INFORMATION RETRIEVAL IN
MULTILINGUAL ENVIRONMENT: PHONETIC
AND SEMANTIC ALONG WITH ONTOLOGY
APPROACHES FOR HINDI AND MARATHI

THE THESIS SUBMITTED BY

SANDEEP CHAWARE
(Gr. No. 71118070002)

FOR THE AWARD OF THE DEGREE OF

DOCTOR OF PHILOSOPHY
IN
ENGINEERING

Under the guidance of

DR. SRIKANTHA RAO

Mukesh Patel School of Technology, Management and Engineering
Vile Parle (West), Mumbai-400056
CERTIFICATE

This is to certify that the work entitled ‘Information Retrieval in Multilingual Environment: Phonetic and Semantic Along with Ontology Approaches for Hindi and Marathi’, is a piece of research work carried out by Mr. Sandeep Manohar Chaware under my guidance and supervision for the degree of Doctor of Philosophy in Engineering in Computer department of MPSTME, SVKM’s NMIMS (Deemed-to-be University), Vile Parle (West), Mumbai-56.

To the best of my knowledge and belief the thesis –

I) Embodies the research work carried out by the candidate himself.
II) Has duly been completed.
III) Fulfills the requirements of the ordinance relating to the PhD degree of University.
IV) Is up to the standard in respect of both, the contents and language for being referred to the expectations.

Date: 
Place: 
Signature of Guide

Forwarded by:
Dr. D.J. Shah
Dean, MPSTME,
SVKM’s NMIMS, Mumbai-56
This is to certify that the thesis titled ‘*Information Retrieval in Multilingual Environment: Phonetic and Semantic Along with Ontology Approaches for Hindi and Marathi*’, for PhD in Engineering submitted to SVKM’s NMIMS (Deemed-to-be University) by Mr. Sandeep Manohar Chaware a bonafide student of MPSTME, Vile Parle (West), Mumbai-56 has been evaluated and recommended for the reward of Doctor of Philosophy in Engineering.

Signature:          Signature:          
Name:               Name: Dr. Srikantha Rao  
Examiner           Research Guide and Supervisor

Date:               Date:               

Dean,  
Dr. D.J. Shah  
MPSTME,  
SVKM’s NMIMS, Mumbai-56
DECLARATION

I hereby declare that the thesis work entitled ‘Information Retrieval in Multilingual Environment: Phonetic and Semantic Along with Ontology Approaches for Hindi and Marathi’ is my own research work carried out under guidance and supervision of Dr. Srikantha Rao.

I further declare that to the best of knowledge the thesis does not contain any part of any work which has been submitted for the award of any degree either in the University or in any other University without proper citation.

Date:

Place:

Signature of Candidate        Signature of Guide
Mr. Sandeep Chaware           Dr. Srikantha Rao

Forwarded by:

Dean,
Dr. D.J. Shah
MPSTME
SVKM’s NMIMS, Mumbai
ACKNOWLEDGEMENTS

I am extremely grateful to Dr. Srikantha Rao, my advisor, for many lessons on how to do research and write research papers, for being very supportive in my work, for guiding into each part of the research work and life in general. Specifically, I am thankful for countless hours he spent with me in explaining each part, sharing his experiences on his research. Also, I am thankful for his insightful suggestions that helped me made the right strategic choices at many crucial decision points along these years.

I am grateful to our driving force Dr. D.J. Shah, Dean, MPSTME, who is constantly inspiring throughout the years. I also thanks to research committee members of MPSTME, Dr. Vijay Raisinghani, Dr. Ketan Shah for the time and energy they have spent in reviewing my research work and their detailed technical feedback.

My special thanks to Dr. Hari Vasudevan, Principal, DJSCoE, for his support for the entire period. I thanks to Dr. H.B. Kekre, Dr. M.V. Deshpande, Dr. K. S. Kinage and Prof. Ahibijit Joshi for their valuable time-to-time suggestions and kind support.

I thank all my friends and everyone who have contributed to this thesis through many fruitful discussions, technical advice, encouraging words and in many other ways.

Finally, I thank my parents, Shri Manohar and Sau. Jinamati, who have given love, support and understanding over all of these years. I am very special thanks to my beloved wife Mrs. Trushita for accepting my style of living during these years. Her inspiration and love have been endless source of energy that helped me in completing this thesis.

- Mr. Sandeep Chaware
Dedicated to ..........................................................................................................

My Beloved Parents

SHRI. MANOHAR AND SAU. JINAMATI CHAWARE
ABSTRACT

Information retrieval (IR) is an important issue and difficult to handle especially in multilingual environment. The IR process begins with entering a query into the system in prescribed format. A query does not uniquely identify a single object but several objects with different degrees of relevancy. An object is an entity from database which represents the information. User queries are matched against the database information. The matched objects are ranked according to the query terms and shown to the user as information.

Multilingual environment (ME) provides a platform where a query can be inputted in one language and the result can be in the same language or different language. This result can be viewed as IR or knowledge. Further this IR helps to make better decision in various scenarios. This is only possible in a system that supports multilingual environment in all aspects. In order to enhance the IR process of a system, phonetic and semantic issues have to be resolved. The phonetic and semantic matching will enhance the efficiency and utilization of resources for IR. International Phonetic Association (IPA) as phonetic standard gives a set of alphabet as a standard way of representing any language string that is utilized as phonetic matching. Ontology is a promising solution to any semantic matching problem. It typically provides a vocabulary that describes a domain of interest and specification of meaning of terms used in the vocabulary.

Several different solutions have been proposed for phonetic and semantic matching so far. Phonetic matching solutions worked either on developing an operator in a system or designing rules as per phonemes of alphabets of a language. The operator uses text-to-phonetic (TTP) system for translation of a text string into its equivalent phonetic form. This form represents the alphabets from IPA. But TTPs for Hindi or Marathi are cumbersome to use. In semantic matching, solutions proposed so far were based on ontology, where some system used top-down approach or some used bottom up approach. There is no fix approach for ontology building. The solutions have been exploited by using external resources such as WordNet, multi-lingual dictionary, inter-link-index (ILI) and so on. Always use of these
resources makes system more complex. The system evaluated to give results as matching or not matching, no further analysis has been done so far.

The thesis proposes the so-called phonetic matching and semantic matching approach for Hindi and Marathi languages. The phonetic matching approach is based on two key ideas. The first is that various way of writing styles of a string are calculated or considered and the strings are translated to English by using mapping methodology. The results are treated as IR for a query. The second idea is that, we transformed each string into its phonetic form according to our proposed phonetic rules for languages and compared with some threshold value to match. If matches, we are performing IR in order to extract the detailed information.

The semantic matching approach is based on two key concepts. The first is, from query and other from formal and informal question-answer approach for a domain has been developed to acquire correct ontology terms from database in order to match two strings. The ontology is used to represent the knowledge. An integrated approach has been proposed to build ontology. The second is an extension in which ontology terms are being matched semantically by using an approach based domain-based dictionary for translation, and/or domain-based synset for each language. A string will be inputted by a user and matched semantically with other string from ontology. If they match then the system has been further extended to an inference system, which we called as IR or knowledge. The performance of both the approaches has been evaluated by taking performance parameters as time to build ontology, ontology terms, precision, recall, f-measure and accuracy.

The basic and iterative phonetic and semantic matching algorithms have been designed and developed. The approach has been evaluated on various real world test cases with encouraging results proving its benefits.

**Keywords:** phonetic matching, semantic matching, writing style, information retrieval, ontology, domain-based dictionary, domain-based synsets.
List of Figures

Figure 1.1: Multilingual Environment ................................................................. 25
Figure 1.2: Hypothetical Example of Shopping Mall ........................................ 29
Figure 1.3: Woman at Grocery Shop with a Query in Mind ................................ 30
Figure 1.4: Hypothetical Example of a Grocery Shop ........................................ 30
Figure 1.5: Hindi Query Keyword and Its Equivalent SQL Query .................... 32
Figure 1.6: Example of Phonetic Comparison of Hindi Strings ......................... 33
Figure 1.7: Simple SQL Query and Inference Query ......................................... 34
Figure 2.1: Matching Process ........................................................................... 37
Figure 2.2: Phonetic Matching Approach-I using Mapping Methodology ........... 39
Figure 2.3: Phonetic Matching Approach-II using Phonetic Rules of Hindi and Marathi .... 39
Figure 2.4: Semantic Matching Approach using Ontology ............................... 40
Figure 3.1: General Architecture of Text-To-Phonetic (TTP) based System .......... 45
Figure 3.2: General Architecture of Code-based System .................................... 45
Figure 3.3: General Architecture of Phonetic Rule-based System .................... 46
Figure 3.4: Multilingual Name Query with MLLike Operator ......................... 48
Figure 3.5: Multilingual Name Query with MLLexEqual Operator ................... 49
Figure 4.1: Various Forms of Ontology ............................................................. 58
Figure 4.2: Graphical Depiction of Ontology .................................................. 58
Figure 4.3: Morteza Poyan Rad Semantic Matching Approach for Documents .......................... 61
Figure 4.4: Atanas Kiryakov Semantic Matching Approach .................................................. 62
Figure 4.5: Synset-based Multilingual Dictionary for Semantic Matching .................................. 64
Figure 4.6: Ontology Matching Approach as Semantic Matching .......................................... 65
Figure 4.7: Semantic Matching for Product Domain Using Ontology ...................................... 66
Figure 5.1: Generalized Phonetic Matching Approach .............................................................. 83
Figure 5.2: System Architecture for Phonetic Matching with IR ............................................. 84
Figure 5.3: Parsing of a String ‘रघूिलला’ ................................................................................. 91
Figure 5.4: SQL Query for a String ‘Raghulila’ and Other Forms ............................................ 92
Figure 5.5: Sample Input Interface in Hindi for Phonetic Matching with IR .............................. 95
Figure 5.6: Result of Sample Query ......................................................................................... 96
Figure 5.7: Indic-Phonetic Approach: System Architecture ...................................................... 97
Figure 5.8: User Interface for Phonetic Matching Approach for Hindi and Marathi ................. 109
Figure 5.9: Three Phonetic Matching Algorithms to Compare .................................................. 110
Figure 5.10: Result of Soundex Approach for Hindi ................................................................. 110
Figure 5.11: Result of Q-Gram Approach for Hindi ................................................................. 110
Figure 5.12: Result of Indic-Phonetic Approach for Hindi ....................................................... 111
Figure 5.13: Phonetic Name-wise Search for Domain for IR .................................................... 111
Figure 5.14: IR after Phonetic Matching .................................................................................. 111
Figure 5.15: Comparison of Phonetic Matching Approaches .................................................... 112
Information Retrieval in Multilingual Environment: Phonetic and Semantic Along with Ontology Approaches for Hindi and Marathi

Figure 6.1: System Architecture for Semantic-Based IR using Ontology ........................................ 116

Figure 6.2: Proposed Overall System Architecture for Ontology Building ..................................... 119

Figure 6.3: Relation Schema to Ontology Translation Rules ............................................................. 122

Figure 6.4: Ontology for ‘पूरी छोले’ .......................................................................................... 124

‘Puri-chhole’ Sub-domain

Figure 6.5: Time Required to build Ontology for Hindi Users ....................................................... 128

Figure 6.6: Time Required to build Ontology for Marathi Users .................................................... 128

Figure 6.7: Comparison of Time in ms Required to build Ontology for Hindi & Marathi Users ................................................................. 128

Figure 6.8: Precision and Recall for Hindi Users for Building Ontology .................................. 129

Figure 6.9: Precision and Recall for Marathi Users for Building Ontology ............................. 129

‘Ragda-patis’ Sub-domain

Figure 6.10: Time Required to build Ontology for Hindi Users ...................................................... 130

Figure 6.11: Time Required to build Ontology for Marathi Users ............................................... 130

Figure 6.12: Comparison of Time in ms Required to Build Ontology for Hindi and Marathi Users .................................................................................................................................. 131

Figure 6.13: Precision and Recall For Hindi Users for Building Ontology .......................... 131

Figure 6.14: Precision and Recall For Marathi Users for Building Ontology ........................ 131

Inference System

Figure 6.15: Precision, Recall, F-Measure and Accuracy for Hindi Users .................................. 138

Figure 6.16: Precision, Recall, F-Measure and Accuracy for Marathi Users .......................... 138
Figure A: Main User Interface for Selection of Sub-domain .................................................. 143
Figure B: Main User Interface for Selection of Language .................................................... 143
Figure C: Question-Answer (Q-A) Interface1 to build Ontology ........................................ 144
Figure D: Question-Answer (Q-A) Interface2 to build Ontology ........................................ 144
Figure E: Ontology for Sub-domain ................................................................................... 145
Figure F: User Interface for Inference Queries for Hindi User .......................................... 146
Figure G: User Interface for Inference Queries for Marathi User ........................................ 146
Figure A1: Hindi Traditional Keyboard Layout for Consonants & Modifiers .................. 152
Figure A2: Hindi Traditional Keyboard Layout for Vowels & Consonants ....................... 152
Figure A3: Hindi Traditional Keyboard Layout for Numbers ............................................ 153
Figure A4: Hindi Traditional Keyboard Layout for Special Characters ......................... 153
Figure B1: ISCII Code Table for Devanagari ................................................................. 155
Figure B2: Unicode Code Table for Devanagari ............................................................ 159
## List of Tables

Table 3.1: Soundex Phonetic Codes ................................................................. 50
Table 3.2: Soundex Code for Hindi as Per Algorithm ....................................... 51
Table 3.3: Q-Gram Example ........................................................................... 52
Table 3.4: Q-Gram Example for Hindi .............................................................. 53
Table 4.1: Comparison of Ontology Building Approaches ................................. 75
Table 4.2: Ontology Building Approaches from Database ................................. 79
Table 5.1: Hindi-to-English Transliteration Mapping Table .............................. 92
Table 5.2: English-to-Hindi Transliteration Mapping Table .............................. 93
Table 5.3: Results of Phonetic Matching for Hindi Strings ............................... 105
Table 5.4: Results of Phonetic Matching for Marathi Strings ............................ 107
Table 5.5: Strings with Codes and Matching Status for Hindi and Marathi .......... 112
Table 6.1: Representation of Ontology with Relationships ............................... 124
Table 6.2: Total Ontology Terms and Time in ms required to build Ontology for Hindi and Marathi Users for ‘puri-chhole’ sub-domain ......................................................... 127
Table 6.3: Total Ontology Terms and Time in ms required to build Ontology for Hindi and Marathi Users for ‘ragda-patis’ sub-domain ......................................................... 130
Table 6.4: Accuracy, Precision and Recall for Ontology Building ..................... 132
Table 6.5: Inference Queries for the Sub-domains for Hindi and Marathi Users .... 137
Table 6.6: Precision, Recall, F-measure and Accuracy in % for Hindi and Marathi Users for Inference Approach ........................................................................ 137
List of Proposed Algorithms

3.4.3.2: Phonetic Matching using Soundex Approach for Hindi or Marathi ............................................. 51

3.4.3.4: Phonetic Matching using Q-Gram Approach for Hindi or Marathi ......................................... 53

Algorithm 5.2.3: Phonetic Matching Algorithm-I .................................................................................. 90

Algorithm 5.3.3: Phonetic Matching Algorithm-II ................................................................................ 100

Algorithm 6.4.4: Ontology Building: The Algorithm ........................................................................... 122

Algorithm 6.5: Semantic Matching: The Algorithm with IR ................................................................. 133
List of Abbreviations

IR – Information Retrieval
IPA – International Phonetic Association
TTP -Text-to-Phonetic/Phonemic
ILI – Inter Link Index
IT – Information Technology
TDIL – Technology Development of Indian Languages
OCR – Optical Character Recognition
SQL- Structured Query Language
CPU – Central Processing Unit
XML – eXtensible Markup Language
RDBMS – Relational Database Management System
OWL – Ontology Web Language
IC – Information Content
ATE – Atomic Text Element
MT – Machine Translation
CLIR – Cross-Language Information Retrieval
MRD – Machine Readable Dictionary
TOVE – Toronto Virtual Enterprise
KB – Knowledge Base
NF- Normal Form

ASCII – American Standard for Code Information Interchange

ASP – Active Server Pages

IME – Input Method Editor

DoE – Department of Electronics

ISCI – Indian Script Code for Information Interchange

DB2 – Database 2

UTF – UCS Transformation Format

NChar – National Char

HD – Hard Disk

PC – Personal Computer

RAM – Random Access Memory

H-E/M-E – Hindi-English/Marathi-English

Q-A – Question-Answer

GB – Giga Byte

CD – Compact Disk

OKBC – Open Knowledge Base Connectivity

GIST - Graphics and Intelligence based Script Technology

ME – Multilingual Environment
Table of Contents

Certificate ................................................................................................................................... 2
Thesis Approval Certificate ....................................................................................................... 3
Declaration ................................................................................................................................. 4
Acknowledgements .................................................................................................................... 5
Abstract ..................................................................................................................................... 7
List of Figures ............................................................................................................................ 9
List of Tables ........................................................................................................................... 13
List of Proposed Algorithms .................................................................................................... 14
List of Abbreviations ............................................................................................................... 15

I. Introduction .......................................................................................................................... 21
  1. The Introduction ............................................................................................................ 22
      1.1 Introduction ............................................................................................................ 22
      1.2 Multilingual Environment ..................................................................................... 24
      1.3 Language Issues for Hindi and Marathi ................................................................. 26
      1.4 Phonetic Matching Issues for Hindi and Marathi .................................................. 27
      1.5 Semantic Matching Issues for Hindi and Marathi ................................................. 27
      1.6 Motivations with Information Retrieval (IR) ........................................................ 28
      1.7 Research Issues Explored ...................................................................................... 32
      1.8 Organization of the Thesis ..................................................................................... 34

2. The Matching Problem ...................................................................................................... 36
   2.1 Introduction .............................................................................................................. 36
   2.2 Problem Statement and Proposed Design ............................................................... 36
   2.3 Scope of the Research ............................................................................................. 41
   2.4 Principal Findings .................................................................................................... 41
   2.5 Summary ................................................................................................................ 42
II State Of The Art: Phonetic and Semantic Matching Approaches

3. Phonetic Matching Approaches ................................................................. 44
   3.1 Introduction ....................................................................................... 44
   3.2 Classification of Matching Approaches .............................................. 44
   3.3 Matching Strategies ........................................................................ 46
   3.4 On-Hand Approaches ................................................................. 47
   3.5 Drawbacks of Existing Phonetic Matching Approaches ................. 54
   3.6 Summary ......................................................................................... 54

4. Semantic Matching Approaches ............................................................ 55
   4.1 Introduction ....................................................................................... 55
   4.2 Theoretical Foundation ............................................................... 55
   4.3 On-hand Approaches ..................................................................... 61
   4.4 Classification of Semantic Matching Approaches ......................... 68
   4.5 Ontology for Matching ................................................................. 71
      4.5.1 Ontology Building Approaches ............................................... 71
      4.5.2 Drawbacks of Existing Ontology Development Approaches .... 75
      4.5.3 Implementation of Ontology Building from Database ............. 76
      4.5.4 Drawbacks of Ontology from Database Approaches ............. 79
   4.6 Drawbacks of Existing Semantic Matching Approaches using Ontology ............................................................................. 79
   4.7 Summary ......................................................................................... 80

III. Proposed Solutions for Phonetic and Semantic Matching

5. Phonetic Matching Approaches ................................................................. 82
   5.1 Introduction ....................................................................................... 82
   5.2 Approach I ...................................................................................... 84
      5.2.1 Matching Methodology .......................................................... 84
      5.2.2 Implementation Details .......................................................... 84
IV. Results and Discussion ................................................................. 140

V. Conclusions and Future Work ..................................................... 147

Appendix .............................................................................................. 150

Appendix – A: IMEs for Hindi and Marathi Languages .................. 151
Appendix – B: Encoding standards such as ISCII and Unicode ........ 154
Appendix - C: Bi-lingual Domain-based Dictionary and Synsets for Hindi, Marathi and English Languages ................................................... 160
Appendix – D: Case study as domains (Shopping Mall and Grocery Shop) .... 168

References ............................................................................................ 170

List of Publications .............................................................................. 176