

Contents

List of Figures	viii
List of Tables	x
PREFACE	xiii
1 Introduction	1
1.1 Information Retrieval	1
1.2 Input to an IR system	3
1.2.1 Document collection	3
1.2.2 Query	3
1.2.3 Relevance Judgments	3
1.3 Different steps in an IR system	4
1.3.1 Tokenization	4
1.3.2 Pre-processing	4
1.3.3 Indexing	6
1.3.4 Search and Ranking	6
1.4 Motivation	7
1.5 Dissertation Overview	8
1.6 Research Goals	9
1.7 Contribution and Impact	11
1.7.1 A study on stopwords in Indian languages IR	11
1.7.2 Creation of a text collection and a study on stemming techniques on Sanskrit	12
1.7.3 A study on word decompounding methods in Indian languages IR .	13
1.8 Structure of the Thesis	14

2	Literature Survey and Analysis of Trends	17
2.1	Stopword Removal	17
2.2	Stemming	24
2.3	Decompounding	30
2.3.1	Corpus-based decompounding methods	30
2.3.2	Machine learning-based decompounding methods	31
2.3.3	Deep learning-based decompounding methods	32
3	Background	36
3.1	Information Retrieval Framework	36
3.1.1	BM25 Model	37
3.1.2	TF-IDF Model	37
3.1.3	In_expB2 Model	38
3.1.4	In_expC2 Model	39
3.1.5	BB2 Model	39
3.1.6	IFB2 Model	40
3.1.7	DLH Model	40
3.1.8	PL2 Model	40
3.1.9	InL2 Model	40
3.1.10	Hiemstra_Language Model	40
3.2	Simulation Setup and Test Collections	41
3.2.1	Simulation Setup	41
3.2.2	Test Collections	41
3.3	Evaluation Strategy	47
3.3.1	Direct Evaluation	47
3.3.2	Indirect Evaluation	48
3.4	Evaluation Metrics	48
3.4.1	Direct Evaluation	48
3.4.2	Indirect Evaluation	50
3.5	Statistical Tests	52
3.5.1	t-test	52
3.5.2	Bonferroni correction	53

3.5.3	Confidence Interval	53
4	Effect of stopwords in Indian language IR	54
4.1	Introduction	54
4.2	Problem Formulation	54
4.3	Experimental Setup	55
4.4	Evaluation	56
4.4.1	Effect of stopword removal on retrieval	56
4.4.2	Interaction between stopwords and document length	61
4.5	Discussion	62
4.6	Summary	64
5	A study on corpus-based stopword lists in Indian language IR	65
5.1	Introduction	65
5.2	Problem Formulation	66
5.3	Methods Used	67
5.3.1	Statistical Method	67
5.3.2	Data Entropy Measure	69
5.3.3	Term Variance-based Approach	70
5.3.4	Term-based Random Sampling Approach	70
5.3.5	Aggregation Method	71
5.4	Experimental Setup	72
5.5	Evaluation	73
5.5.1	Effect of non-corpus-based and corpus-based stopword list on retrieval	73
5.5.2	Effect of corpus-based stopwords on retrieval	77
5.5.3	Length of stopword lists	80
5.6	Summary	81
6	Effect of stemming in Sanskrit IR	82
6.1	Introduction	82
6.2	Sanskrit Morphology	83
6.3	Contribution	85
6.3.1	Light Stemmer	86

6.3.2	Aggressive Stemmer	87
6.3.3	Algorithm for stemming	88
6.4	Test Collection	90
6.4.1	Data Collection	91
6.4.2	Data Processing	92
6.4.3	Topic Creation	92
6.4.4	Pool Creation	94
6.4.5	Relevance Judgments	95
6.5	Experimental Setup	96
6.6	Evaluation	96
6.6.1	Direct Evaluation	96
6.6.2	Indirect Evaluation	99
6.7	Summary	104
7	A case study on decomposing in Indian language IR	105
7.1	Introduction	105
7.2	Problem Formulation	106
7.3	Contribution	107
7.3.1	Corpus-based decomposing approaches	107
7.3.2	Machine learning-based decomposing approaches	109
7.3.3	Deep learning-based decomposing approaches	112
7.4	Experimental Setup	118
7.5	Evaluation	119
7.5.1	Effect of corpus-based decomposing on retrieval	119
7.5.2	Effect of machine learning-based decomposing on retrieval	120
7.5.3	Effect of deep learning-based decomposing on retrieval	121
7.5.4	Retrieval Effectiveness Analysis: some insights	122
7.6	Discussion	126
7.7	Summary	128
8	Discussions	129
8.1	Observations	129
8.2	Limitations	134

9	Conclusions and Future Work	135
9.1	Summary and Contribution	135
9.2	Possible Research Directions	136
9.2.1	Building text collection	136
9.2.2	Stopwords	137
9.2.3	Stemming	137
9.2.4	Decompounding	137
9.2.5	Other Text Processing Applications	137
	References	138
	Appendix	151
	List of Publications	154