Cross Lingual Information Retrieval

Special Emphasis for Sibling Language Hindi-Gujarati

Author

Dr. Kalyani A. Patel

A

ASTER

Publishing

London * New York * New Delhi

First Published 2024

Disclaimer-

The author and the publisher have taken every effort to the maximum of their skill, expertise and knowledge to provide correct material in the book. Even then if some mistakes persist in the content of the book, the publisher does not take responsibility for the same. The publisher shall have no liability to any person or entity with respect to any loss or damage caused, or alleged to have been caused directly or indirectly, by the information contained in this book. Hence, the book should be taken as general guide only.

The publisher has fully tried to follow the copyright law. However, if any work found to be similar, it is unintentional and the same should not be used as defamatory or to file legal suit against the author/publisher. If the readers find any mistakes, we shall be grateful to them to point out those to us so that they can be corrected in the next edition.

ASTER Publishing
London, New York, New Delhi
E-mail : asterpublishing@gmail.com

in association with
Apple Books, India
E-mail : applebookspd@gmail.com

British Library Cataloguing in Publication Data
ddc 025.04

A catalogue record for this is available from the British Library

ISBN 978-93-92007-81-1

Typeset by SR Digitals, New Delhi
Printed in India at Trident Press, Noida
Printed on paper from sustainable resources.

PREFACE

India is one of the richest countries, with eighteen constitutional languages which are written in ten different scripts in the field of linguistics. Along with this vernacular linguistic richness, about a billion people use these languages as their first language. A huge amount of regional news and cultural information is usually found on the web in these languages and inaccessible to people of other regions within the country, essentially because of a language barrier. This in turn has sharpened the focus on the demand to cross language barrier at a rapid pace, with no end in sight for this growth in appetite.

Indian languages are highly inflectional with a rich morphology, relatively free word order, and default sentence structure as SOV (Subject Object Verb). Many of them are structurally similar which are better known as sibling languages. The study on cross lingual information extraction and retrieval for similar structured languages with special emphasis to Hindi and Gujarati has been undertaken by us keeping in view the following points.

- Gujarati is still lesser touched language. With a Gujarat Government's emphasis on the basics of e-governance and the development of the Information Technology sector, Gujarati language is slated to attain the significance it aspires towards Digital Era. Therefore, it is required to overcome language barrier from Gujarati to any other Indian language or English language and vice versa.
- The structurally similar language Hindi is considered as an Interlingua to cross language barrier for Gujarati because
    - Hindi is a National language and the transformation applications form Hindi language to Indian languages, Hindi language to English language and vice versa has been developed and tested.

  - o Machine translation between closely related languages is easy and efficient.

The principal aim of this research study is to boost up Gujarati on Internet via Hindi as an Interlingua, considering fascinating factor of closely related languages. It is an encouragement of cross lingual information retrieval systems in true sense. The performance of cross lingual information extraction and retrieval system heavily depends on how the query is processed. The study focuses on disambiguated query expansion and translation tasks of query processing. The major contribution of the book can be outlined as follows:

- Query Expansion using Hindi WordNet
- Word Sense Disambiguation using Lesk like algorithm [Lesk,1996].
- Word-for-Word Translation
  - o Development of a generalized model for storage of typologically different words in Hindi and Gujarati.
  - o Design and development of rule-base for
    - § String-pattern-based transformations of morphologically rich and lexically similar words in Hindi and Gujarati.
    - § Mapping of lexical, grammatical and structural divergent words between Hindi and Gujarati languages.
- Development of a system called GH-MAP for rule based token mapping for translation from Hindi to Gujarati and vice-versa.
- Development of search engine prototype to retrieve documents in Hindi-Gujarati.
- Development of evaluation software to calculate PER, METEOR and BLEU score considering into-Hindi / into-Gujarati characteristics.
- A set of machine judgment like F-score, PER, BLEU and METEOR score, has been calculated and compared to establish the relevance of model.

The book is organized in eleven chapters.

Chapter 1 provides research prologue which includes motivation, aim, objectives, and benefits of research. It also includes description of the work already done. Chapter 2 provides overview of information processing, to appreciate the resemblance among information extraction, information retrieval, cross lingual information retrieval, query processing and multilingual information retrieval. Chapter 3 describes possible approaches of machine translation along with the related work done by various researchers. Chapter 4 includes comparative study of Hindi and Gujarati language to understand the similarities and differences between the both. It provides an inspiration to create effective mapping rules amongst languages within the same group with fewer efforts to overcome the language barrier. Chapter 5 provides overview of Hindi and Gujarati Cross Lingual Information Extraction and Retrieval System along with the related work done by various researchers. Chapter 6 describes Query Expansion

Engine along with algorithm and output. Chapter 7 describes Query Translation system along with related work. Provide overview of GH-MAP system. Chapter 8 describes rule base and data structure used for GH-MAP. Chapter 9 describes translation mechanism used by GH-MAP. Also presents algorithm and output of GH-MAP system. Chapter 10 explains evaluation and discusses result. Chapter 11 concludes the contribution of research work in this book. Future work possibilities are discussed in this chapter. Appendix A contains WX alphabet notation used for transliteration Hindi and Gujarati characters, for readers' convenience. Appendix B contains GH-MAP: User Interface screen shots for addition of new rules. Appendix C contains Example of a block of 50 Sentences. Appendix D contains screen shot for BLEU, METEOR and PER scores; exemplary sentences are given in appendix C.

— ***Kalyani Patel***

ACKNOWLEDGEMENT

First and foremost I want to thank Late. Prof. Dr. Rekha Govil, who had not only introduced me to the field of cross lingual information extraction and retrieval but had also helped me during the first phase of the study, she had been a great source of inspiration all along.

I would like to express my sincere thanks to my guide Dr. Jyoti Pareek. She as a mentor and guide taught me the basics of research and stayed with me right till the end. Her efforts, comments, advices and ideas developed my thinking, and improved my way of presentation. Without her constant encouragement, keen interest, inspiring criticism and invaluable guidance, I would have not accomplished my work. I admit that her efforts need much more acknowledgement than expressed here.

I acknowledge and thank the Banasthali University.

I would like to express my sincere thanks to Prof. Pushpak Bhattacharyya, who allowed me to use Hindi WordNet, developed at Center for Indian Language Technology at IIT Bombay. I would like to express my thanks to Dr. B.R. Patel (Ex. Principal and well known critic in Gujarati Language and Literature), Dr. Yogendra Vyas (Ex. Head of Department of Linguistics, Gujarat University), Dr. Arvind Bhandari(Head of Department of Linguistics, Gujarat University), for many fruitful discussions regarding doubt and query in my research. I express my gratitude to Dr. Sarla Achyutam, Director, K.S. School of Business Management, Dr Savita Gandhi , Jt. Director, K.S. School of Business Management and adore to my faculty colleagues and staff members for moral support.

I express my hearty thanks to Mrs. Parul Patel(Professor, Nirma Institute of Technology), an amazing friend who allowed me to stay with her when I had visited library at IIT Bombay for literature survey. I would like to express special thanks to Mrs. Sonal Jain (Associate Professor, GLSICT), who read the manuscript and gave

honest comments. I wish to extend my warmest thanks to Mrs. Payal Mehta, who proof read the manuscript. I am thankful to my students Ms. Parthi Gandhi and Ms. Sona Shah for helping me in programming.

I would like to express special thanks to my cousin Mr. Deven Patel for his remote help in providing all electronic documents.

This research work would have never been possible without the blessings of my parents, who have been there for me with unconditional support throughout my childhood and academic career. I convey my sincere regards to my husband and sister for the sacrifices they have made, for the patience they have shown, and for the love and blessing they have showered. Most imperative of all, I would like to express my profound sense of gratitude and appreciation to my daughter Prutha.

I would like to thank everybody who has been important to the successful realization of book, as well as expressing my apology that I could not mention personally one by one.

Finally, I thank GOD for everything.

— Kalyani Patel

CONTENTS