MULTI-DOCUMENT ABSTRACTIVE TEXT SUMMARIZATION USING MULTI-SENTENCE COMPRESSION AND TEXT SIMPLIFICATION

RAKSHA AGARWAL

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2022

MULTI-DOCUMENT ABSTRACTIVE TEXT SUMMARIZATION USING MULTI-SENTENCE COMPRESSION AND TEXT SIMPLIFICATION

by

RAKSHA AGARWAL
DEPARTMENT OF MATHEMATICS

Submitted
in fulfillment of the requirements of the degree of
DOCTOR OF PHILOSOPHY
to the

Indian Institute of Technology Delhi
October 2022

CERTIFICATE

This is to certify that the thesis entitled, **Multi-Document Abstractive Text Summarization using Multi-Sentence Compression and Text Simplification** being submitted by **Ms. Raksha Agarwal** for the award of the degree of **Doctor of Philosophy** is a record of *bona fide* research work carried out by her in the Department of Mathematics of Indian Institute of Technology Delhi.

Ms. Raksha Agarwal has worked under my guidance and supervision. She has fulfilled the requirements for the submission of this thesis, which to my knowledge has reached the requisite standard. The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma. .

Date:

Prof. Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology Delhi
New Delhi-110016, India

ACKNOWLEDGEMENTS

I am deeply indebted to Mother Saraswati; the Goddess of Knowledge. Due to Her divine grace I have been able to carry out this research work.

This thesis would not have been possible without the support of many people. It is impossible to acknowledge all of them here, but there are few people who deserve a special mention. I am forever indebted to Prof. Niladri Chatterjee for providing me an opportunity to carry out the Ph.D. work under his supervision at IIT Delhi. It is a privilege to work with a person who can guide your research and also train you for all the aspects which Ph.D. entails. His unwavering support has been the driving force of motivation even during the most difficult times. He taught me research writing, research methodologies and most importantly he helped me out in dealing with failures in experiments that came several times during this duration.

I am extremely grateful to the Department of Mathematics and IIT Delhi authorities for providing me with excellent state of the art facilities and infrastructure to work in during my Ph.D. program. I thank Head, Department of Mathematics and DRC Chairperson for their help at various stages of my research work. I am especially thankful to my Student Research Committee members: Prof. B. S. Panda, Prof. Aparna Mehra and Prof. Sumeet Agarwal for their insightful inputs, which strengthened my research work.

I would like to acknowledge the Council of Scientific and Industrial Research, India, for providing financial support through their Shyama Prasad Mukherjee (SPM) fellowship program during my research period.

I would like to use this platform to thank the wonderful people I met during my Ph.D. and who made this entire learning experience in IIT Delhi productive and fun. I would like to thank my senior colleagues, Dr. Shradhha Chaudhary, Dr. Nidhika Yadav, Dr. Susmita Gupta, Dr. Kartikay Gupta, Dr. Anubha Goel, Dr. Puneet Parsicha, Dr. Lipsy and Dr. Sanjay Kumar for the academic as well as non-academic guidance they provided to me. I would also like to thank Dr. Priyamvada, Ms. Soniya Takshak, Ms. Pooja, Dr. Pooja Goyal, Dr. Hariom, Mr. Navnit Yadav, Dr. Pradeep Singh, Dr. Vaibhav and my juniors Ms. Shivani Chaudhary, Ms. Kushagri Tandon, Mr. Aayush Singha Roy. I am likewise thankful to all those who have directly or indirectly helped me to finish my dissertation study.

No amount of thanks and credit is sufficient enough to reflect the support I received from Mr. Pratik Sinha in this academic endeavour. I would like to extend my special thanks to Mrs. Archna Kumari. She was my pillar of strength through thick and thin of the Ph.D. journey. I would also like to thank Ms. Surbhi for the unfailing support and motivation she provided me. I thank Ms. Ambika Sharma for unconditional support whenever I approached her for the same.

All my achievements are outcomes of the silent sacrifice and patience of my family. First and foremost, I thank my father Late Mr. Kailash Kumar Agarwal and my sister Late Mrs. Rashmi Aggarwal. You left this earth too soon but I know you are looking upon me from heaven. I do not find words to thank my mother Mrs. Manju Agarwal for the love, care, and support that she has given me all my life. I would like to thank her for allowing me to progress on the path of research while not being bothered by my duties as a daughter. I wish to thank my brothers Mr. Rajiv Sanwatia and Mr. Ravi Sanwatia for being immensely supportive and having my back at all time. I also thank my brother-in-law Mr. Amit Aggarwal and sister-in-laws Mrs. Kiran Sanwatia and Mrs. Suruchi Sanwatia for encouraging me throughout this journey. I thank Yuvika, Toshank, Tishika, Aarohi, Vedika and Aarav for the joy they have brought in our life. I express my gratitude to Mr. Rajesh Agarwal and his family for supporting me selflessly. It would not have been possible for me to pursue this academic life without their help.

At last, I feel extremely thankful to almighty for giving me such a life where I have been able to make choices that I have made and pursue the goal of higher knowledge and wisdom.

Raksha Agarwal

ABSTRACT

Automatic Text Summarization systems are of high demand due to the overwhelming amount of textual information generated on the internet every minute. Historically sentence ranking approach used in Extractive text summarization has received huge attention from NLP researchers. However, different shortcomings associated with extraction has of late created a lot of enthusiasm among the researchers towards Abstractive summarization. But this task is quite challenging because it requires deep understanding of the input text and intensive natural language generation techniques for preserving the information and semantics of the input. The focus of the present research is on performing multi-document abstractive text summarization using two sentence rewriting techniques, namely, multi-sentence compression and sentence simplification.

Our research tries to find answer to several queries pertaining to Abstractive text summarization: a) Effective representation of sentences to facilitate summary generation. b) Examining the usefulness of syntactic simplification as a pre-processing step before employing summarization techniques c) Developing an appropriate sentence selection process from the simplified text both in a deterministic way, and in a fuzzy way. In particular, the latter is used in the event of query-focused summarization.

A Word Graph-based representation of sentence clusters is used for generation of multi-sentence compression. In the proposed *AlignFuse* algorithm node alignment between semantically similar words is performed to faciliate phrase stitching from similar sentences. Sentences of the input text often contain additional clauses that are not relevant enough to be included in the summary. As a consequence, DEPSYM++, a syntactic simplification algorithm based on dependency tree structure is proposed to split long sentences into simpler and smaller sentences before applying abstract generation strategies. Input text representation using a multilayer weighted sentence network along with VoteRank sentence node scoring has been used in *VoteSumm* summarization system. Its efficacy has been examined with both raw and simplified inputs. A comparative study with different sentence node ranking schemes is also conducted. In order to perform query-focused text summarization, *QSummFuzzy* has been developed in this research. It uses a combination of FuzzyRank, derived from a Mamdani fuzzy inference system, network based TextRank and keyphrase information for sentence ranking. Coreference resolution has been used for sentence rewriting.

बहु-वाक्य संक्षिप्तीकरण और वाक्य सरलीकरण से बहु-दस्तावेज़ अब्स्ट्रक्टिव पाठ सारांशीकरण

शोध प्रबंध सार

हर मिनट इंटरनेट पर भारी मात्रा में पाठ्य सूचना उत्पन्न होने के कारण स्वचालित टेक्स्ट सारांश प्रणाली की अत्यधिक मांग है। ऐतिहासिक रूप से एक्स्ट्रेक्टिव टेक्स्ट सारांशीकरण में प्रयुक्त वाक्य रैंकिंग दृष्टिकोण ने एनएलपी (NLP) शोधकर्ताओं का बहुत ध्यान आकर्षित किया है। हालांकि, एक्सट्रैक्शन से जुड़ी विभिन्न कमियों ने हाल ही में अब्स्ट्रक्टिव सारांशीकरण के प्रति शोधकर्ताओं में अति उत्साह पैदा किया है। लेकिन यह कार्य काफी चुनौतीपूर्ण है क्योंकि इसमें इनपुट टेक्स्ट की गहरी समझ और इनपुट की सूचना और शब्दार्थ को संरक्षित करने के लिए गहन प्राकृतिक भाषा निर्माण (NLG) तकनीकों की आवश्यकता होती है। वर्तमान शोध का ध्यान दो वाक्य पुनर्लेखन तकनीकों अर्थात् बहु-वाक्य संक्षिप्तीकरण और वाक्य सरलीकरण का उपयोग करके बहु-दस्तावेज़ अब्स्ट्रक्टिव पाठ सारांशीकरण करने पर है।

हमारा शोध अब्स्ट्रक्टिव टेक्स्ट सारांशीकरण से संबंधित कई प्रश्नों के उत्तर खोजने का प्रयास करता है:
अ) सारांश संरचना को सुगम करने के लिए वाक्यों का प्रभावी प्रतिनिधित्व।
ब) संक्षेपण तकनीकों को नियोजित करने से पहले पूर्व-प्रसंस्करण चरण के रूप में वाक्यात्मक सरलीकरण की उपयोगिता की जांच करना ।
स) नियतात्मक तरीके से और फ़ज़ी (Fuzzy) तरीके से एक उपयुक्त सरलीकृत वाक्य चयन प्रक्रिया का विकास करना। विशेष रूप से, बाद वाले का उपयोग क्वेरी-केंद्रित सारांशीकरण की स्थिति में किया गया है ।
बहु-वाक्य संक्षिप्तीकरण उत्पन्न करने के लिए वाक्य समूहों के वर्ड ग्राफ़ (Word Graph) -आधारित प्रतिनिधित्व का उपयोग किया गया है । प्रस्तावित AlignFuse एल्गोरिथ्म में समान वाक्यों से बने व्याक्यांशों के मेल के लिए समान शब्दार्थ वाले शब्दों के बीच नोड एकत्रीकरण किया गया है । इनपुट टेक्स्ट के वाक्यों में अक्सर अतिरिक्त खंड होते हैं जो सारांश में शामिल किए जाने के लिए पर्याप्त प्रासंगिक नहीं होते हैं। एक परिणाम के रूप में निर्भरता वृक्ष पद विच्छेदन पर आधारित , DEPSYM++, एक वाक्यात्मक सरलीकरण एल्गोरिथ्म प्रस्तावित किया गया है । इसका इस्तेमाल सारांशीकरण की रणनीतियों को लागू करने से पहले लंबे वाक्यों को सरल और छोटे वाक्यों में विभाजित करने में किया गया है । वोटसम (VoteSumm) सारांश प्रणाली में इनपुट टेक्स्ट प्रतिनिधित्व के लिए एक बहुपरत भारित वाक्य नेटवर्क का उपयोग करते हुए वोटरैंक (VoteRank) वाक्य नोड स्कोरिंग का प्रयोग किया गया है।अनिर्मित और सरल इनपुट दोनों के साथ इसकी प्रभावकारिता की जांच की गई है। विभिन्न वाक्य नोड रैंकिंग योजनाओं के साथ एक तुलनात्मक अध्ययन भी किया जाता है। क्वेरी-केंद्रित टेक्स्ट सारांशीकरण करने के लिए इस शोध में QSummFuzzy को विकसित किया गया है। यह वाक्य रैंकिंग के लिए ममदानी फ़ज़ी इनफरेंस (Mamdani Fuzzy Inference) सिस्टम से प्राप्त फ़ज़ीरैंक, नेटवर्क-आधारित टेक्स्टरैंक, की-फ्रेज़ जानकारी के संयोजन का उपयोग करता है। इसमें वाक्य पुनर्लेखन के लिए कोरफेरेंस रिज़ॉल्यूशन का उपयोग किया गया है।

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ATS	Automatic Text Summarization
AMR	Abstractive Meaning Representation
BART	Bidirectional and Auto-Regressive Transformer
BERT	Bidirectional Encoder Representations from Transformers
BLEU	BiLingual Evaluation Understudy
CNN/DM	CNN/DailyMail
CR	Compression Ratio
DEPSYM++	DEpendency Parse based SYntactic siMplification
DUC	Document Understanding Conferences
EACS	Embedding Average Cosine Similarity
EW	English Wikipedia
FKGL	Flesch-Kincaid Grade Leve
GMS	Greedy Matching Score
HAC	Hierarchical Agglomerative Clustering
ILP	Integer Linear Programming
LSA	Latent Semantic Analysis
LSTM	Long Short Term Memory Networks
METEOR	Metric for Evaluation of Translation with Explicit ORdering
ML	Machine Learning
MSC	Multi-Sentence Compression
NLP	Natural Language Processing
NLE	Natural Language Engineering
POS	Part Of Speech
PPDB	ParaPhrase DataBase
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RI	Random Indexing
RNN	Recursive Neural Network
RSG	Rich Semantic Graphs
SARI	System output Against References and Input sentence
SCU	Summary Content Unit

SEW	Simple English Wikipedia
SVD	Singular Value Decomposition
SVM	Support Vector Machines
TAC	Text Analysis Conferences
TC	Turk Corpus
TF-IDF	Term Frequency-Inverse Document Frequency
TMF	Triangular Membership Function
WSM	Word Space Model

LIST OF SYMBOLS

d	Damping Factor
r	Sentence Relatedness score in SICK Dataset
λ	Sentence Clustering Threshold for HAC
τ_{sim}	Threshold for cosine similarity between BERT word vectors
T	Total number of paths between START and END node of the Word Graph
k	Number of paths extracted from the Word Graph
M	Minimum summary sentence length
W	Summary length limit
K	Threshold for cosine similarity between sentences
α	Layering constant