Enhanced Techniques to Improve Speech Query based Tamil to English Cross-Language Text Retrieval System using Ensemble Models

Thesis submitted in Partial Fulfilment of the

Degree of Doctor of Philosophy (Ph.D.)

By

P.Iswarya

11PH34

Department of Computer Science

Avinashilingam Institute for Home Science and Higher Education for Women,

Coimbatore – 641043

August 2015

CERTIFICATE

I certify that the thesis entitled **"Enhanced Techniques to Improve Speech Query based Tamil to English Cross-Language Text Retrieval System using Ensemble Models** "submitted for the degree of **Doctor of Philosophy (Ph.D.)** by **Mrs. P.Iswarya,** is the record of research work carried out by her during the period from June 2011 to August 2015 under my guidance and supervision, and that this work has not formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or other Titles in this University or any other University or Institution of Higher Learning.

10/8/15

Signature of the Head of the Department

Dr.G.PADMAVATHI
M.Sc., M.Phil., Ph.D.
Professor and Head
Department of Computer Science
Avinashilingam Institute for Home Science
and Higher Education for Women
Coimbatore - 641 043

10/08/15

Signature of the Supervisor

Dr. (Mrs) V. RADHA
Professor, Dept. of Computer Science
Avinashilingam Institute for Home Science
and Higher Education for Women
Coimbatore - 641 043

10.08.15

Signature of the Dean

Dean, Faculty of Science
Avinashilingam Institute for Home Science
and Higher Education for Women
(Deemed University)
Coimbatore-641 043

DECLARATION

I declare that the thesis entitled **"Enhanced Techniques to Improve Speech Query based Tamil to English Cross-Language Text Retrieval System using Ensemble Models "** submitted by me for the degree of **Doctor of Philosophy (Ph.D.)** is the record of work carried out by me during the period from June 2011 to August 2015 under the guidance of **Dr. (Mrs.) V.Radha,** Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, and has not formed the basis for the award of any Degree, Diploma, Associateship, Fellowship, Titles in this University or any other University or other similar Institution of Higher Learning.

P. [signature] 10/08/15

Signature of the Candidate

ACKNOWLEDGEMENT

*First and Foremost I would like to Thank **God**. You have given me the power to believe in myself and pursue my dreams. I could never have done this without the faith I have in you, the Almighty.*

*I would like to thank the Chancellor **Dr. T.S.K. MEENAKSHI SUNDARAM**, M.A., M.Phil., Ph.D., Avinashilingam Institute for Home Science and Higher Education for Women, for providing an opportunity to conduct the study.*

*I express my heartfelt gratitude to **Dr. (Mrs.) SHEELA RAMACHANDRAN**, M.Sc., P.G. Dip., Ph.D., Vice Chancellor, Avinashilingam Institute for Home Science and Higher Education for Women, for her supportive attitude and encouragement.*

*I offer my profound gratitude to **Dr. (Mrs.) A.VENMATHI**, M.Sc., Dip.Ed., M.Phil., Ph.D., Registrar , Avinashilingam Institute for Home Science and Higher Education for Women, for the Administrative Support rendered throughout the span of the research work.*

*I express my sincere thanks to **Dr. (Mrs.) A.PARVATHI,** M.Sc., Dip.Ed., M.Phil., Ph.D., Dean, Faculty of Science , Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her spontaneous and timely help and amenities provided for the successful completion of this research work.*

*I also extend my thanks to **Dr. (Mrs) G. PADMAVATHI**, M.Sc., M.Phil., Ph.D., Professor and Head, Department of Computer Science , Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her support, encouragement and cooperation rendered towards the completion of this research.*

I take immense pleasure to express my sincere and deep sense of gratitude to my Supervisor ***Dr. (Mrs). V. RADHA****,* M.Sc., P.G.D.C.A., P.G.D.O.R., B.Ed., M.Phil., Ph.D., *Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her sustained enthusiasm, creative suggestions, motivation and exemplary guidance throughout the course of my doctoral research. Apart from the subject of my research, I learnt a lot from him, which I am sure, will be useful in different stages of my life. I solemnly submit my honest and humble thanks to her for bringing my dreams into reality.*

I am very much grateful to ***Mr. KARNAMURTHI*** *for correcting all the grammatical mistakes in the write up. I deeply appreciate their timely help and constructive criticism which brought the document to shape.*

I accord my warm thanks to all the ***MEMBERS OF THE FACULTY, NON-TEACHING STAFF, AND RESEARCH SCHOLARS*** *from the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for their cooperation.*

A special thanks to my family. Words cannot express how grateful I am to my ***PARENTS*** *for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank all of my* ***FRIENDS*** *who supported me in writing, and incented me to strive towards my goal. At the end I would like express appreciation to my beloved husband* ***Mr. R. GANESAN*** *who spent sleepless nights with and was always my support in the moments when there was no one to answer my queries.*

P. ISWARYA

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADABOOST	ADAptive BOOSTing
AMOSA	Archived Multi-Objective Simulated Annealing
ANN	Artificial Neural Network
AP	Average Precision
ARC	Associative Rule-based Classifier
ARC-BC	Association Rule-based Categorizer- By Category
ARM	Association Rule Mining
ASR	Automatic Speech Recognition
ASRS	Automatic Speech Recognition System
ATSQR	Automatic Tamil Speech Query Recognition
BIM	Binary Independence Model
BM	Best Match
CLEF	Cross-Language Evaluation Forum
CLIR	Cross-Language Information Retrieval
CLTR	Cross-Language Text Retrieval
CS	Cluster Score
CSL	Computer Speech Laboratory
CSR	Continuous Speech Recognition
CWFT	Context Words Features of a Token
CWSD	Conventional Word Sense Disambiguation Algorithm
D4	Daubechies wavelet-4
DCM	Dirichlet Compound Multinomial
DCT	Discrete Cosine Transformation
DFR	Deviation From Randomness
DFT	Discrete Fourier Transform
DT	Decision Tree
DTQ	Descriptive Topic Queries

LIST OF ABBREVIATIONS

DTW	Dynamic Time Wrapping
DWPT	Discrete Wavelet Packet Transformation
DWT	Discrete Wavelet Transform
EBMT	Example Based Machine Translation
ECLTR	Existing Cross-Language Text Retrieval
EFS	Ensemble Feature Selection
EHAKNN	Ensemble Hybrid ARC and IKNN Classifier
ESMA	Ensemble SVM-based Morphological Analyzer
FIRE	Forum for Information Retrieval and Evaluation
FIRE	Forum for Information Retrieval Evaluation
HAKNN	Hybrid Association Rule Mining and Improved KNN
HLDA	Hetroscedastic Linear Discriminant Analysis
HMM	Hidden Markov Model
HQT	Hybrid Query Translation
HTK	Hidden markov model Tool Kit
IARC	Improved Associative Rule-based Classifier
IDWT	Inverse Discrete Wavelet Transform
IKNN	Improved KNN
ILIR	Indian Language Information Retrieval
IR	Information Retrieval
IT	Information Technology
IWR	Isolated Word Recognition
KL	Kullback-Leibler
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LPCC	Linear Predictive Cepstral Co-efficient
LVCSR	Large Vocabulary Continuous Speech Recognition

LIST OF ABBREVIATIONS

MA	Morphological Analyzer
MAD	Median Absolute Deviation
MAP	Mean Average Precision
MAP	Mean Average Precision
MFCC	Mel Frequency Cepstral Co-efficients
MFR	Multiple Frame Rate
MFS	Multiple Frame Size
MLLR	Maximum Likelihood Linear Regression
MLP	Multi-Layer Perceptron
MLTR	Mono-Lingual Text Retrieval
MOO	Multi-Objective Optimization
MSE	Mean Square Error
MSE	Mean Square Error
MT	Machine Translation
NER	Named Entity Recognition
NICO	Neural Inference COmputation
NII	National Institute for Informatics
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NTQ	Narrative Topic Queries
OOV	Out-of-Vocabulary
PCA	Principal Component Analysis
POS	Part-of-Speech
POSHES	POS Tagging using Hybrid SVM-WNN Ensemble Classifier with Ensemble Feature Selection

LIST OF ABBREVIATIONS

POSSA	POS Tagging using SVM classifier with Simulated Annealing and AMOSA
PRF	Pseudo-Relevance Feedback
PRP	Probability Ranking Principle
QE	Query expansion
QT	Query Translation
S	Conventional SVM Classifier
SA	Simulated Annealing
SCP	Support-Confidence based algorithm
SDK	Software Development Kit
SDR	Spoken Document Retrieval
SES	Ensemble SVM Classifier enhanced with SOM
SMA	SVM-based Morphological Analyzer
SNR	Signal to Noise Ratio
SNR	Signal to Noise Ratio
SOM	Self Organizing Map
SRDSTQ	Silence Removal from Denoised Speech Tamil Query
SS	SVM classifier enhanced with SOM
STE	Short Time Energy
STTQ	Short Title Topic Queries
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression
TDIL	Technology Development for Indian Languages
TECLTR-S	Tamil-English Cross-Language Text Retrieval for Speech query

LIST OF ABBREVIATIONS

TECLTR-T	Tamil-English Cross-Language Text Retrieval for Text query
TF-IDF	Term Frequency-Inverse Document Frequency
TQTS	Tamil Query Translation System
TQTS	Tamil Query Translation System
TREC	Text Retrieval Conference
VSM	Vector Space Model
W2SHT	Wavelet Denoising based on Switching Soft and Hard Thresholding
WHT	Wavelet Denoising based on Hard Thresholding
WNN	Wavelet Neural Network
WOPP	Without Pre-Processing
WPD	Wavelet Packet Decomposition
WPLPCC	Wavelet Packet Based LPCC
WPMFCC	Wavelet Packet Based MFCC
WPP	With Pre-Processing
WSD	Word Sense Disambiguation
WSDCS	Word Sense Disambiguation Algorithm enhanced with Clustering and Sense-collocation Dictionary
WSDPCS	Enhanced WSD with Part-of-speech and Clustering based Sense-collocation
WST	Wavelet Denoising based on Soft Thresholding
WWQT	Word by Word Query Translation
WWW	World Wide Web
ZCR	Zero Crossing Rate

ABSTRACT

Identification of techniques to improve the process of searching and retrieving the interested documents from large databases is a challenging task in information retrieval systems. Cross-Language Information Retrieval (CLIR), where queries and documents to be retrieved are in different languages, has become one of the major topics within the world of information retrieval community.

The main objective of this research work is to design and develop Tamil-English Cross-Language Text Retrieval (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 achieve this objective, the research methodology was designed in three phases, where the first phase focused on Tamil query speech recognition, second phase focused on query translation (from Tamil to English) and the third phase concentrated on relevant English documents.

Speech recognition was performed in three steps. The first step performed noise removal, using wavelet-based denoising algorithm using Switching Soft and Hard Thresholding and silence removal based on two criteria, short time energy and zero crossing rate combined with automatic threshold estimation and rule-based algorithms. Feature extraction was performed using wavelet packet based Mel Frequency Cepstral Co-efficients and Linear Predictive Cepstral Co-efficients, which was standardized to same length using Self Organizing Map. The speech recognition is performed using an ensemble SVM classifier.

The query translation was performed using rule-based machine translation and statistical machine translation. Rule-based machine translation includes tokenization, Part-of-Speech (POS) tagging, chunking, morphological analysis, named entity recognition, word sense disambiguation and translation. The statistical machine translation is applied to carry out transliteration procedure.

The proposed POS tagging algorithm performs feature extraction and selections using three feature selection algorithms, whose output are combined to form an optimal

feature set. This is used by a fast SVM classifier to produce an optimized training vector that trains WNN classifier to improve the accuracy of POS tagging. The morphological analysis was performed using an ensemble SVM classifier with trained models. The model segments the words into morphemes, and it is tagged using corresponding labels. The word sense disambiguation was performed using an ensemble K-Means clustering algorithm with sense-collocation dictionary. The root words are translated using knowledge sources, and grammatical categories are replaced by implementing tense marker and gerund ending rules. Untranslated named entities are transliterated using statistical machine translation. Finally rule-based re-ordering is implemented to get correct structure of English sentence.

The final phase of the research work performs document retrieval using an ensemble Hybrid classification system using improved K-Nearest Neighbor classifier enhanced through the use of associative rule mining. The ranking of the retrieved documents was done using Okapi BM25 algorithm.

Experiments to evaluate the performance of the proposed algorithms were performed using Tamil queries from Forum for Information Retrieval and Evaluation (FIRE) dataset 2011. Three types of queries namely, short term title queries; descriptive title queries and narrative title queries were used. Performance metrics like Signal to noise ratio, Mean Square error, Speech recognition accuracy and speed were used to evaluate algorithms in Phase I. Phase II algorithms were evaluated using Precision, Recall, F- Measure, Accuracy and Speed. Phase III algorithms were analyzed using Precision, Recall, Mean Average Precision (MAP), F-Measure, Accuracy and Speed of retrieval.

Experimental results revealed that the proposed Tamil-English Cross -Language Text Retrieval with speech-based and text-based queries produced improved results, when compared with the conventional and existing algorithms, while using the various proposed algorithms in each phase.