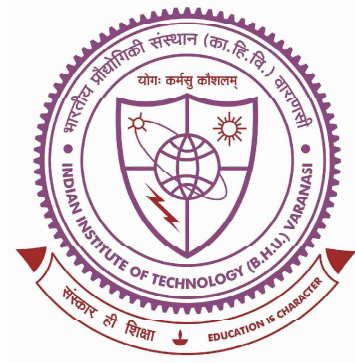


Pre-processing in Indian language IR



Thesis submitted in partial fulfilment
for the award of degree

Doctor of Philosophy

by

Siba Sankar Sahu

**DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING**

**Indian Institute of Technology
(Banaras Hindu University)
Varanasi**

Roll No: 17071012

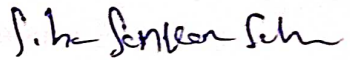
Year of Submission: 2024

DECLARATION

I, Siba Sankar Sahu, certify that the work embodied in this thesis is my own bona fide work and carried out by me under the supervision of Dr. Sukomal Pal from July-2017 to April-2023, at the Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.


Date:

Place: Varanasi


Siba Sankar Sahu

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.


Dr. Sukomal Pal
पर्यवेक्षक/Supervisor
IIT(BHU), Varanasi
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)
(Banaras Hindu University)
वाराणसी/Varanasi-221005

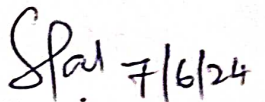

Signature of Head of Department/Coordinator of School

साचार्य व विभागाध्यक्ष
Professor & Head
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)
(Banaras Hindu University)
वाराणसी/Varanasi-221005

To
The
Supreme Power God
and
My beloved family

CERTIFICATE

It is certified that the work contained in the thesis titled **Pre-Processing in Indian Language IR** by **Siba Sankar Sahu** has been carried out under my supervision and that this work has not been submitted elsewhere for a degree. It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy and SOTA for the award of Ph.D. Degree.


Supervisor

Dr. Sukomal Pal

Associate Professor

Department of Computer Science and Engineering

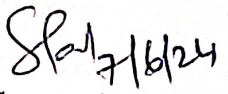
Indian Institute of Technology (BHU) Varanasi

Varanasi, INDIA, 221005

CERTIFICATE

This is to certify that the revised thesis titled **Pre-Processing in Indian Language IR** is being submitted by **Siba Sankar Sahu** in partial fulfillment for the award of Ph.D. in Department of Computer Science and Engineering, IIT (BHU), Varanasi is a record of bonafide work carried out by him.

Date of Submission: 16/03/2024


Supervisor

Dr. Sukomal Pal

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology (BHU) Varanasi

Varanasi, INDIA, 221005

COPYRIGHT TRANSFER CERTIFICATE

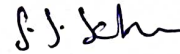
Title of the Thesis: Pre-Processing in Indian Language IR
Name of Student: Siba Sankar Sahu

Copyright Transfer

The undersigned hereby assigns to the Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the Doctor of Philosophy.

Date: 16/03/2024

Place: Varanasi



Siba Sankar Sahu

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

Acknowledgments

Though only my name appears on the cover of this dissertation, so many great people have contributed to its production. I owe my gratitude to all who have made this thesis possible and because of whom my post-graduate experience has been one I will cherish forever.

I take this opportunity to express my profound gratitude and deep regards to my supervisor Dr Sukomal Pal, Associate Professor, Department of Computer Science and Engineering, IIT(BHU), Varanasi, for his exemplary guidance, monitoring and constant encouragement throughout this dissertation.

I am obliged to the faculty members of the Department of Computer Science and Engineering, IIT(BHU), Varanasi, in particular, Prof. Sanjay Kumar Singh and staff members Prof. Anil Kumar Tripathi, Dr Anil Kumar Singh, Dr Pratik Chattopadhyay and Dr Sandip Ghosh, Associate Professor, Department of Electrical Engineering, IIT(BHU), Varanasi. They have provided constant support during my research program.

A very special gratitude goes out to my colleagues and friends, with special mention to Dr Jayraj Singh, Dr Tribikram Pradhan, Dr Pradeepika Verma, Dr Chintoo Kumar, Dr Anita Saroj, Mr Supriya Chanda, Mr Preetam Pal, Mr Biswajit Parmanik, Mr Dinesh Pargai, and Mrs Akanksha Mishra. In addition, I am grateful to Mr Sushil Kulkarni, who has provided me with moral and emotional support during my research program.

I extend special thanks to the non-teaching staff in the department, particularly Mr Ravi Kumar Bharti, Mr Shubham Pandey, Mr Prakhar Kumar, Mr Manoj Kumar Rai, Mr Viplav Biswas, and Mr Akhilesh Kumar Pal.

I thank especially my parents, Mr Budhia Sahu and Mrs Bideshi Sahu, my uncle and aunt, Mr Ananda Sahu and Mrs Santilata Sahu, my sister Priyadarshini and my brother Subrat, Subhasish, Debasish for their constant support and encouragement, without which this assignment would not have been completed at all. I am grateful to my other family

members who have supported me.

I take this opportunity to sincerely acknowledge the Ministry of Human Resource Development, Government of India, for providing a PhD Fellowship for financial assistance. It is said that it is no coincidence whom you meet on the journey. I have had the fortune to meet many wonderful people on the journey, capable and ready to guide, help, and advise me. I sincerely thank all of them who contributed to helping me to see the light at the end of every scary tunnel during my PhD.

Finally, I bow in great reverence to almighty God, the most gracious, the most merciful, whose bounteous blessings enabled me to accomplish this thesis.

Date: _____

Siba Sankar Sahu

List of Tables

2.1	Summary of effect of stopword lists in text processing task	22
2.1	Summary of effect of stopword lists in text processing task	23
2.2	Summary of effect of stemming techniques in text analysis task	28
2.2	Summary of effect of stemming techniques in text analysis task	29
2.3	Summary of effect of decomposing models in text processing task	34
2.3	Summary of effect of decomposing models in text processing task	35
3.1	Shows the statistics of test collection	42
3.2	Confusion Matrix	49
4.1	MAP, R-prec and $P@10$ in Marathi (39 TDN queries)	56
4.2	MAP, R-prec and $P@10$ in Bengali (50 TDN queries)	57
4.3	MAP, R-prec and $P@10$ in Gujarati (46 TDN queries)	57
4.4	MAP, R-prec and $P@10$ in Hindi (50 TDN queries)	57
4.5	MAP, R-prec and $P@10$ without and with stopword removal in Sanskrit (50 TDN queries)	58
4.6	MAP scores of the effect of stopword in short vs. long docs in Marathi . . .	62
4.7	MAP scores of the effect of stopword in short vs. long docs in Bengali . . .	62
4.8	MAP scores of the effect of stopword in short vs. long docs in Gujarati . . .	62
4.9	MAP scores of the effect of stopwords in short vs. long docs in Hindi . . .	63
4.10	MAP scores of the effect of stopwords in short vs. long documents in Sanskrit	63
5.1	MAP scores for different methods of stopword evaluation in the Bengali (50 T queries)	74
5.2	MAP scores for different methods of stopword evaluation in the Marathi (39 T queries)	75

5.3	MAP scores for different methods of stopword evaluation in the Gujarati (50 T queries)	75
5.4	MAP scores for different methods of stopword evaluation in the Hindi (50 T queries)	76
5.5	MAP scores for different methods of stopword evaluation in the English (50 T queries)	76
5.6	Length of stopword list used for evaluation in Indian languages	80
6.1	Examples of Sanskrit masculine noun (man) declensions	84
6.2	Examples of root word generation by the light stemmer in Sanskrit	87
6.3	Examples of root word generation by the aggressive stemmer in Sanskrit	87
6.4	Statistics of Test Collection	91
6.5	Statistics of pooled documents	95
6.6	Result-summary of the verb-based stemmer	97
6.7	Result-summary of the light stemmer	97
6.8	Result-summary of the aggressive stemmer	97
6.9	Result-summary of the GRAS stemmer	97
6.10	Result summary of direct-based evaluation	98
6.11	Mean average precision (MAP) of IR models vs. stemmers in Title-only (T) queries	100
7.1	Show the Parameter of different Encoder-Decoder models	118
7.2	MAP scores of different corpus-based decompounding evaluation in Marathi (39 T queries)	119
7.3	MAP scores of different corpus-based decompounding evaluation in the Hindi (50 T queries)	120
7.4	MAP scores of different corpus-based decompounding evaluation in the Sanskrit (50 T queries)	120
7.5	MAP scores of different hybrid machine learning-based decompounding evaluation in the Marathi (39 T queries)	121
7.6	MAP scores of different hybrid machine learning-based decompounding evaluation in the Hindi (50 T queries)	121

7.7	MAP scores of different hybrid machine learning-based decomponing evaluation in the Sanskrit (50 T queries)	122
7.8	MAP scores of different deep learning-based decomponing evaluation in the Marathi (39 T queries)	123
7.9	MAP scores of different deep learning-based decomponing evaluation in the Hindi (50 T queries)	123
7.10	MAP scores of different deep learning-based decomponing evaluation in the Sanskrit (50 T queries)	124
8.1	MAP scores before and after pre-processing steps in Marathi retrieval . . .	133
1	List of suffixes stemmed by light stemmer	152
2	Additionally a list of suffixes stemmed by aggressive stemmer	152
3	List of prefixes used in different Indian languages IR	153

List of Figures

1.1	Block diagram of an information retrieval system	2
1.2	Structure and organization of thesis	16
3.1	An example of topic in Marathi	43
3.2	Marathi topic translation in English	43
3.3	An example of topic in Bengali	43
3.4	An example of a topic in Gujarati	44
3.5	Gujarati and Bengali topic translation in English	44
3.6	An example of topic in Hindi	44
3.7	Hindi topic translation in English	45
3.8	An example of a topic in English	45
3.9	An example of a topic in Sanskrit	46
3.10	Sanskrit topic translation in English	46
3.11	An example of a document in Marathi	47
4.1	A query-by-query evaluation in Marathi by BM25 model	59
4.2	A query-by-query evaluation in Bengali by BM25 model	59
4.3	A query-by-query evaluation in Gujarati by In_expB2 model	60
4.4	A query-by-query evaluation in Hindi by Hiem_LM model	60
4.5	A query-by-query evaluation in Sanskrit by In_expB2 model	61
5.1	A query by query evaluation in the Bengali by BM25 model	77
5.2	A query by query evaluation in the Marathi by TF-IDF model	77
5.3	A query by query evaluation in the Gujarati by BM25 model	78
5.4	A query by query evaluation in the Hindi by TF-IDF model	78
5.5	A query by query evaluation in the English by TF-IDF model	79

6.1	An example of a document in Sanskrit	92
6.2	English translation of the above document in Sanskrit	93
6.3	An example of a topic in Sanskrit within a topic-set	94
6.4	English translation of the above Sanskrit topic	94
6.5	MAP for different trunc ‘n’	101
6.6	A query by query evaluation in light stemming by Hiemstra_language Model	102
6.7	A query by query evaluation in aggressive stemming by Hiemstra_language Model	102
7.1	Sandhi-window (AnaNg) as the prediction target in a compound word . . .	114
7.2	Model Architecture for Sandhi Split - Stage 1	114
7.3	A basic RNN Architecture	115
7.4	A single cell LSTM Architecture	116
7.5	A single cell GRU Architecture	117
7.6	A query by query evaluation in the Marathi by In_expC2 model	125
7.7	A query by query evaluation in the Hindi by In_expC2 model	125
7.8	A query by query evaluation in the Sanskrit by BB2 model	126