CHAPTER 8

SUMMARY AND CONCLUSION

As international markets and rapidly expanding trans-national information networks interact, an imperative need to access information written in several languages is becoming increasingly vital. Cross Language Information Retrieval (CLIR), the detection and retrieval of relevant documents in one natural language using queries expressed in another, provides an important capability to meet this requirement. In this research work, the CLIR systems are proposed for persons who are able to read documents in foreign languages, but have difficulty in formulating foreign queries but have comfortable knowledge in formulating regional queries. The foreign language considered is English and the regional language considered is Tamil. The proposed CLIR system accepts query input in two forms, speech and text, and then uses them to retrieve English documents.

The scope of this research work is to design a Tamil-English CLIR system that uses Tamil Query in the form of speech (or text) to retrieve English documents. In order to design such a system, the CLIR is designed using three major steps, namely, speech recognition, query translation and document retrieval. Each of these steps is treated as a separate phase that is interconnected, using a simple input-output interface.

The first phase of the research work proposed ATSQR (Automatic Tamil Speech Query Recognition) System that performed Tamil speech recognition in three steps. The first step was pre-processing, which was used to improve the quality of the input speech signal. It consisted of two main algorithms, namely, noise removal and silence removal. Noise removal was performed using a wavelet-based denoising algorithm using Switching Soft and Hard Thresholding (W2SHT). The second algorithm performed Silence Removal from Denoised Speech Tamil Query (SRDSTQ) using two criteria, namely, short time energy and zero crossing rate combined with an automatically calculated threshold and rule-based procedure. The second step was feature extraction, where the extraction of the conventional MFCC and LPCC features was enhanced through the use of wavelet packet decomposition. Finally, the third step performed the recognition task to covert the speech query to Tamil text. For this purpose, an ensemble-SVM classifier created using Adaboost algorithm with SOM was proposed.

The second phase of the study performs query translation, to convert the Tamil query in text form to English query. The main steps involved in hybrid MT are tokenization, POS tagging, chunking, named entity recognizer, morphological analysis, word sense disambiguation, chunking, named entity and transliteration and machine translation. After tokenization, the research work used an algorithm that performed POS tagging in two steps. The first step extracts 11 features, from which optimal set was retrieved using an ensemble feature selection algorithm combining results from decision tree, F-Score and Linear Discriminant Algorithms. The result was then used, in the second step, by a hybrid SVM-WNN ensemble classifier to produce accurate tagging. The morphological analysis was performed, using an ensemble SVM classifier that consists of components, namely, noun/verb analyzer, proper noun analyzer, pronoun analyzer and other analyzer. The word sense disambiguation was performed using an ensemble K-Means clustering-based algorithm with sense-collocation dictionary. During translation, instead of a word-by-word approach, MT based method was used.

The final phase of the research work is concerned with the relevant document retrieval that is closely matching the query words. For this purpose, a hybrid method was proposed. This hybrid method was built by combining the advantages of improved KNN and associative rule classifier.

Using the above algorithms, two CLTR systems were designed. They are TECLTR-S and TECLTR-T. TECLTR-S was based on speech-based query and uses algorithms proposed in all phases of the research methodology. TECLTR-T was based on text-based query and uses algorithms proposed in Phase II and Phase II of the research methodology.

Several experiments were conducted to evaluate the proposed algorithms using FIRE dataset. Three types of queries, namely short term title queries, descriptive title queries and narrative title queries were used to analyze the performance of the algorithms. In Phase I of the research work, performance metrics like Signal to Noise Ratio, Mean Square Error and speed were used to evaluate the proposed noise removal and silence removal algorithms. The accuracy metric was used to evaluate the proposed speech recognition system and also to study the effect of the enhanced feature extraction algorithms on recognition. The performance of the algorithms proposed in the second phase was evaluated using metrics like precision, recall, F Measure,

accuracy and speed. Error analysis of the results was also performed. Metrics like F Measure, Macro F Measure, accuracy, MAP, R-Precision and speed were used during the analysis of algorithms proposed for document retrieval tasks (Phase III)

FINDINGS

From the experiments, the following facts were ascertained.

1. Inclusion of noise and silence removal algorithms, using hybrid thresholding algorithm, improved the speech quality and had a positive impact on the efficiency of speech recognition. Similarly, the wavelet packet based MFCC and LPCC feature extraction algorithms also proved to be advantageous while improving the performance of the Tamil speech recognizer. Further improvement was also envisaged through the use of the ensemble SVM classifier incorporated with SOM. The proposed Tamil query recognition system showed a maximum accuracy of 93.10%, indicating that the system has achieved its objective of improving the process of Tamil speech query recognition accuracy.

2. Analysis of the POS Tagging using Hybrid SVM-WNN Ensemble Classifier with Ensemble Feature Selection (POSHES) showed a positive increase with all performance metrics (precision, recall, F-Measure, accuracy) when compared with the existing algorithm (POS Tagging using SVM classifier with Simulated Annealing and AMOSA, POSSA). The proposed POSHES algorithm showed a maximum accuracy of 97.84% indicating the enhancement operations incorporated through ensemble feature selection, hybridization of SVM-WNN and ensemble technology have improved the task of POS tagging.

3. The results also showed that the usage of POS tags also increased the performance of morphological analysis. The performance analysis of the ensemble SVM-based classifier showed an increase in the process of morphological analysis with all performance metrics, namely, precision, recall, F-Measure and accuracy. A maximum accuracy of 95.75% was achieved by the proposed ESMA algorithm.

4. The proposed word sense disambiguation algorithm enhanced with clustering and sense-collocation dictionary showed an average accuracy of 95.66%, which increased the efficiency of the CWSD algorithm by 1.94%.

5. The performance of the translation process was evaluated using F-Measure and accuracy. The results proved that the inclusion of the proposed POS tagging algorithm, morphological analysis algorithm and word sense disambiguation algorithm has improved the performance of query translation. The proposed HQTPMW using Tamil speech query and Tamil text query showed a maximum of 91.83% and 93.74% accuracy was achieved with STTQ queries, 88.56% and 98.64% with DTQ and 86.12% and 90.55% with NTQ respectively. This proves that the process of query translation, while using the various enhancement algorithms, is successful and meets the objective formulated.

6. The text retrieval process using Ensemble ARC and IKNN Hybrid Classifier (EHAKNN) was evaluated using six performance metrics, namely, F-Measure, macro F Measure, accuracy, MAP, precision and R-precision. All the metrics showed an increase in performance when compared to the existing and single classifier system. The proposed EHAKNN algorithm showed a maximum accuracy of 92.33% and 93.99% while using speech-based and text-based STTQ queries respectively, 94.88% with DTQ and 96.78% with NTQ. The high accuracy obtained proved that algorithm is efficient in retrieving relevant documents using Tamil queries, thus meeting the research objective.

7. The performance analysis TECLTR-S and TECLTR-T, the combined CLTR system using all the proposed enhanced algorithms was performed using three performance metrics, namely, MAP, precision and R-precision. The results pertaining to all the three metrics showed that the performance of proposed TECLTR-S is comparatively better than the ECLTR system. In the TECLTR-T system title, descriptive and narrative queries achieve performance of 97.24%, 89.39% and 90.65% for monolingual system respectively. The performance of the TECLTR-T has outperformed the existing system with all the three types of queries.

8. While considering the time complexity of the proposed algorithms in terms of execution speed, it was evident that all the proposed algorithms produced results in slower space than its corresponding existing and conventional algorithms, which were enhanced in this research work. The reason behind this is the use of ensemble technology which has high computational complexity and hence needs more execution time.

DIRECTIONS FOR FUTURE RESEARCH

The following points can be considered in future to improve the proposed CLTR systems.

- Ensemble static pruning algorithm or dynamic pruning algorithm or a combined static and dynamic pruning algorithm can be incorporated to handle the problem of computation complexity and high execution time.

- The training time of the classifiers can be further reduced, by parallelizing the training process. This is feasible, by identifying operations that are independent to each other and propose a parallel architecture to improve the performance.

- Semantic or ontology based text retrieval can be probed and combined with the proposed classification algorithm.

- The proposed system can be applied for other languages like Kannada, Malayalam and their applicability and performance can be analyzed.