. Majority voting algorithm is then used to find the optimal clustering set from the different partitions created, thus estimating the optimal K value for clustering. The advantage of this approach is that the estimation of this K value is embedded during the process of clustering and requires no extra optimization procedures. The next step then uses a sense-collocation dictionary to associate collocations with sense words. The advantage of using automating the extraction step is that it can save search time while considering large number of ambiguous words in a language and reduces manual errors.

All the above pre-processing steps analyze the given input to gain maximum information regarding the words in the document sentences and query sentence. The next important step is translation; it is a task of finding translation equivalents for the Tamil words. Translation is defined as a process of changing or converting a word, sentence, text etc. to another language without changing the underlying meaning or message. Although language changes, meaning does not change. The MA returns root words and grammatical categories of a Tamil sentence. The root words obtained from MA is translated using bilingual dictionary and semantic dictionary. The remaining grammatical morphemes are translated using tense marker and gerund ending rules. The untranslatable named entities that are not translated using dictionaries are handled by transliteration process. Finally to obtain correct structure of English sentence Tamil to English re-ordering rule is applied.

Transliteration is the process of representing letters, and words of one language in the corresponding characters (or the closest of such characters) of another language. Although language changes, pronunciation as well as alphabetical arrangement more or less remains unchanged. Transliteration is first performed to convert named entities and numbers. Named entities include the identification of names of people, location and companies / organizations, while numbers may refer to time/date stamp and amount references. This step is performed using a character transformation procedure which converts each Tamil character to its English equivalent. During the first pass retrieval, a Tamil-English Character Mapping Table (http://www.azhagi.com/az-tamil-modern.html) is used. The second pass retrieval uses statistical transliteration model for handling named entities, to correct the errors occurred during first pass.

Detailed description of the methodology used for the translation of Tamil Query to English Query is presented in Chapter 5, Design of Query Translation System. The experimental results that evaluate the performance of the proposed algorithms are presented in Chapter 7, Results and Discussion.

3.4. PHASE III: TEXT RETRIEVAL

Phase III of the study is involved in techniques that focus on improving the actual retrieval of the documents. The first method is Query Expansion, which is defined as the task of reformulating the translated query from Phase II, by selecting or adding terms to the query, using

information obtained from the analysis of the returned documents. The main goal here is to minimize query-document mismatch and to maximize retrieval performance. Inclusion of query expansion in CLTR, in general, can improve the retrieval performance by 4% to15% (Adriani, 2002). In this research, the method proposed by Lee and Croft (2013) is used for query expansion.

The second step is document retrieval, which is a task that matches the expanded query against the document corpora and produces a set of documents that are relevant to the query. For this purpose, this research work proposes a hybrid Ensemble classifier, which consists of two steps, namely, design of the hybrid classifier and design of the ensemble model.

The hybrid model uses two classifiers, namely, K-Nearest Neighbour (KNN) classifier and associative rules (Mangai *et al.*, 2013). In this research work, both these classifiers are first optimized to improve its performance and then combined to improve the retrieval process. The KNN classifier is a simple but effective method for text categorization. This classifier has the advantage of being non-parametric and easy to implement. However, the time complexity of this classifier increases with the corpus size. To solve this problem, the working of the KNN classifier is modified, to include a simple one-pass clustering algorithm. A clustering score is used to identify the cluster which is close to the query words, and nearest neighbor search is performed only with that cluster, thus reducing the similarity computations and time complexity.

Associative classifier combines association rule mining and classification, and is one of the most frequently used method in the field of document retrieval. This classifier performs document retrieval in two steps, namely, frequent item set generation and association rule generation, which are used to retrieve documents. In this research work, the associative classifier proposed by Antonie and Zaiane (2002) called ARC-BC (Association Rule based Categorizer By Category) is used. The main drawback of this system is the number of association rules generated by this system, which is very high. To solve this issue, the database coverage pruning technique is used.

The hybrid classification model, designed using these two enhanced classifier consists of the following three stages:-

- ✓ Stage 1: pre-processing
- ✓ Stage 2: Calculating feature weights of terms using association rules
- ✓ Stage 3: Classifying document using enhanced KNN algorithm

Stage 1 performs pre-processing which converts the high document corpora into high dimensional feature set. The optimal features are selected using Term Frequency/Inverse Document Frequency (TF/IDF) and Ward's minimum variance measure (Mangai *et al*., 2012). This step also performs discretization which converts the selected features into a n-dimensional feature vector, where n denotes the number of features. Stage 2, using minimum support, minimum confidence along with the feature vector generates association rules and calculates a weight for each feature. The resultant weighted feature vector is used as an input by the KNN classifier to predict the class which has relevant documents to the query words.

During the design of ensemble classifier, different minimum support and minimum confidence values are first used to generate different sets of associative rules. Similarly, by varying the K value of KNN classifier, a set of KNN classifiers are generated. These are then combined to form a set of hybrid classifiers, which are combined, using majority voting scheme to categorize and retrieve the documents. Finally, Okapi BM25 ranking algorithm (http://*en.wikipedia.org/wiki/Okapi_BM25*) is used to rank the retrieved documents in terms of its relevancy to query words.

Detailed description of the methodology used for the retrieval of relevant English texts is presented in Chapter 5, Design of Text Retrieval System and the experimental results that evaluate the performance of the proposed algorithms are presented in Chapter 7, Results and Discussion.

3.5. PERFORMANCE EVALUATION

Several experiments were conducted to evaluate the performance of the various techniques proposed in each phase of the research work and also the proposed CLTR system. The automatic speech recognition system proposed in Phase I of the research work was

evaluated, using the Tamil queries from FIRE dataset 2011 (http://www.isical.ac.in/~fire/data.html). Performance metrics like Signal to Noise Ratio (SNR), Mean Square Error (MSE), word level recognition accuracy and sentence level recognition accuracy, along with speed of recognition were used to analyze the efficiency of the proposed speech recognition system.

In Phase II, the operation of the three algorithms, namely, word sense disambiguation, morphological analysis and POS tagging, were enhanced to improve the process of translation and transliteration. Performance metrics like precision, recall, F-Measure and speed were used to evaluate the performance of the respective operation. Two more metrics, namely, accuracy and error rate were also used to analyze the overall performance of the translation process.

In Phase III of the research work, the retrieval performance is determined using quality metrics, namely, Precision, Recall, Mean Average Precision (MAP), F-Measure, Accuracy and Speed of Retrieval. The experimental results were compared with the conventional and existing algorithms. The effect of each technique proposed on document retrieval was also analyzed.

Experimental results proved that the enhancement algorithms proposed in each phase have positive effect on the performance of the CLTR system, indicating that the objectives of the research work have been achieved. The results also prove that the proposed CLTR system is comparable with the standard quality required by the recent applications like WWW and e-libraries.

The results of the experiments conducted to evaluate the algorithms proposed in each phase of the research work are presented in Chapter 7, Results and Discussion.

3.6. CHAPTER SUMMARY

The main objective of this research work is to design a Tamil-English CLTR system, using a speech query in Tamil to perform English document retrieval. The system consists of five major steps, namely, Tamil speech recognition to obtain the query, Tamil query translation, query expansion, document retrieval and ranking. Several techniques were proposed to enhance the operation of these steps. These enhanced techniques when applied sequentially, aim to improve the overall performance of the proposed TECLTR-T and TECLTR-S. The speech

module introduced in TECLTR-S consists of various steps like pre-processing, feature extraction and recognition. Detailed description of the proposed techniques in each of these steps is presented in the next chapter, Chapter 4, **Design of Tamil Speech Query Recognition System.**