

**INTERACTIVE MALAYALAM QUESTION ANSWERING SYSTEM IN
HEALTH DOMAIN USING NEURAL WORD EMBEDDING AND
HYBRID LARGE LANGUAGE MODELS**

A Thesis Submitted to
UNIVERSITY OF CALICUT
in partial fulfilment of the requirements for the award of the degree of
DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE
Under the Faculty of Science

By
LIJI S.K.

Under the guidance of
Dr. MUHAMED ILYAS P
Principal
Sullamussalam Science College, Areekode



P.G & RESEARCH DEPARTMENT OF COMPUTER SCIENCE
Sullamussalam Science College, Areekode - 673639
Affiliated to University of Calicut
Malappuram Dist., Kerala, India
February 2025



DECLARATION

I, Liji SK, hereby declare that this thesis entitled “**Interactive Malayalam Question Answering System in Health Domain Using Neural Word Embedding and Hybrid Large Language Models**“ is based on the original work done by me under the supervision of Dr. Muhamed Ilyas P, Principal, Sullamussalam Science College, Areekode, Kerala.

I confirm that,

- The work presented in this Thesis has not been submitted previously for the award of any degree either to this University or to any other University or Institution.
- I have followed the guiding principles given by the University in organizing the thesis.
- Whenever I have used materials (theoretical analysis, data, figures, and text) from other sources, I have given due credit to them by citing them in the thesis and giving their particulars in the references.



Liji S K.

Areekode
5 February 2015



Ref:

Date:

CERTIFICATE

This is to certify that the thesis entitled “**Interactive Malayalam Question Answering System in Health Domain Using Neural Word Embedding and Hybrid Large Language Models**” , submitted by **Mrs. Liji SK**, PG and Research Department of Computer Science, Sullamussalam Science College Areekode, Malappuram for the award of the degree of Doctor of Philosophy, is a record of bonafide work carried out by her under my supervision, as per the Calicut University code of academic and research ethics. The contents of this thesis have not been submitted either in part or in full, for the award of any other degree or diploma in this Institute or any other Institute or University.

The thesis is revised as per the modifications and recommendations reported by the adjudicators. Soft copy attached is the same as that of the revised copy. The thesis is submitted as such to the University of Calicut with reference to the letter number No. 30179/RESEARCH-C-ASST-1/2024/Admn Dated 01.12.2024.

Dr. Muhamed Ilyas P.
Research Guide & Principal,
Sullamussalam Science College,
Areekode, Malappuram, Kerala,
India, 673639

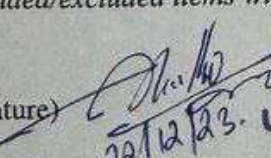
Place: Areekode

Date: 5/02/2025


**UNIVERSITY OF CALICUT
CERTIFICATE ON PLAGIARISM CHECK**

1.	Name of the research scholar	Liji S K		
2.	Title of thesis/dissertation	Interactive Malayalam Question Answering System in Health Domain Using Neural Word Embedding and Hybrid Large Language Models		
3.	Name of the supervisor	Dr.Muhamed Ilyas P		
4.	Department/Institution	PG and Research Department of Computer Science Sullamussalam Science College, Areekode,673639.		
5.	Similar content (%)identified	Introduction/ Review of literature	Materia ls and Method s	Result/ Discussion/ Summary/ Conclusion
		4%	8%	1%
	Acceptable maximum limit (%)	10	10	10
6.	Software used	iThenticate		
7.	Date of verification	28/12/2023		

*Report on plagiarism check, specifying included/excluded items with % of similarity to be attached.

Checked by (with name, designation & Signature)  **Dr. Nasirudheen. T**
Assistant Librarian
University of Calicut, Kerala.

Name and signature of the Researcher

Name & Signature of the Supervisor 

M. MUHAMMAD ILYAS P
PRINCIPAL
SULLAMUSSALAM SCIENCE COLLEGE
AREEKODE, UGRAPURAM (PO)
SULLAMUSSALAM DISTRICT, KERALA

The Doctoral Committee* has verified the report on plagiarism check with the contents of the thesis, as summarized above and appropriate measures have been taken to ensure originality of the Research accomplished herein.


Name & Signature of the HoD/HoI (Chairperson of the Doctoral Committee)
PRINCIPAL

SULLAMUSSALAM SCIENCE COLLEGE
AREEKODE, UGRAPURAM (PO)
SULLAMUSSALAM DISTRICT, KERALA

Acknowledgment

First and foremost, I express my gratitude to the **God** Almighty for providing me with the strength, knowledge, and opportunity to undertake and complete this research. This research would not have been possible without the assistance, inspiration, and support of all of my advisors, collaborators, and colleagues.

I want to express my deepest gratitude to my research supervisor **Dr. Muhamed Ilyas P**, Principal, Sullamussalam Science College, Areekode, for granting me such an excellent opportunity to conduct this research in this esteemed institution. His positive attitude and support inspired me and gave me confidence throughout my research journey

Special thanks to **Dr. Mohamed Basheer KP**, Ph.D. Nodal Officer at Sullamussalam Science College, Areekode, for his motivation and unwavering support.

I appreciate **Dr. Shameem Kappan**, HOD, Dept. of Computer Science, Sullamussalam Science College, Areekode, for his support and the resources provided.

Heartfelt thanks go to my RAC members, **Dr. Lajish VL**, Associate Professor & Head, Department of Computer Science, University of Calicut., **Dr. Binu P Chako**, Principal, Prajyoti Niketan College, Thrissur and **Dr. Vasudevan C**, Professor, Dept. of Library & Information Science University of Calicut, whose interest and advice significantly contributed to the completion of this work. I am grateful for their suggestions and support throughout the research.

I am highly thankful to **Dr. Premjith B**, Assistant Professor, Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore for the support and suggestions throughout the research work.

I thank **Dr. Rajeev R R**, Head at International Centre for Free and Open Source Software (ICFOSS), Trivandrum, Kerala for the insightful discussions and motivations.

I wholeheartedly express my thanks for the support and encouragement, I received from my research colleagues and friends. My sincere thanks to the College faculty and staff members for their cooperation and support.

Deep gratitude goes to my husband, **Ramadas V**, and my lovely son, **Aadisrey V**, for their unwavering support and endurance in completing this work. Their love and encouragement were crucial throughout this journey. I also thank my entire family for their constant support and encouragement, which played a vital role in completing this research.

Contents

1	INTRODUCTION	1
1.1	Introduction	1
1.2	Question Answering and Information Retrieval	2
1.3	Natural Language Processing	4
1.4	Features Malayalam Language	5
1.5	Motivation and Challenges for the Research Work	7
1.6	Problem Statement and Objectives	8
1.7	Main Contributions	9
1.8	Research Design and Work Flow	10
1.9	Organisation of the Thesis	12

2	Review Of Literature and Background Theory	15
2.1	Introduction	15
2.2	Review of Literature	16
2.2.1	Review on Information Retrieval and Question Answering	17
2.2.2	Review on Malayalam Information Retrieval and Question Answering	40
2.2.3	Review on Large Language Models	42
2.3	Research Gap	46
2.4	Back Ground Theory	47
2.4.1	Question Answering System	47
2.4.2	Architecture of a Question Answering System	51
2.4.3	Neural Network Models for NLP	54
2.4.4	Large Language Models	56
2.5	Conclusion	57
3	Corpus Creation, Pre-processing and Modelling	58

3.1	Introduction	58
3.2	Data Collection and Corpus Creation	59
3.3	Data Annotation for Question Answering	61
3.4	Data Pre-processing of Queries and Documents	65
3.4.1	Tokenization	66
3.4.2	Stop Word Removal	66
3.4.3	Stemming and Lemmatization	67
3.4.4	Bag of Words	68
3.5	Vector Space Representation and Document Modelling	68
3.5.1	Term Frequency (TF) Modelling	71
3.5.2	Inverse Document Frequency (IDF) Modelling	71
3.5.3	Term Frequency-Inverse Document Frequency (TF-IDF) Modelling	72
3.6	Conclusion	72
4	Word Embedding Based Question Answering Model	74

4.1	Introduction	74
4.2	Word Embedding	75
4.2.1	Word2Vec	77
4.3	Architecture of Word Embedding Based Question Answering Model	84
4.3.1	Pre-processing of Queries and Documents	84
4.3.2	Feature Vector Creation and Semantic Modeling	86
4.3.3	Document Ranking and Answer Retrieval	86
4.4	Algorithm for Word Embedding Based Question Answering	88
4.5	Result and Discussion	89
4.6	Case study: COVID-19 Question Answering System Using Word Embedding	91
4.7	Performance Evaluation and Analysis	92
4.8	Conclusion	94
5	BERT Based Models for Question Answering	95

5.1	Introduction	95
5.2	Bidirectional Encoder Representation from Transformers (BERT) Model	96
5.2.1	Model Architecture	96
5.2.2	Self Attention	98
5.2.3	Different BERT Models	100
5.3	Question Answering using Different BERT Models	101
5.4	BERT Models with Reader-Retriever Pipeline for Question Answering	103
5.5	Result and Discussion	107
5.6	Conclusion	111
6	BioQA BERT- Proposed BERT Model for Question Answering Using LoRA and LLaMA	112
6.1	Introduction	112
6.2	LoRa and LLaMA for Question Answering	112
6.2.1	LoRA-Low-Rank Adaptation Algorithm	113

6.2.2	LLaMA-Large Language Model Meta AI	113
6.2.3	LoRA and LLaMA for Question Answering	114
6.3	BioQA BERT- Proposed BERT model with LoRA and LLaMA	114
6.4	Result and Discussion	116
6.5	Conclusion	119
7	Performance Analysis and Discussion of Various Proposed Models	120
7.1	Introduction	120
7.2	Comparative Analysis	121
7.3	Conclusion	123
8	Conclusion	124
9	Recommendations	126
9.1	Introduction	126
9.2	Recommendations for Future Research	126

9.3 Conclusion	128
--------------------------	-----

List of Figures

1.1	Work Flow of the System	11
1.2	Thesis Outline	13
2.1	General Architecture of Question Answering System	52
3.1	Sample Collected Data Set	59
3.2	Sample Health Data	60
3.3	Data Annotation Using Haystack	62
3.4	Data Annotation Process	62
3.5	Annotated Data in JSON Format	63
3.6	Annotated Data in Excel Format	64

4.1	Sample Word Embedding	76
4.2	CBOW-Architecture	79
4.3	Skip-Gram-Architecture	82
4.4	Block Diagram of Word Embedding Based Question Answering System . .	85
4.5	Word Embedding for Queries and Documents	87
4.6	Word Embedding for MQuAD	88
4.7	Sample Queries and Retrieved Results	90
4.8	Sample Queries and Retrieved Results	91
5.1	Encoder Decoder Stacks in Transformers	97
5.2	Transformer Model Architecture	98
5.3	Working of Question Answering Process in BERT	102
5.4	Block Diagram of Customized BERT Model	104
5.5	Sample Result from BERT Model	107
5.6	F1 Score of Different BERT Models	109

5.7	Performance Analysis of the System With and Without Reader-Retriever Pipeline	110
6.1	Architecture of Proposed BioQA-BERT	115
6.2	Result from BERT / Input to LLaMA Model	117
6.3	Final Results from BioQA-BERT Model	118

List of Tables

3.1	Description of MQuAD Corpus	65
4.1	Performance Analysis of the System	93
4.2	Features of the Model	94
5.1	Features of Different BERT Models	101
5.2	Performance Analysis of Different BERT Models	108
5.3	F1 Score of Different BERT Models	108
5.4	System Specifications	110
6.1	Performance Analysis of Different proposed Models	118
6.2	System Specifications	119

7.1 Performance Analysis of Different Question Answering Models . . 122

List Of Abbreviations

ACR Anaphora Catephora Resolution

AE Answer Equivalence

AI Artificial Intelligence

ALBERT A Lite Bidirectional Encoder Representations from Transformers

ASR Automatic Speech Recognition

BEM BERT Matching Measure

BERT Bidirectional Encoder Representations from Transformers

Bi-LSTM Bidirectional Long Short-Term Memory

Bi-RNN Bidirectional Recurrent Neural Network

BioQA-BERT Bio Question Answering BERT

CBOW Continuous Bag Of Word

CL Computational Linguistics

CLIR Cross-Language Information Retrieval

CoCUAD Covid-19 Question Answering Data set

Co-reference Resolution CR

CNN Convolutional Neural Networks

CS Computer Science

CSV Comma Separated Values

CQA Community Question Answering

CQA Computational Question Answering

DEHG Document Entity Heterogeneous Graph

DistilBERT Distil Bidirectional Encoder Representations from Transformers

DL Deep Learning

EMR Electronic Medical Records

ERINE Enhanced Representation through Knowledge Integration

FN False Negative

FP False Positive

GPT Generative Pre-trained Transformers

GLM General Language Model

GNN Graph Neural Networks

GRU Gated Recurrent Unit

HAR Hierarchical Attention Retrieval

IDF Inverse Document Frequency

iNLTK Natural Language Toolkit for Indic Languages

IE Information Extraction

IR Information Retrieval

JSON Java Script Object Notation

KB Knowledge Base

LLaMA Large Language Meta AI

LLM Large Language Model

LM Large Language

LaMDA Language Model for Dialogue Applications

LoRA Low Rank Adaptation

LODQA Linked Open Data Quality Assessment

LSTM Long Short Term Memory

MAP Mean Average Precision

MCQA Multiple Choice Question Answering

ML Machine Learning

MLM Masked Language Model

MLP Multi Layer Perceptron

MRR Mean Reciprocal Rank

MT Machine Translation

MQuAD Malayalam Question Answering Data set

NER Named Entity Recognition

NDCG Normalized Discounted Cumulative Gain

NLP Natural Language Processing

NLTK Natural Language Tool Kit

NLU Natural Language Understanding

NN Neural Networks

NSP Next Sentence Prediction

OWL Ontology Web Language

POS Part Of Speech

PRF Pseudo Relevance Feedback

QA Question Answering

QANER Question Answering using NER

QAN Question Answering cross attention Network

QAS Question Answering System

QE Query Expansion

RNN Recurrent Neural Network

RNP Relation Network Parser

RoBERTa Robustly Optimized Bidirectional Encoder Representations from Transformers

SDG Stochastic Gradient Descent

SDM Sequential Dependence Model

SGD Stochastic Gradient Descent

SK Skip-Gram

SM Statistical Modeling

StructBERT Struct Bidirectional Encoder Representations from Transformers

SQuAD Stanford Question Answering Data set

TDIL Technology Development for Indian Languages

TF Term Frequency

TF-IDF Term Frequency Inverse Document Frequency

TREQS TRanslate Edit Model for Question to SQL

TP True Positive

URL Uniform Resource Locator

VSM Vector Space Model

WebGLM Web General Language Model

WOL Web Ontology Language

WSD Word Sense Disambiguation

Abstract

Human-computer interaction has been a mission for mankind since the emergence of Artificial Intelligence (AI) and Machine Learning (ML) techniques. Currently, Information Retrieval (IR) and Question Answering (QA) stand out as the most blooming research areas. A vast volume of data is scattered across the web, with a large number of users simultaneously seeking information. So computers must comprehend and analyze natural languages for accurate and effective communication. Natural Language Processing (NLP) and Natural Language Understanding (NLU) techniques address the challenges in man-machine interaction. Both Information Retrieval (IR) and Question Answering (QA) systems act as interfaces for communication with computers. In Information Retrieval, a set of documents is fetched from an extensive document collection. At the same time, Question Answering aims to provide the most relevant and appropriate answers to users' queries.

This study aims to create an interactive Malayalam Question Answering System (QAS) in the health domain that can effectively answer users' queries related to their health problems and wellness. To do this, here employ advanced NLP techniques such as Word Embedding, hybrid transformer-based models, and Large Language Models (LLM). In the first phase, Word Embedding was used to capture the semantic relationships be-

tween words, with Continuous Bag Of Word (CBOW) and Skip-gram models. Then, fine-tuned different Bidirectional Encoder Representations from Transformers (BERT) models with health data sets for the Question Answering process. Secondly, a new hybrid BERT model was introduced by combining a Reader-Retriever pipeline architecture with different BERT models. Finally, proposed a new model Bio Question Answering BERT (BioQA-BERT) that uses the Low Rank Adaptation (LoRA) fine-tuning algorithm and the latest Large Language Model (LLM)- Large Language Meta AI (LLaMA) with the optimized BERT model RoBERTa.

In the initial phase, a Word Embedding-based Question Answering model was introduced, with an annotated SQuAD format Malayalam Question Answering Data set (MQuAD), SQuAD means Stand-ford Question Answering Data set. In this study employed two-Word Embedding algorithms - CBOW and Skip-Gram. Notably, the Skip-Gram algorithm exhibited the highest accuracy when applied to the MQuAD corpus. The primary objective of this system is to bridge the gap between human language understanding and machine reasoning. By exploiting Word Embedding, the system provides precise and relevant answers to user's queries. One pivotal advancement that has significantly contributed to the effectiveness of the Question Answering system is the incorporation of Word Embedding techniques, specifically demonstrated through the superior performance of the Skip-Gram algorithm on the MQuAD corpus.

A detailed analysis of various BERT models for Question Answering was conducted. Different pre-trained BERT models, namely A Lite Bidirectional Encoder Representations from Transformers (ALBERT), Distil Bidirectional Encoder Representations from Transformers (DistilBERT), Struct Bidirectional Encoder Representations from Transformers (StructBERT), and Robustly Optimized Bidirectional Encoder Representations

from Transformers (RoBERTa), were utilized and fine-tuned using the MQuAD corpus. A hybrid Reader-Retriever pipeline architecture was incorporated with different BERT models. The Question Answering model, BioQA-BERT was implemented by integrating the Low Rank Adaptation LoRA algorithm and LLaMA with the optimized BERT model RoBERTa. Experimental results indicated a notable increase in the F1 score after introducing LoRA and LLaMA with BERT models. In conclusion, while this study provides significant insights into the state and challenges of Malayalam Question Answering, especially in the health domain, it is constrained by the inherent complexities of the language and the limited precedent in the field.

In conclusion, this research constitutes a substantial contribution to understanding the current state and challenges within the domain of Malayalam Question Answering, particularly in the context of healthcare. The comprehensive exploration of these challenges illuminates the intricacies intrinsic to the Malayalam language, offering a subtle perspective on the difficulties encountered in the development of robust Question Answering systems for this linguistic environment. The intricate nature of the Malayalam language poses obstacles to achieving optimal performance, necessitating sophisticated solutions attuned to its linguistic idiosyncrasies. Additionally, the limited precedent in the field of Malayalam Question Answering introduces an extra layer of complexity, highlighting the pioneering nature of this research venture. As the landscape evolves, future research endeavors will benefit from addressing these challenges, thereby building upon the foundational insights derived from this study and contributing to the continual advancement of Malayalam Question Answering systems, particularly the health domain.

Keywords: Question Answering System, Health Question Answering, Natural Language Processing, Information Retrieval, Word Embedding, Transformers, BERT, Large Language Models.

സംഗ്രഹം

ആർട്ടിഫിഷ്യൽ ഇന്റലിജൻസ്, മെഷീൻ ലേണിംഗ് ടെക്നിക്കുകൾ എന്നിവയുടെ ആവിർഭാവം മുതൽ മനുഷ്യനും, കമ്പ്യൂട്ടറും തമ്മിലുള്ള ആശയവിനിമയം മനുഷ്യരാശിക്ക് ഒരു സ്വപ്നം ആണ്. നിലവിൽ, വിവര വീണ്ടെടുക്കൽ, ചോദ്യോത്തരങ്ങൾ എന്നിവയാണ് ഏറ്റവും സജീവമായ ഗവേഷണ മേഖലകൾ. ധാരാളം ഉപയോക്താക്കൾ ഒരേസമയം വിവരങ്ങൾ തേടുന്നതിനാലും ധാരാളം ഡാറ്റ വെബിലുടനീളം ചിതറിക്കിടക്കുന്നതിനാലും. കൃത്യവും ഫലപ്രദവുമായ ആശയവിനിമയത്തിനായി കമ്പ്യൂട്ടറുകൾ സ്വാഭാവിക ഭാഷകൾ മനസ്സിലാക്കുകയും വിശകലനം ചെയ്യുകയും വേണം. നാച്ചുറൽ ലാംഗ്വേജ് പ്രോസസ്സിംഗ് (എൻഎൽപി) ടെക്നിക്കുകളും നാച്ചുറൽ ലാംഗ്വേജ് അനലിസിസ് ഡിംഗ് (എൻഎൽയു) ടെക്നിക്കുകളും മനുഷ്യ - മെഷീൻ ആശയവിനിമയത്തിലെ വെല്ലുവിളികളെ അഭിസംബോധന ചെയ്യുന്നു. വിവര വീണ്ടെടുക്കൽ (ഐആർ), ചോദ്യോത്തര (ക്യുഎ) സംവിധാനങ്ങൾ കമ്പ്യൂട്ടറുകളുമായുള്ള ആശയവിനിമയത്തിനുള്ള ഇന്റർഫേസുകളായി പ്രവർത്തിക്കുന്നു. വിവര വീണ്ടെടുക്കലിൽ, ഒരു വലിയ ഡോക്യുമെന്റ് ശേഖരത്തിൽ നിന്ന് ഒരു കൂട്ടം ഡോക്യുമെന്റുകൾ ലഭിക്കുന്നു, അതേസമയം ചോദ്യ ഉത്തര സംവിധാനങ്ങൾ ഉപയോക്താവിന്റെ ചോദ്യങ്ങൾക്ക് ഏറ്റവും പ്രസക്തവും ഉചിതവുമായ ഉത്തരങ്ങൾ നൽകാൻ ലക്ഷ്യമിടുന്നു.

ആരോഗ്യമേഖലയിലെ സംവേദനാത്മക മലയാളം ചോദ്യോത്തര സംവിധാനത്തിന്റെ വിപുലമായ അന്വേഷണത്തിലാണ് ഈ പഠനം ശ്രദ്ധ കേന്ദ്രീകരിക്കുന്നത്. ഉപയോക്താക്കളുടെ ആരോഗ്യവും ക്ഷേമവും സംബന്ധിച്ച ചോദ്യങ്ങളെ ഫലപ്രദമായി അഭിസംബോധന ചെയ്യാൻ കഴിവുള്ള ശക്തമായ സംവിധാനം വികസിപ്പിക്കുക എന്നതാണ് ലക്ഷ്യം. ഈ ലക്ഷ്യം നേടുന്നതിന്, വേഡ് എംബഡിംഗ്, ഹൈബ്രിഡ് ട്രാൻസ്ഫോമർ അഡിഷ്ണൽ മോഡലുകൾ പോലുള്ള നൂതന എൻഎൽപി ടെക്നിക്കുകൾ ഉപയോഗിക്കുന്നു. തുടക്കത്തിൽ, തുടർച്ചയായ ബാഗ് ഓഫ് വേഡ്സ് (സിബിഒഡബ്ല്യു), സ്കിപ്പ്-ഗ്രാം മോഡലുകൾ എന്നിവ ഉപയോഗിച്ച് വാക്കുകൾ തമ്മിലുള്ള സെമാന്റിക് ബന്ധങ്ങൾ കണ്ടെത്താൻ വേഡ് എംബഡിംഗ് ഉപയോഗിച്ചു. തുടർന്ന്, ട്രാൻസ്ഫോമർ ആർക്കിടെക്ചറിന്റെ ശക്തി പ്രയോജനപ്പെടുത്തി, ക്യൂഎ ടാസ്കിനായി ബെർട്ട് മോഡൽ നന്നായി ക്രമീകരിച്ചു. കൂടാതെ, ബെർട്ട് മോഡലിൽ വേഡ് എംബഡിംഗ്, റീഡർ-റിട്രീവർ പൈപ്പ് ലൈൻ ആർക്കിടെക്ചർ എന്നിവ ഉൾപ്പെടുത്തി ഞങ്ങൾ ഒരു പുതിയ ഹൈബ്രിഡ് ബെർട്ട് മോഡൽ ബയോക്യൂഎ-ബെർട്ട് അവതരിപ്പിക്കുന്നു. കൂടാതെ, ലോറ അൽഗോരിതവും

ഏറ്റവും പുതിയ ലാർജ്ജ് ലാംഗ്വേജ് മോഡലായ എൽഎൽഎഎംഎ എന്നിവ ബെർട്ട് മോഡലുമായി സംയോജിപ്പിച്ചു.

ആദ്യ ഘട്ടത്തിൽ, വേഡ് എംബഡിംഗ് അടിസ്ഥാനമാക്കിയുള്ള ചോദ്യോത്തര സംവിധാനം നിർദ്ദേശിച്ചു. വ്യാഖ്യാനിച്ച, മലയാളം ചോദ്യോത്തര ഡാറ്റാ സെറ്റ് ഉപയോഗിച്ച് ഒരു വേഡ് എംബഡിംഗ് മോഡൽ വികസിപ്പിച്ചു. വേഡ് എംബഡിംഗ് അൽഗോരിതം സീബോവ് , സ്കിപ്ഗ്രാം എന്നിവ ഈ പഠനത്തിനായി ഉപയോഗിക്കുന്നു. അവയിൽ, സ്കിപ്ഗ്രാം അൽഗോരിതം ഞങ്ങളുടെ ഡാറ്റാ സെറ്റിൽ ഏറ്റവും ഉയർന്ന കൃത്യത പ്രകടമാക്കി. മനുഷ്യ ഭാഷാ ധാരണയും യന്ത്ര യുക്തിയും തമ്മിലുള്ള വിഭജനം കുറയ്ക്കാൻ ഈ സംവിധാനം ലക്ഷ്യമിടുന്നു . ഉപയോക്താക്കളുടെ ചോദ്യങ്ങൾക്ക് കൃത്യവും പ്രസക്തവുമായ ഉത്തരങ്ങൾ നൽകാൻ സിസ്റ്റത്തിന് കഴിയും.ചോദ്യോത്തര സംവിധാനത്തിന്റെ ഫലപ്രാപ്തിക്ക് ഗണ്യമായ സംഭാവന നൽകിയ ഒരു പ്രധാന മുന്നേറ്റം വേഡ് എംബഡിംഗ് ഉൾപ്പെടുത്തിയതാണ്.

അടുത്ത തലത്തിൽ, ചോദ്യോത്തരത്തിനായി വ്യത്യസ്ത ബെർട്ട് മോഡലുകളുടെ വിശദമായ വിശകലനം നടത്തി. ഇവിടെ ഞങ്ങൾ വ്യത്യസ്ത മുൻകൂട്ടി പരിശീലനം ലഭിച്ച ബെർട്ട് മോഡലുകൾ ഉപയോഗിച്ചു - ആൽബർട്ട്, ഡിസ്റ്റിൽബെർട്ട്, സ്ട്രക്റ്റ്ബെർട്ട്, റോബെർട്ട്. തുടർന്ന് ഹെൽത്ത് ഡൊമെയ്നിൽ സജ്ജീകരിച്ച മലയാളം ചോദ്യോത്തര ഡാറ്റാ ഞങ്ങളുടെ അനോട്ടേറ്റഡ് സ്കാഡ് ഫോർമാറ്റ് ഉപയോഗിച്ച് അവ നന്നായി ക്രമീകരിച്ചു. ഇവിടെ ഞങ്ങൾ ഓരോ ബെർട്ട് മോഡലിലും വേഡ് എംബഡിംഗ്, ഒരു ഹൈബ്രിഡ് റീഡർ-റിട്രീവർ പൈപ്പ് ലൈൻ ആർക്കിടെക്ചർ എന്നിവ ഉൾപ്പെടുത്തി. ലോറാക് അഡാപ്റ്റേഷൻ (ലോറ) അൽഗോരിതം, ബെർട്ട് മോഡലിനൊപ്പം എൽഎൽഎഎംഎ എന്നിവ ഉൾപ്പെടുത്തിക്കൊണ്ട് ഒരു ചോദ്യോത്തര മോഡൽ ബയോക്യൂഎ-ബെർട്ട് കണ്ടു പിടിച്ചു. ബെർട്ട് മോഡലുകളുമായി ലോറ, എൽഎൽഎഎംഎ എന്നിവ അവതരിപ്പിച്ചതിനുശേഷം എഫ്1 സ്കോറിൽ വർദ്ധനവ് പരീക്ഷണ ഫലങ്ങൾ സൂചിപ്പിച്ചു. ഉപസംഹാരമായി, മലയാളം ചോദ്യോത്തരത്തിന്റെ അവസ്ഥയെക്കുറിച്ചും വെല്ലുവിളികളെക്കുറിച്ചും, പ്രത്യേകിച്ച് ആരോഗ്യമേഖലയിൽ, കാര്യമായ ഉൾക്കാഴ്ചകൾ ഈ പഠനം നൽകുന്നുണ്ടെങ്കിലും, അത് ഭാഷയുടെ അന്തർലീനമായ സങ്കീർണ്ണതകളാലും ഈ മേഖലയിലെ പരിമിതമായ കീഴ്വരകളാലും ഇത് ബന്ധപ്പെട്ടിരിക്കുന്നു.

Chapter 1

INTRODUCTION

1.1 Introduction

Communicating with computers using natural languages has been a longstanding mission of mankind since the advent of Artificial Intelligence (AI) and Machine Learning (ML) methods. Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) are some of the techniques that enable this goal. In recent times, Large Language Models (LLM)s [18] have gained significant impact on this. Some of the most common NLP applications include Text Summarization, Machine Translation (MT), Sentiment Analysis, Automatic Speech Recognition (ASR), Information Retrieval (IR), and Question Answering(QA), etc. Among them, Information Retrieval (IR) and Question Answering (QA) are two of the most thriving research areas.

A huge volume of data is scattered across the web and a large number of people are engaged in information access simultaneously. Computers must understand and analyze natural languages for accurate and effective communication. NLP and NLU techniques are used to address the challenges related to man-machine interaction, and Question Answering Systems (QAS) are used for human-computer interaction. Information Retrieval (IR) system aims to retrieve a collection of document sets. On the other hand, the Question Answering System (QAS) focuses on retrieving precise answers to users' queries[62].

Question Answering employs a diverse set of methods to enable computers to understand and respond to users' queries effectively. NLP techniques, including Tokenization, Part-Of-Speech Tagging, Stemming, and Lemmatization, provide the foundation for language comprehension by breaking down and analyzing text. Word Embedding techniques, such as CBOW, and Skip-Gram, are used to represent words as vectors, capturing semantic relationships and enhancing language understanding. Deep learning techniques, such as Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), and advanced Transformer models like BERT and LLM, revolutionize Question Answering by enabling neural networks to grasp intricate language structures and context, achieving state-of-the-art performance. Combining these methods allows Question Answering Systems to navigate the complexities of language, providing accurate and contextually relevant responses to users' queries.

1.2 Question Answering and Information Retrieval

Question Answering (QA) and Information Retrieval (IR) are two foundational components of Natural Language Processing (NLP) that play a crucial role in facilitating

effective communication. Information Retrieval involves the process of retrieving pertinent documents from a vast collection. Techniques such as indexing, ranking, and retrieval models are employed to align user queries with relevant documents [92]. IR systems are commonly deployed in search engines, providing users with a list of documents that match their queries.

In contrast, Question Answering goes beyond the retrieval of documents, focusing on providing specific answers to users' queries. Unlike Information Retrieval, Question Answering Systems aim to comprehend the intent behind questions and furnish concise and accurate responses. These systems utilize various techniques, including Natural Language Understanding (NLU), Information Extraction (IE), and Knowledge Representation [115], to analyze and process questions to retrieve the most appropriate answers. QA systems cater to a broad spectrum of inquiries, from simple fact-based questions to complex queries demanding deeper understanding and reasoning.

Both Information Retrieval and Question Answering Systems are pivotal in facilitating information access and interaction with computers. While Information Retrieval offers a broader set of relevant documents, QA systems strive to directly answer users' specific questions[93]. These technologies find extensive application in search engines, virtual assistants, customer support systems, and other contexts where users require swift and accurate access to information. Advances in machine learning and deep learning have significantly enhanced the performance of both IR and QA systems, enabling more effective and precise answer retrieval capabilities.

1.3 Natural Language Processing

Natural Language Processing (NLP) is the branch of Artificial Intelligence (AI) and Computer Science (CS), that has undergone dramatic changes in its applications and technologies. It deals with how efficiently computers can understand text and speech in the same manner as humans. Natural Language Processing combines the features of Computational Linguistics (CL), Statistical Modeling (SM), Machine Learning (ML), and Deep Learning (DL) techniques. These capabilities enable computers to identify, process, and understand human languages in text or speech, grasping the meaning and intent of human thought. Various NLP techniques are used to break down natural languages to make them in computer-understandable format[82]. Some of the important NLP tasks include:

- Part Of Speech (POS) tagging: This process identifies the part of speech of a given text based on its context.
- Named Entity Recognition (NER): This technique identifies words or phrases as meaningful entities.
- Word Sense Disambiguation (WSD): This approach selects the meaning of a word with multiple meanings [111].
- Co-reference Resolution (CR): This determines if two words refer to the same entity[139].
- Finally, Natural Language Understanding (NLU) is the process of infusing structured information into natural languages [53].

1.4 Features Malayalam Language

Malayalam, a Dravidian language, recognized as both the official and spoken language of Kerala, finds its place within the southern group of languages. The linguistic tapestry of Malayalam is intricately woven, drawing influences from languages such as Tamil and Sanskrit [109]. Within this rich linguistic mosaic, which encompasses nouns, verbs, and adjectives, one encounters a profound inflection of words. A nested complexity characterizes this inflection, adding layers to identifying morphological features.

Malayalam stands out for its linguistic nature marked by intricacy and agglutination. In agglutinative languages like Malayalam, the formation of words involves combining a sequence of morphemes [101], introducing a higher level of intricacy to its structural composition [135]. Some notable features of the Malayalam language are discussed below:

- **Agglutinative Structure:**

Malayalam is an agglutinative language, meaning that it forms words by combining root words with various affixes. This feature can impact tokenization and morphological analysis during NLP tasks[101].

- **Complex Morphology:**

Malayalam exhibits complex morphological variations, especially in verb conjugations and noun inflections. NLP systems need to handle the rich morphological structure to accurately understand and generate text[135].

- **Vowel Harmony:**

Malayalam employs vowel harmony, where the vowels in a word tend to harmonize in terms of nasality. This aspect should be considered during phonetic analysis and

pronunciation modeling.

- Script and Orthography:

Malayalam has its script, which is an abugida (a writing system where consonant-vowel sequences are written as a unit). Handling this script in NLP involves addressing the unique characteristics of the Malayalam writing system.

- Sandhi Rules:

Malayalam employs sandhi rules, which involve changes in the sounds of adjacent words when they are combined. NLP systems need to account for these rules during text processing to maintain linguistic accuracy.

- Politeness Levels:

Malayalam, like many other languages, has distinct levels of politeness in speech. Understanding and incorporating these nuances is crucial in sentiment analysis, chatbots, and other applications that involve human-computer interaction.

- Code-Switching:

Code-switching is common in Malayalam, where speakers seamlessly switch between Malayalam and English or other languages within a conversation. NLP models should be designed to handle multilingual contexts and variations.

- Idiomatic Expressions:

Malayalam has a rich collection of idiomatic expressions and colloquialisms. Proper understanding of these expressions is essential for accurate sentiment analysis and natural language understanding.

- Dialectical Variations:

Malayalam exhibits dialectical variations across different regions of Kerala. NLP models should be robust enough to recognize and handle these variations for applications like speech recognition and machine translation[83].

- Lack of Standardized Resources:

Unlike some major languages, Malayalam may have limited standardized resources, such as large annotated corpora, which can pose challenges for training and evaluating NLP models.

Understanding and accommodating these features in NLP systems designed for Malayalam is crucial for developing effective applications that cater to the linguistic nuances of Malayalam speakers. [83]. This linguistic complexity poses substantial language processing challenges, especially in constructing a Question Answering System explicitly for the Malayalam language. The endeavor to embark upon such a task is laborious and presents a formidable challenge.

1.5 Motivation and Challenges for the Research Work

Currently, a limited number of research studies have been conducted on Question Answering in the Malayalam language, particularly at the semantic level. Addressing the specific challenge of semantic Malayalam Question-answering within the healthcare domain is essential and holds significant influence. While substantial research exists for Question Answering in English and some other global languages, it is crucial to recognize the linguistic diversity in India. Not all residents are proficient in English. Therefore, creating Question Answering Systems (QAS) in regional Indian languages such as Malayalam and Tamil can have a profound impact and is highly relevant.

In Malayalam, the majority of words, including nouns, verbs, adjectives, and adverbs, undergo heavy inflections, providing information such as a person, number, tense, and

mood, respectively. These inflections are often nested, adding a layer of complexity to the identification of morphological features. Malayalam boasts a productive morphology that allows the creation of complex words, introducing a high level of ambiguity. Due to the complexity of the Malayalam language, implementing a semantic Question Answering System (QAS) is a tedious and time-consuming task. Existing QAS designed for other languages may not be suitable for Malayalam due to its intricate nature, making language processing more challenging.

1.6 Problem Statement and Objectives

This research work is dedicated to developing a specialized Malayalam Question Answering System (QAS) designed for the health domain, with the primary objective of effectively addressing users' health-related inquiries. In this ambitious undertaking, integrate advanced semantic-level processing techniques into both queries and documents, aiming to provide users with accurate and contextually relevant answers. To accomplish this challenging task, used various NLP techniques, Word Embedding algorithms, and hybrid BERT models by integrating LoRA, an efficient fine-tuning algorithm, and the latest Large Language Model, LLaMA for enhanced performance.

The main objectives of this study are

- To collect Malayalam health-related documents and annotate them to create a standard Malayalam Question Answering Data set- MQaAD.
- To generate Word Embedding for the collected health data set and develop a Question Answering model.

- To re-train and optimize various BERT models, integrate a Reader-Retriever pipeline architecture with BERT models, and then compare their performance.
- To develop a new model, the BioQA-BERT model, by integrating the LoRA algorithm and LLaMA with the optimized BERT model.

1.7 Main Contributions

The primary contributions of this research work are multifaceted, encompassing various stages of data preparation, model development, and optimization, all aimed at enhancing the performance of Malayalam Question Answering Systems (QAS).

The key contributions of this research work include:

- Development and annotation of a Malayalam Question Answering Data set - MQuAD in the standard SQuAD format.
- Creation of Word Embeddings and development of a Question-Answering (QA) model.
- Re-train and fine-tune different BERT models by integrating Reader Retriever pipeline architecture, then compare the performance.
- Development of a new model BioQA-BERT by incorporating LoRA and LLaMA with the optimized BERT model RoBERTa.

This research significantly advances Malayalam Question Answering Systems by creating and annotating a dedicated data set, generating customized Word Embeddings,

and investigating BERT models with the Reader-Retriever pipeline. The introduction of the BioQA-BERT model goes beyond conventional approaches by integrating innovative attention mechanisms, enhancing linguistic role-aware processing. Collectively, these contributions improve the efficacy of natural language understanding and processing within the Malayalam language context.

1.8 Research Design and Work Flow

The research design and the overall workflow of this system encompass a multi-stage process, carefully designed to ensure efficiency and accuracy in handling complex Question Answering tasks. It unfolds through several key stages, each contributing to the system's robustness and effectiveness. Figure 1.1 serves as a visual guide, encapsulating the intricate stages of the workflow. It provides a clear depiction of the system's architecture, showcasing the interconnected processes from query and document pre-processing to the various levels of Question Answering tasks and ultimate answer retrieval.

- **Query and Document Pre-processing:** The journey begins with the pre-processing of both queries and documents. Raw textual data undergoes a series of transformations to ensure uniformity, cleanliness, and optimal representation for subsequent stages. This involves steps such as tokenization, stemming, and stop word removal. The goal is to create a refined and standardized dataset that serves as the foundation for subsequent analyses.
- **Feature Vector Creation and Modeling:** Once the pre-processing is complete, the system moves into the modeling phase. Here, various Natural Language Processing (NLP) techniques and Neural Network models are used. These models are trained

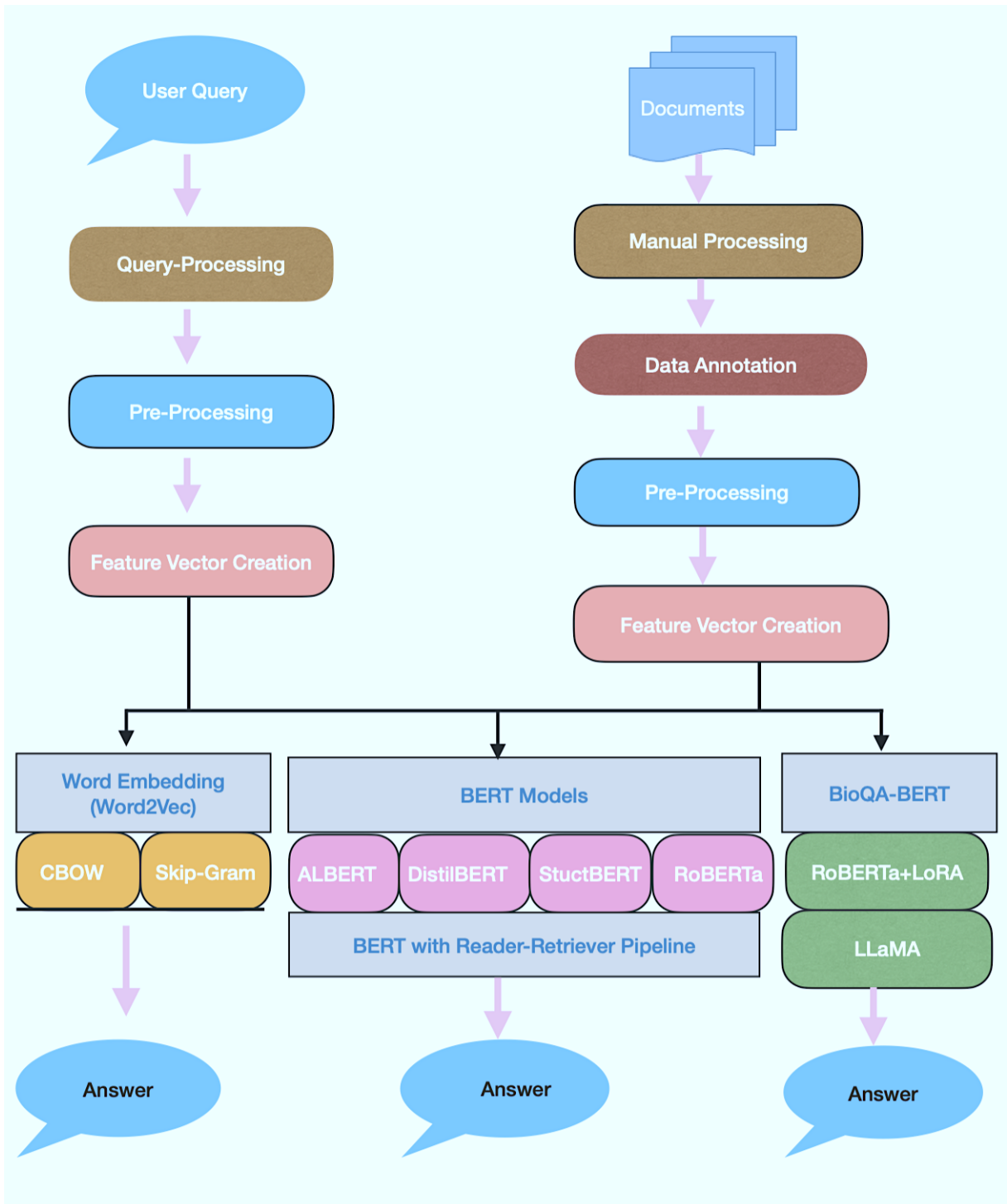


Figure 1.1: Work Flow of the System

on the pre-processed data to understand the intricate patterns, relationships, and semantic relationships within the text.

- **Question Answering Tasks:** The system excels in handling different levels of Question Answering tasks. This involves a hierarchy of challenges, ranging from straightforward factual questions to more complex inquiries that require deeper comprehension and reasoning.
- **Answer Retrieval:** An integral part of the system is the answer retrieval stage. This involves the systematic retrieval of relevant information from the pre-processed documents based on the input query. The system employs advanced algorithms and techniques to pinpoint the most accurate and contextually appropriate answers. The overall workflow of the entire system is depicted in Figure 1.1.

There are three levels of implementation: Word Embedding-based Question Answering, BERT-based model, and hybrid BERT model with LoRA and LLaMA.

The workflow of this system shows a meticulous approach to handling complex Question Answering tasks. Through stages of pre-processing, document modeling, and answer retrieval, the system navigates the intricacies of natural language, offering a comprehensive solution for extracting meaningful answers from a corpus of documents.

1.9 Organisation of the Thesis

This section presents the overview of the entire thesis and an insight into the work contributions. The overall structure of the thesis is shown in Figure 1.2.

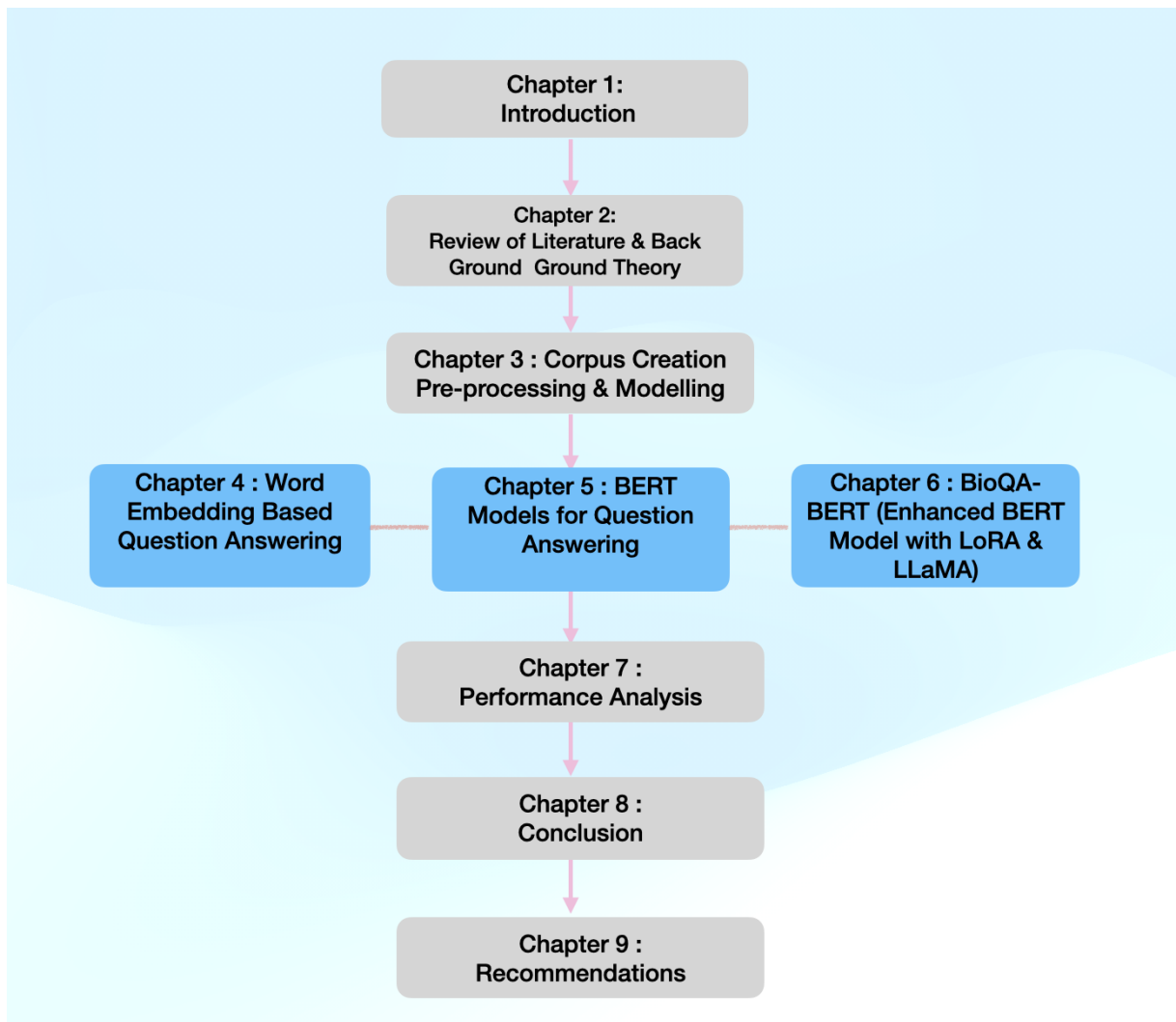


Figure 1.2: Thesis Outline

- Chapter 1 is the introduction, and the rest of the thesis is organized as follows.
- Chapter 2 provides a detailed review of research in the field of Information Retrieval (IR) and Question Answering (QA), along with a discussion of background theories such as Machine Learning (ML), Deep Learning (DL), and transformer models for building Question Answering Systems (QAS).

- Chapter 3 describes the development and annotation of the Malayalam Question-Answering data set MQuAD, followed by various pre-processing techniques and document modeling for Question Answering.
- Chapter 4 details the development of a Word Embedding and similarity measures-based Question Answering model.
- Chapter 5 encompasses the development of a hybrid Question Answering model by incorporating a Reader-Retriever pipeline architecture with BERT models.
- Chapter 6 describes developing an enhanced BERT model - BioQA-BERT for Question Answering, by integrating LoRA and LLaMA with BERT models.
- Chapter 7 discusses the performance analysis and discussion of various proposed Question Answering models.
- Chapter 8 summarizes the present work, emphasizes its importance, and outlines major contributions.
- Finally, Chapter 9 provides recommendations and a road map for future research.

Chapter 2

Review Of Literature and Background Theory

2.1 Introduction

In this chapter, the focus is on reviewing the existing research works related to the implementation of a Question Answering System (QAS) and providing a background study for the work at hand. This review serves as a critical foundation for understanding the current state of the field and the context in which the present research is situated. The principal aim of this phase is to perform an exhaustive examination of the literature about the deployment of Question Answering Systems (QAS).

This comprehensive examination establishes a robust background study, laying the groundwork for the research at hand. By investigating the wealth of prior investigations,

we aim to gain a deep understanding of the current state of the field, identify key trends, and ascertain the contextual landscape within which our present research is situated.

The review encompasses a wide spectrum of QAS, spanning various domains and methodologies, providing valuable insights into the evolution of this field over time. By critically assessing the strengths, limitations, and innovations of previous studies, this aims to impose the collective knowledge to inform and enhance the design, methodology, and contributions of the research. This chapter, therefore, plays a pivotal role as a foundational element, ensuring that the work is well-informed and positioned to make meaningful contributions within the broader context of the Question Answering process.

2.2 Review of Literature

This section thoroughly examines the extensive research efforts that have shaped the domain of Question Answering Systems (QAS). Excavating into the historical trajectory of these systems, Try to clarify significant milestones and breakthroughs that have shaped the current landscape, drawing insights from Mishra’s influential survey on the subject.[87]. Encompassing a broad spectrum of approaches, methodologies, and technologies utilized in prior works, explore diverse paradigms, including Rule-based Information Retrieval, Machine Learning, and Deep Learning methods. This examination critically assesses the strengths and weaknesses inherent in each approach, shedding light on their respective contributions to the evolution of Question Answering Systems.

Moreover, this chapter explores deep various application domains where Question Answering Systems (QAS) have found practical utility. Ranging from healthcare and educa-

tion to customer support and Information Retrieval, the review provides an over-arching view of the diverse use cases and contexts in which these systems have been deployed. By examining the applications, the systems gain valuable insights into the adaptability and versatility of Question Answering Systems, showcasing their impact across different fields [73].

In addition to investigating the application domains, the review extends to the consideration of different data sets and evaluation metrics employed in prior research. This analysis is crucial for comprehending the benchmarks and standards against which the performance of Question Answering systems has been measured. Understanding the methodologies for evaluating these systems provides a foundation for assessing their effectiveness and contributes to the establishment of best practices within the field. Altogether, this chapter aims to distill a rich understanding of the historical, methodological, and contextual dimensions of Question Answering Systems, setting the stage for the innovative contributions and insights to be unveiled in the subsequent sections of this research.

2.2.1 Review on Information Retrieval and Question Answering

In the research conducted by Nadia Soudani, Ibrahim Bounhas, et.al [133], an innovative Arabic semantic IR system was elucidated, using a text-mining approach. The authors introduced a versatile semantic search methodology based on semantic spaces, offering a generic framework for effective Information Retrieval. Their work involved a comprehensive experimental study, focusing on NLP tools for the Arabic language. This comparative analysis shed light on the efficacy of various NLP tools in the context of semantic search, providing valuable insights into their strengths and limitations.

Furthermore, the study looks into the exploration of linguistic resources and their impact on the performance of semantic search. By investigating the role of linguistic choices, the authors highlighted the significance of fine-tuning semantic search engine results. The research emphasized that the thoughtful utilization of linguistic resources can significantly enhance the precision and relevance of the semantic search outcomes. This underscores the relationship between language features and the effectiveness of semantic search, reinforcing the idea that linguistic considerations play a pivotal role in refining and optimizing Information Retrieval systems. The findings of the study contribute to the broader discourse on semantic search methodologies, particularly in the context of Arabic language processing, providing valuable insights for researchers and practitioners in the field.

Incorporating a knowledge-based approach, the system seamlessly integrated a Question Answering module designed for Arabic Semantic Disambiguation with a dictionary. The WSD process was skillfully executed through the application of a Sense Recognition algorithm. To explore diverse avenues for Semantic Information Retrieval, the system conducted experiments employing semantic spaces. These experiments encompassed rigorous tests utilizing various morphological analyzers and linguistic resources, contributing to a comprehensive evaluation of semantic disambiguation methodologies within the Arabic language context. The Mean Average Precision (MAP) for the system falls in ranges between 0.97% and 7.52%.

In a separate study, Shengxian, Yanyan, et. al [11]. introduced an innovative deep neural network architecture designed for semantic matching, as documented in their work. The proposed architecture, named MV-LSTM, is characterized by multiple positional sentence representations, offering a novel approach to capturing semantic relationships. Central to this architecture is the utilization of a Bidirectional Long Short-Term Mem-

ory (Bi-LSTM) network, strategically employed to model the intricate interactions between sentence representations. This modeling process incorporates three operations: bi-linear, and tensor layer, providing a comprehensive understanding of the semantic interplay within the given context.

To further refine the captured interactions, the k-max pooling strategy is introduced, enabling the selection of the top k-strongest interactions. The final output is generated through a Multi Layer Perceptron (MLP), contributing to the network's ability to discern intricate semantic patterns. The model undergoes training using Back-propagation and Stochastic Gradient Descent (SGD), optimizing its performance for the intended semantic matching tasks.

They tested out the MV-LSTM architecture in tasks like Question Answering (QA) and sentence completion, focusing on how well it handles understanding the meaning of words and phrases. The results were impressive, showing a big 11.4% improvement compared to the usual method. This highlights how effective the MV-LSTM architecture is at grasping and using the meaning behind words to boost performance in these language tasks. So, not only does this study bring in a cool new way of designing neural networks, but it also shows how well it works in real-world situations, especially when it comes to understanding the meaning of sentences

In their research, Saravanakumar Kandasamy et. al. [55] presented a methodology aimed at enhancing open-domain Question-answering, as documented in their work. The proposed method comprises two pivotal components: query processing and document processing. Within the aspect of query processing, the authors used different methods such as POS tagging, Parsing, Keyword extraction, Syn set identification NER and Similarity

Measurement. These techniques collectively contribute to the generation of alternative queries, fostering a more nuanced and expansive understanding of user inquiries. On the document processing front, the method involves intricate processes like Uniform Resource Locator (URL) weight calculation and Latent Semantic Analysis (LSA), refining the retrieval of answers. The precision of the system is evaluated as 0.77 and the Mean Reciprocal Rank (MRR) at 0.79.

Similarly, Arora, Foster et. al. [7] focused on improving Information Retrieval within the context of the WebAP data set by employing Query Expansion (QE) methods. Their diverse approaches encompass Pseudo Relevance Feedback (PRF), and Word Embedding of query words, strategically addressing the challenge of differences in query-document terms. The Word Embedding is accomplished through a feed-forward neural network, utilizing contextual information to predict a word. This innovative approach enhances the representation of query terms, mitigating the semantic gap and ultimately improving the effectiveness of information retrieval systems. The system evaluation was performed in terms of Normalized Discounted Cumulative Gain (NDCG) and the MRR. They are evaluated as 0.16 and 0.36 respectively.

In the research conducted by Nouha, Rim et. al. [95], the focus was on a Computational Question Answering (CQA) system. To overcome the lexical gap between questions, they employed a method based on the Word Embedding algorithm Word2Vec for the semantic modeling of words in a continuous vector space. The questions underwent ranking based on cosine similarity, where a previously posed question with a high similarity score to the newly queried question would be retrieved, along with its corresponding answer. The performance was evaluated using MAP, resulting in scores ranging from 0.39 to 0.45 across different models.

In another paper, Othman, et. al.[96] explain the intricate task of similar question retrieval within the realm of Community QA. The primary challenge addressed is locating historical questions that exhibit semantic equivalence to new queries, a task further complicated by the expanding archives of community forums and similar questions. Existing methods, often reliant on supervised models with extensive training data sets and manual feature engineering, encounter limitations due to their specificity, overlooking word order and insufficiently capturing syntactic and semantic information inherent in questions. This research aims to overcome these challenges by exploring innovative approaches to learning semantic representations that can enhance the effectiveness of similar question retrieval in the dynamic landscape of Community-driven Question Answering Systems.

The researchers advocate for a deep learning methodology employing Neural Networks (NN), specifically using a Siamese architecture with LSTM networks. This model incorporates an attention mechanism designed to discern word attention disparities during the modeling of questions. Moreover, Convolutional Neural Networks (CNN) are investigated within the Siamese architecture to facilitate pertinent question retrieval. The study explores multiple similarity measures to assess their efficacy in predicting semantic similarity between pairs of questions. The proposed approach undergoes rigorous evaluation through experiments conducted on extensive data sets encompassing both English and Arabic languages.

In their research, Shenghui Wang and Rob Koopman [151] conducted a comparative analysis of Word Embeddings, specifically examining Word2Vec and GloVe, against their novel Ariadne method. They used the document embedding method, Doc2Vec to perform Information Retrieval tasks on the Medline dataset. The reported average recall rates for the Doc2Vec and Ariadne methods were 93.3% and 86.3%, respectively. Significantly,

the findings demonstrated that Ariadne performs comparably well to Doc2Vec in specific Information Retrieval scenarios. Particularly noteworthy is Ariadne’s proficiency in providing contextual information about a word, making it a favorable choice in certain contexts.

In a different work, Prathyusha Kanakam, S. Mahaboob Hussain, and D. Suryanarayana [54] presented an algorithm designed for querying the semantic web. Their approach employs the SPARQL querying language and incorporates Linked Open Data Quality Assessment (LODQA) for information search, transforming query documents into numerical vectors. The Web Ontology Language (WOL) is utilized to describe relationships among different classes and corresponding relationships with SPARQL facilitating the retrieval of the most relevant answers. The study follows the High-Performance Linguistics algorithmic process, contributing to the effectiveness and precision of the proposed semantic web querying system.

In their study, Reshma PK and Lajish VL [110] introduced a semantic Information Retrieval model for the Calicut University domain, employing ontology through the utilization of Protege. This model uses ontology to facilitate the comparison of conceptual information across two knowledge bases on the web. Ontology, in this context, formally articulates a set of terms representing crucial concepts, including classes of objects and their interrelationships, thereby encapsulating a specific domain of knowledge. The WOL is employed as the tool for constructing these ontologies, involving distinct stages such as ontology capture, and ontology coding. The procedural method encompasses defining classes and their hierarchical structures, specifying object properties, defining instances of ontologies, and executing queries through DLQuery. The system’s performance is evaluated through recall and precision parameters reported as 56% and 87% respectively.

Pratibha, Parul et.al. [9] explored the development of a Cross-Language Information Retrieval (CLIR) system specifically tailored for English to Hindi language translation. In their research, they conducted experiments using documents retrieved from Google and Bing search results. A key focus of their work was the implementation of a two-level word sense disambiguation model designed to disambiguate Hindi words within the context of the English language. To enhance the efficiency of both translation and disambiguation processes, the researchers introduced a valuable new component, namely the analyzer, into the fundamental CLIR architecture.

This innovative approach aimed to optimize the system's performance in handling the intricacies of language translation and word sense disambiguation. The MAP values obtained for Bing and Google queries were reported as 0.35 and 0.45, respectively. These results underscored the efficacy of their two-level word sense disambiguation model and the overall system architecture in facilitating accurate and meaningful cross-language information retrieval between English and Hindi.

D Thenmozhi and Chandrabose Aravindan [142] spearheaded the creation of a CLIR system catering specifically to Tamil-English translation within the agriculture domain. Their innovative approach integrated the use of Ontology and Word Sense Disambiguation to enhance the precision and relevance of retrieved information. This system demonstrates a Mean Average Precision (MAP) of 95.36%, highlighting its effectiveness in facilitating accurate cross-language information retrieval.

In a parallel vein, Singh and Mishra [127] developed a semantic Information Retrieval system geared towards legal cases, as outlined in their work. This system leverages the Ontology merger and extends the GAIA methodology to incorporate comprehensive in-

formation about legal cases. Notably, the model goes beyond simple retrieval, providing reasoning capabilities to enhance the depth and context of the information retrieved. This advanced semantic Information Retrieval system offers a robust framework for navigating the complexities of legal case data, showcasing the potential for innovative methodologies to enrich information retrieval in domain-specific contexts.

In their work, Piyush Mital and Saurabh Agrawal [88] introduced a question-answering system that relies on graphs, with a specific application to Wikipedia documents. Their methodology encompassed the development of an information retrieval and information extraction system by converting unstructured document sets into structured graphs and also integrating natural language question processing. To improve the system's understanding of input queries and generate graph elements, the researchers employed a range of NLP techniques, including semantic role extraction, phrase chunking, and concept extraction.

The system's performance was evaluated, yielding notable results. The precision, recall, and average accuracy metrics were reported as 85.45%, 86.28%, and 80.1%, respectively. These outcomes underscore the efficacy of the graph-based Question Answering System in effectively transforming unstructured textual information into structured graphs, enabling accurate and insightful responses to natural language queries, particularly within the context of Wikipedia documents.

Dwaipayan Roy and Debasis Ganguly [118] introduced a query expansion technique based on Word Embeddings for the Information Retrieval system. They performed their experiment with Wikipedia documents. They utilized two distinct models, namely i) Word2Vec and ii) Fast-Text. For learning they used subword information. The calcula-

tion of word similarity employed Jaccard similarity metrics. To match query terms with embedded documents, they applied the Composition method with Indexing Units. The MAP for the Word2Vec and Fast-Text models was evaluated as 0.23 and 0.24, respectively.

Additionally, their findings indicated that Word2Vec demonstrated effectiveness on stemmed collections, while Fast-Text excelled on unstemmed collections. This research not only introduced a novel approach to query expansion but also provided insights into the comparative performance of different Word Embedding models in the specific context of Information Retrieval from Wikipedia documents.

Shomi Khan, Khadiza Tul et. al. [126] undertook efforts to enhance the answer extraction process within the Bengali question-answering system. In their study, they showcased the utilization of a WordNet and web document hierarchy for answer matching, by using Anaphora Catephora Resolution (ACR). WordNet, in this context, is employed as a lexical database that enriches the understanding of word relationships and meanings.

The proposed methodology incorporates a hierarchical structure of web documents, leveraging semantic matching techniques to improve the accuracy of answer extraction. Notably, Anaphora Catephora Resolution (ACR) is employed to address co-referential relationships within the text, enhancing the system's ability to identify and extract relevant answers. The average accuracy of the system was observed to be 74%, underscoring the effectiveness of their approach in elevating the performance of the Bengali question-answering system. This research contributes valuable insights into advancing answer extraction methodologies in multilingual contexts.

A novel query expansion framework based on learning-to-rank techniques was introduced by Bo Xu, Hongfei Lin, and Yuan Lin [157] for biomedical information retrieval. They chose candidate expansion terms by incorporating the MeSH thesaurus into a co-occurrence-based term selection process inside this framework. To refine these expansion words, feature vectors representing the terms were created by defining and extracting resource- and corpus-based term features. The development of term-ranking models was therefore made easier by using these vectors as inputs for learning-to-rank techniques. Several learning-to-rank techniques were investigated in the study to train these models. The Average MAP was used to evaluate the system's performance, and the result was 0.35.

They suggested learning-to-rank techniques to optimize the pseudo-relevance feedback method, a traditional query expansion strategy. The goal of this optimization was to improve the set of expansion words' accuracy. The combined efforts of these projects progress the domain of biological information retrieval by presenting new techniques for query expansion that take advantage of learning-to-rank techniques to achieve better results.

Manasamithra P. and H. C. Vijayalakshmi [80] presented a novel technique that uses a hybrid approach to convert natural language questions into queries that can be understood by the system. This hybrid approach uses the B-tree, an effective data structure, to store keywords that act as a knowledge base, smoothly integrating both keyword-based and semantic-based techniques. A dependency parser is used in the proposed system's semantic analysis to enable an extensive understanding of the user's natural language input.

In the experimentation phase, the system was applied to an employee database, demonstrating its practicality and effectiveness. The analysis revealed a significant reduction in execution time, with an impressive 86% decrease when employing the B-tree data structure. This outcome underscores the efficiency and optimization brought about by the hybrid approach, showcasing its potential to enhance the processing speed and overall performance of natural language query conversion systems. The proposed methodology not only advances the field of natural language processing but also introduces a practical solution for transforming user queries into a format comprehensible to computer systems.

Weiguo, Jeffrey et. al. [168], presented a sophisticated semantic Question Answering System for knowledge graphs. Their solution leverages the strength of both the knowledge graph and a large text corpus, utilizing an innovative and methodical methodology to understand natural language inquiries through a wide range of templates. Semantic graphs are then used to execute these templates, and a Stochastic Gradient Descent (SDG) is used to help identify target templates. Effective entity-level and structural-level disambiguation techniques were used by the researchers; these techniques are essential for transforming natural language inquiries into structured queries.

In the final stage, a SPARQL query is constructed based on the processed natural language question, leading to the retrieval of the corresponding answer. Two well-known knowledge graphs were used in the study: Freebase and the Wikipedia text corpus-Dbpedia. Impressive average precision values were achieved by the system; Dbpedia and Freebase achieved 84.67% and 82.19%, respectively. This research represents a significant stride in the development of semantic question-answering systems, showcasing its potential to effectively navigate and extract valuable insights from knowledge graphs.

A Semantic Sequential Dependence Model(SSDM) combining semantic data and the conventional Sequential Dependency Model(SDM) was developed by Fan Fang, Bo-wen Zhang et.al [32] for use in biomedical article search. Automatic synonym acquisition was achieved by using Word Embeddings with the word2vec and skip-gram models. For semantic integration, the neural network-based SSDM language model was employed. Additionally, the researchers used the K-nearest neighbors (KNN) classification algorithm to create a thesaurus.

The next step involved extracting query keywords from the generated thesaurus and replacing them with synonyms. After that, the synonyms were used to produce possible sequences that shared the same semantics as the original question. To get the findings that were received, these sequences were entered into the SDM. This innovative approach, combining semantic information and traditional modeling techniques, showcased its effectiveness in enhancing biomedical article searches. The incorporation of neural network-based models and semantic understanding enriches the depth of query processing, contributing to improved retrieval outcomes in the context of biomedical literature.

A unique deep learning architecture called DeepRank was presented by Liang Pang, Yanyan Lan, et. al. [98], to be used for relevance ranking in information retrieval. Their method, which consists of three main parts, mimics the relevance rating process used by humans. First, a query-centric context is used to find relevant sites. Next, local relevance that is, the significance between each query-centric context and the query itself is ascertained by utilizing two-dimensional Gated Recurrent Units (2D-GRU) in conjunction with a Convolutional Neural Network (CNN). Ultimately, a global relevance score is generated by combining the local relevances using an RNN.

The DeepRank model undergoes training using the Stochastic Gradient Descent (SGD) method. The effectiveness of the model was evaluated through experiments conducted with LETOR4.0 and large-scale Chinese click-through data, resulting in Mean Average Precision (MAP) scores of 0.49 and 0.41, respectively. This research demonstrates the potential of the DeepRank architecture in enhancing relevance ranking for Information Retrieval tasks, offering a valuable contribution to the field of deep learning-based retrieval models.

Ming Zhu, Aman Ahuja, et. al. [170] investigated the development of a neural network model intended for document rating in the context of healthcare Question Answering. Deep attention techniques at the word, phrase, and document levels are used in the suggested model to improve its capacity to identify complex correlations in the data. To further aid in model training and evaluation, the researchers added a substantial data set of healthcare query answers.

The neural network model, named Hierarchical Attention Retrieval (HAR), encompasses various components to optimize document retrieval for health-related queries. Word embedding, which generates k-dimensional word vectors, and the encoder, which uses a bi-directional Recurrent Neural Network(Bi-RNN) to encode temporal correlations inside the query and document words, are some of these components. 3) A bi-directional attention system that determines how relevant each query word is to each document word; 4) A query attention mechanism that uses self-attention to encode queries of varying lengths into fixed-size embeddings; 5) Document hierarchical inner attention mechanism, generating a fixed-dimensional representation of documents using sentence-level embeddings.

The system optimization involves a negative sampling mechanism. The evaluation of the model was performed on a healthcare dataset, termed HealthQA, with metrics such as MRR reaching 87.87%, and recall reaching 96.84%. This research contributes a robust neural network model tailored for healthcare Question Answering, showcasing its effectiveness in capturing hierarchical dependencies and providing valuable insights within the healthcare domain.

Zhuyin and Jamie [25] introduced a transformer model- BERT, to enhance text comprehension in the realm of Information Retrieval. BERT is specifically applied for ad-hoc document retrieval, where the input comprises concatenated query and document tokens, separated by a special token ['SEP'] to delineate the two segments. The token embeddings are obtained by embedding the concatenated tokens, and the position embedding is incorporated to preserve word order.

The tokens traverse multiple layers of transformations, generating new contextualized embeddings for each token at each layer through weighted sums of all other token embeddings. Attention matrices play a crucial role in determining the weights, with stronger attention indicating a closer relationship between words. For each query-document pair, the output embedding of the first token is used, and it is fed into an MLP to forecast the relevance likelihood. To improve performance even more, search knowledge can be added to this model. Two standard datasets, Clueweb 09-B, and the Robust-04 news corpus, were used for the evaluation. The NDCG was reported as 0.52 and 0.29 for Robust-04 and Clueweb 09-B, respectively, in terms of accuracy. According to the findings, BERT performs better on the Robust-04 data set than it does on the Clueweb 09-B data set.

A Graph Embedding-based Product Search (REPS) ranking model was developed by Yuan, Dong et.al. [77] specifically for e-commerce search applications. Click-graph features are effortlessly integrated by the system into a single neural ranking framework. The model starts with a simple neural network architecture as the basic framework and then integrates a graph embedding technique to improve its retrieval performance.

To extract semantic features from the query and product description phrases, the procedure entails vectorizing these terms, feeding them into CNN layers, and applying max-pooling layers for dimension reduction. Then, semantic feature vectors are transformed into the query's vector space using an MLP, which produces a relevance score. During the training phase, either CNN or RNN is used to introduce graph embedding.

The evaluation of the model was conducted using the CIKM Cup-2016 Track-2 data set, with MAP, MRR, and NDCG metrics yielding values of 0.46, 0.49, and 0.53, respectively. This research showcases the effectiveness of the REPS model in enhancing product search ranking within e-commerce scenarios.

The TRanslate-Edit Model for Question-to-SQL(TREQS) is a revolutionary approach that Ping Wang, Tian Shi et.al.[150] developed specifically for Question Answering on Electronic Medical Records. To produce SQL queries in response to provided questions, the TREQS model makes use of a sequence-to-sequence architecture, a commonly used deep learning paradigm. Interestingly, the model includes task-specific look-up tables and an attentive copying mechanism to carry out the required adjustments.

The researchers assembled a large dataset, MIMICSQL, using publicly accessible Electronic Medical Records(EMR) to aid with their research. There are two sets in this dataset: template questions make up the first set, while natural language questions make up the second. To demonstrate the effectiveness of the suggested methodology, a comprehensive series of tests spanning both template and natural language queries were conducted on the MIMICSQL dataset.

The TREQS model adopts an RNN Sequence-to-Sequence (Seq2Seq) framework for Question-to-SQL generation. The encoder processes a sequence of Word Embeddings for input tokens, transforming them into a sequence of encoder-hidden states. During testing, a beam search algorithm is employed for SQL generation. The model exhibits significant performance improvements on both the development and testing data sets, surpassing other models by an average of 30 percent in accuracy. The average accuracy of the model was evaluated as 0.97.

Using graph embedding approaches, Yuan Zhang, Dong Wang, and Yan Zang [41] present a novel way to improve neural information retrieval models. They concentrate on using a neural network and Graph Embedding ranking model to solve document ranking and information retrieval problems, especially in the Product Search space.

The model integrates graph embedding techniques, which involve representing terms of queries and product descriptions as vectors. These vectors undergo semantic feature extraction through CNN layers, with subsequent dimension reduction achieved by max-pooling layers. Semantic feature vectors are then transformed into a shared vector space with the query using a Multi-Layer Perceptron (MLP), which finally yields a relevance score.

The integration of graph embedding during the training phase, either using CNN or RNN contributes to the model's ability to effectively retrieve relevant information in the context of product search. This approach is poised to enhance the performance and efficiency of information retrieval systems, particularly in scenarios related to product search and recommendation.

In their paper, Tom Young, Devamanyu et.al. [162] reviewed different deep learning models that are used in different NLP tasks and go through their evolution. They discussed the neural networks based on dense vector representations and their success, such as Word Embedding and deep learning methods. They also went over the main deep learning models and techniques, like recurrent and convolutional neural networks, that can be used in language processing.

In their recent research, Fan Fang, Bo-wen Zhang et.al. [32] introduced an innovative approach by integrating semantic information with the conventional Sequential Dependence Model (SDM). Their methodology involves leveraging Word Embedding-based synonym mapping to enhance the process of information retrieval. The research underscores the significance of semantic understanding in the context of document search and retrieval. By incorporating Word Embedding-based synonym mapping, the model gains the ability to capture nuanced relationships between words and their synonyms, thereby improving the accuracy and relevance of retrieved information.

This fusion of semantic information with the established SDM provides a more comprehensive and contextually aware framework for information retrieval. The integration of Word Embeddings for synonym mapping is a pivotal element, allowing the model to discern semantic similarities and enrich the retrieval process. Overall, this approach rep-

resents a noteworthy advancement in the field of information retrieval, offering a nuanced and effective means of enhancing search capabilities.

In their research endeavor, Bo Xu, Hongfei Lin, and Yuan Lin [157] presented an innovative biomedical information retrieval system that harnesses the power of query expansion through learning-to-rank methods. Their approach integrates various techniques, including the co-occurrence term selection method, pseudo-relevance feedback method, and query expansion method, to optimize the information retrieval process in the biomedical domain.

The focal point of their work lies in leveraging learning-to-rank methods to enhance the performance of the information retrieval system. The co-occurrence term selection method plays a crucial role in selecting candidate expansion terms, while the pseudo-relevance feedback method is employed to refine the set of expansion terms. This intricate combination of methodologies aims to improve the system's ability to retrieve relevant and accurate biomedical information.

By incorporating these advanced techniques, the proposed system endeavors to provide more nuanced and contextually relevant results in the challenging domain of biomedical information retrieval. This research contributes to the ongoing efforts to optimize information retrieval methodologies, particularly in specialized domains such as bio-medicine.

Dwaipayan, Debasis, et. al. [118] investigated how word normalization and collection decisions affected Word Embedding learning and, in turn, ad hoc retrieval performance. Two measures were presented in the study that are intended to measure how similar word

vector embedded spaces are to one another. Notably, the results showed that fast-Text performs well on an unprocessed collection and the word2vec method performs well on stemmed collections on average. These findings were explained by the inherent qualities of the corresponding embedding techniques.

Shomi Khan, Khadiza Tul et.al. [59] introduced a Bangla Question-Answering System, exploring its application in both Bangla and English languages. The system utilized anaphora-cataphora resolution to enhance semantic matching. Additionally, Archana S.M, Naima Vahab, et. al. [6] devised a rule-based Malayalam Question Answering System. This system identified question words and associated bhakti (devotion or focus) with the answer, retrieving answers based on the question module and answer module.

In their research, Vaishali Singh et. al [129] put forth a personalized approach to Question Answering that involves end-user modeling. This personalized approach tailors the retrieval of information based on user-specific information and areas of interest. The personalization of retrieved data is achieved through the application of various similarity measures, including attribute values similarity and entity values similarity, which consider the specific preferences and characteristics of individual users.

In their study, Sheetal.S, et. al [131] introduced an innovative method for assessing document similarity employing a graph model. Their approach involves a modified technique applied to WordNet and Wikipedia. The method utilizes a weighted conceptual graph, representing text documents through the coexistence of terms. To construct the graph, they employed co-reference resolution to determine associations among feature terms and assign weights to the graph edges based on term importance.

Swathilakshmi Venkatachalam, et. al. [148] introduced a system for Tamil news content retrieval that utilizes ontology-based information extraction and summarization. Through Information Extraction, the system retrieves specific information from natural language and integrates it into an ontology. The ontology is organized into two distinct domains. Subsequently, a multi-document text summarizer generates a concise overview of significant events. Finally, the query extractor retrieves data from the database and presents it to users based on their queries.

Navjot Kaur, et. al. [58] designed a semantic information retrieval system focused on the music domain. They employed string ontology for semantic information retrieval, incorporating query reformulation techniques to facilitate multilingual information retrieval. The system’s approach involves leveraging semantic structures to enhance the accuracy and relevance of retrieved information in the music domain.

A Natural Language Processing Question-Answering System, NLP-QAS built to deal with BERT models was presented for Question Answering in a work by Phuc et. al. [28]. The phrase selection and answer detection phases make up the two stages of the system. The SQuAD dataset was analyzed by the study utilizing the suggested NLP-QAS. Lemmatization, punkt, and punkt with SDP were among the variants used for sentence selection. Notably, during the sentence selection stage, Punkt with SDP proved to be the most effective technique. Several Relation Network Parser (RNP) techniques, including RC, SPOS, and SNER, were investigated during the answer discovery step.

Choi, Byeongmin, et. al. [23] propose to use a knowledge graph information extraction method in conjunction with the most recent pre-trained language model, namely the lite bidirectional encoder representations from transformers, ALBERT. We also present a new

method, schema graph expansion, that we apply to modern language models. We assess the influence of incorporating knowledge graph-based knowledge extraction methods into state-of-the-art pre-trained language models. Furthermore, provides proof that on the Commonsense QA data set, our suggested model performs better than the current KagNet and MHGRN models.

The authors, Yue Feng, Zhen Han et. al. [33] from Baidu Research, introduce the Document Entity Heterogeneous Graph (DEHG) Network as a novel approach to integrate different sources of information and conduct reasoning on heterogeneous information for open-domain Question Answering.

Experimental results on the HybirdQA data set demonstrate that DEHG outperforms state-of-the-art methods, showcasing its effectiveness in handling multi-hop reasoning on integrated heterogeneous information. The research contributes to the field by homogenizing structured and unstructured knowledge in open-domain Question Answering and analyzing complex relations of heterogeneous information. The proposed DEHG model provides promising insights and advancements in the domain of open-domain Question Answering.

The paper Bulian, Cbuck et. al. [17]presents a systematic analysis of the limitations of token-level equivalence measures in evaluating Question Answering Systems. The authors introduce the concept of Answer Equivalence(AE) and provide data-driven analysis to reveal the limitations of existing measures. They propose a BERT Matching Measure (BEM) measure as a better approximation for AE. The paper concludes by highlighting the limitations of their work and suggesting future research directions.

CoQUAD is a COVID-19 Question Answering data set system that is intended to help researchers, clinicians, and the general public acquire genuine scientific information in the form of quality assurance (QA). Raza, Shaina et.al. [108] discusses CoQUAD. The system is made up of a Retriever that answers questions by retrieving documents from the document store, and a Reader that uses MPNet to extract the precise response to each query from the documents that the Retriever returns. Utilizing a gold-standard data set for evaluation, the system outperformed state-of-the-art models. Along with the methods for preparing the data sets, the publication also includes information on the inclusion and exclusion criteria for the literature used in the system.

Kim, Yohan et. al. [61] investigate address the labor-intensive process of manually looking through large amounts of textual data to find information about infrastructure damage. The research recommends a Question Answering technique that automatically pulls damaged information from this kind of data, making the process of efficient infrastructure planning more streamlined. For information retrieval, the suggested method makes use of a natural language model, namely the bidirectional encoder representations from transformers BERT. The model was trained and evaluated on a subset of more than 100 reports obtained from the National Hurricane Center, resulting in 533 question-answer pairings. For the hurricane and earthquake data sets, additional testing using different question-answer pairs with information related to earthquakes was produced. F1 scores of the system were evaluated as 90.5% and 83.6%, respectively.

The literature review by Roy, Pradeep [119] on Computational Question Answering(CQA) synthesizes findings from 133 articles, focusing on three modules: questions, answers, and users. Key research themes involve question-and-answer quality, as well as expert identification. Notable platforms studied include Yahoo! Answers, Stack Exchange, and Stack

Overflow, yet cross-platform investigations are limited. ML dominates over DL, but DL adoption in CQA research is increasing. The study highlights challenges like question un-answerability, answer quality prediction, and finding expert users. The user module addresses finding experts and reputation collectors. Limited use of ensemble and multi-modal techniques is noted. Encouraging cross-platform inquiries and comparing CQAs in different languages is suggested for future research.

In the paper Pal, Anki [97] introduces MedMCQA, a novel large-scale dataset for Multiple Choice Question Answering (MCQA) specifically tailored to real-world medical entrance exams. The data set comprises over 194,000 high-quality MCQs from AIIMS and NEET PG entrance exams, covering 2,400 healthcare topics and 21 medical subjects. The questions exhibit an average token length of 12.77 and significant topical diversity. Each sample includes a question, correct answer(s), and other options, demanding a comprehensive language understanding and testing reasoning abilities across diverse medical subjects and topics. The study offers a detailed solution explanation alongside the data set specifics.

Recent advancements [76] Liu, Andy, et. al. have demonstrated the effectiveness of prompt-based learning in few-shot NER with pre-trained language models; enhanced label efficiency is achieved by employing prompts as task direction. Nevertheless, earlier prompt-based techniques for few-shot NER have drawbacks including increased computing complexity, a restricted capacity for zero-shots, a reliance on human prompt engineering, and a deficiency in prompt robustness. They suggest a brand-new prompt-based learning strategy for Question Answering using NER - QaNER, to overcome these drawbacks. NER problems are converted into the QA formulation using a revised strategy. NER prompts are generated specifically for QA models. Prompt tuning using QA models is carried out

by utilizing a few annotated NER examples and zero-shot NER is achieved by prompting the QA model. QaNER exhibits faster inference, tolerance to hyper-parameters, insensitivity to prompt quality, and considerably superior performance in low-resource settings and zero-shot capabilities when compared to previous approaches.

2.2.2 Review on Malayalam Information Retrieval and Question Answering

This section provides a brief insight into the research works that have already taken place in the field of Malayalam Question Answering. There is limited research happening on Question Answering in Malayalam, especially at the semantic level. Additionally, there is no notable research work on semantic Malayalam Question Answering.

Arjun Babu and Sindhu L [8] have developed an Information Retrieval System for Malayalam Using Query Expansion Techniques. They used cosine similarity and vector space modeling for answer retrieval. They also used synonym mapping to improve the accuracy of the system.

Bindu, M S and Dr. Sumam Mary Idicula [13] present a novel approach to question-answering systems tailored for Malayalam. Advanced deep learning methods like Gated Recurrent Unit (GRU), Memory Network models, and LSTM are all used in their work. The investigation utilizes the diverse tasks provided by the Facebook bAbI data set,

including questions with multiple supporting facts, inductive and deductive reasoning, and coreference resolution, among others, to train and evaluate the proposed system. The findings reveal that the Memory Network model stands out, achieving an impressive average accuracy of 80% in accurately retrieving answers in Malayalam.

Archana, Naima et. al. [6] addresses the development of a Question Answering System for Malayalam, an agglutinative language. Malayalam sentences exhibit morphophonemic modifications, specifically using vibhakthi, to denote relationships to verbs or other nouns. By comparing the vibhakthi and POS tags of question words with terms in the provided corpus, the QAS seeks to deliver word-level replies. It may also be extended to examine related Karaka roles for more effective question analysis.

Reji Rahmath K and P.C. RaghuRaj [106] presented a sophisticated system intended to provide accurate responses to user-posed natural language queries. Operating on a specified collection of raw material in natural language, this system is an advanced example of information retrieval (IR). The study suggests a Question Answering (QA) system based on deep learning that is specially designed for the Malayalam language. The system is capable of analyzing the Malayalam language with precision since it makes use of state-of-the-art methods including GRU, Memory Network models, and LSTM.

To evaluate and validate the performance of the proposed QA system, the researchers utilized the Facebook bAbI dataset, which encompasses 20 tasks featuring questions with multiple supporting facts, inductive and deductive reasoning, coreference resolution, and more. The system was trained and tested on this comprehensive data set. Notably, the experimental results revealed that the Memory Network model outperformed the other implemented models, achieving the highest average accuracy of 80%. This underscores

the efficacy of the proposed deep learning-based QAS in retrieving precise answers in Malayalam, demonstrating its potential for addressing natural language queries with a high degree of accuracy.

2.2.3 Review on Large Language Models

This section discusses various Question Answering and Information Retrieval works that happened by using Large Language Models. Some of the recent works are described below.

In a work, Dan Su, Xan Xu, et. al. [137] propose a multi-task learning framework that influences large pre-trained language models to enhance the generalization ability of Question Answering Systems. The approach fine-tunes a pre-trained language model on multiple Reading Comprehension data sets and uses a multi-task learning objective to jointly optimize the model for different QA tasks. Experimental results show that the proposed approach outperforms the BERT-Large baseline on broader QA tasks, achieving an average F1 score of 68.98 and an Exact Match score of 56.59.

A study on Large-Scale Relation Learning for Question Answering over Knowledge Bases using Pre-trained Language Models is presented by Yuanmeng, Rumei, et. al. [159]. They advocate the utilization of BERT-like pre-trained language models to enhance the scoring function for ranking candidate answers. The paper introduces three relation-learning tasks aimed at facilitating the alignment of natural language expressions with relations in the Knowledge Base(KB). Comparative analyses with various baselines reveal the superior performance of the proposed approach. However, the integration of

pre-trained language models introduces potential risks, including the introduction of extra data biases and sensitivity to adversarial examples. Hence, users must exercise caution and ensure the preparation of these resources aligns with task relevance, unbiased representation, and ethical considerations.

In a paper Singal, Karan et. al. [130] discuss the development of a large language model called Med-PaLM 2, which is designed to answer medical questions with expert-level accuracy. They Med-PaLM 2 to previous models and demonstrate its superior performance on a range of medical Question Answering tasks. Also discuss potential applications of this technology in the medical field, such as improving patient outcomes and reducing healthcare costs, while acknowledging the challenges that still need to be addressed, such as ensuring the model’s ethical and legal compliance. Overall, the paper highlights the exciting progress that has been made in using artificial intelligence to improve healthcare and suggests that further research in this area could have significant benefits for patients and healthcare providers alike.

To enhance Question Answering performance, Yasunaga, Hongyu, et. al. [161] present a novel model called QA-GNN that blends knowledge graphs and pre-trained language models. The proposed model addresses the challenges of identifying relevant knowledge and performing joint reasoning through innovative techniques. The authors evaluate the model on various benchmarks and demonstrate its superior performance compared to existing models. The paper also acknowledges the support of various organizations and funding agencies.

In a work, Lazaridou, Angeliki et. al. [69] unique few-shot capabilities inherent in Large Scale Language Models (LSLM) to address certain challenges related to ground-

ing in factual and up-to-date information. Inspired by semi-parametric Language Models (LM), which base their decisions on externally retrieved evidence, this work employs few-shot prompting to teach LMs to incorporate information obtained from the web through Google Searchâa comprehensive and continuously updated knowledge repository. Our method does not require fine-tuning or the introduction of additional parameters, making it applicable to any LM and providing a robust baseline. Remarkably, we discover that LMs conditioned on web-derived information outperform closed-book models of similar, or even larger, sizes in open-domain Question Answering. Additionally, observe that enhancing inference-time computing by utilizing multiple retrieved evidence to generate multiple answers, followed by a reranking stage using scores from the same LMs, yields improved performance and mitigates the lower efficacy of smaller few-shot LMs. In conclusion, our findings suggest that rather than solely pursuing larger models, it may be advantageous to redirect focus toward discovering more effective ways to utilize models, such as enhancing prompting or increasing inference-time computing.

In their paper, Shao, Zhenwei, et. al. [124], Presenting Prophet, a framework designed to provide knowledge-based visual Question Answering on the GPT-3 with answer heuristics. The primary goal is to fully harness the potential of GPT-3 when the provided input information is insufficient. The study reveals that confidence scores play a crucial role in the Prophet’s performance, and even without image captions, they maintain stability. Interestingly, the importance of the prompt head is found to be relatively less significant, suggesting GPT-3’s ability to comprehend the task directly from in-context examples. Additionally, the introduction of extra information, such as object tags, results in a slight performance drop. The authors also present a Prophet variant called Prophet-MC, tailored for multiple-choice tasks, which achieves notably higher accuracy, highlighting the broader applicability of Prophet to related tasks.

In a work by Xing hang Hu [44] a novel approach to answer selection in Community Question Answering (CQA) using pre-trained and large language models. The proposed Question-Answer cross attention Network (QAN) with pre-trained models achieves state-of-the-art performance on two data sets. The authors also introduce external knowledge to further improve the accuracy of answer selection. The paper provides details on the methodology, experiments, and results, highlighting the effectiveness of the proposed approach. Experiments on the SemEval2015 and SemEval2017 data sets demonstrate that the suggested strategy provides state-of-the-art performance and outperforms current approaches. The introduction of external knowledge further improves the accuracy of answer selection, demonstrating the effectiveness of the proposed approach.

In a paper, Hu, Shen [43] presents LoRA, a low-rank adaption method that preserves good model quality while cutting the number of trainable parameters for downstream tasks in big language models. The authors provide empirical evidence that rank deficiency is a common phenomenon in language model adaptation and that LoRA can effectively address this issue. They demonstrate the effectiveness of LoRA on several benchmark data sets and compare it to other adaptation techniques. The proposed method facilitates swift task-switching when implemented as a service by sharing a significant portion of the model parameters. This approach proves to be an efficient adaptation strategy that avoids introducing inference latency or diminishing the input sequence length.

WebGLM, an innovative web-enhanced Question Answering system that uses the General Language Model (GLM), was introduced by Xiao, Liu, et. al. [72]. The main goal is to improve a large language model (LLM) that has already been trained with effective web search and retrieval capabilities for real-world use. With an emphasis on the LLM-augmented retriever, bootstrapped generator, and human preference-aware scorer,

WebGLM is aimed to overcome the shortcomings of WebGPT (OpenAI) and guarantee increased precision, effectiveness, and economy. We also suggest methodical standards for assessing web-enhanced quality assurance solutions. Extensive multi-dimensional human assessments and quantitative ablation experiments demonstrate how much better WebGLM’s designs work than those of other systems. Notably, in human evaluation, WebGLM utilizing the 10-billion-parameter GLM (10B) performs better than similarly sized WebGPT (13B) and even achieves results comparable to WebGPT (175B).

2.3 Research Gap

Upon exploring the existing literature, it is evident that the majority of research works have happened in other global languages like English. While there are a limited number of research works dedicated to native languages such as Tamil and Kannada, it is noteworthy that the Malayalam language has yet to witness the emergence of prominent semantic Question Answering Systems.

One notable gap in existing Malayalam Question Answering is that all are based on Machine Learning and Deep Learning techniques. No Large Language Model-based Question Answering models exist, particularly those related to the health domain. As the exploration of natural language processing expands, the identification of such gaps not only highlights areas for future research but also emphasizes the pressing need for the development of comprehensive language models and Question Answering systems that cater to the linguistic diversity and domain-specific needs of languages like Malayalam.

2.4 Back Ground Theory

This section extensively examines the theories and models essential for constructing a Question Answering System, through the integration of foundational machine learning principles, advanced deep learning architectures, and specialized Question Answering models.

2.4.1 Question Answering System

A Question Answering System (QAS) represents a sophisticated application of Artificial Intelligence and Natural Language Processing, meticulously designed to address user queries effectively. These systems come in various manifestations, including chatbots, virtual assistants, and standalone applications[34]. The training process involves exposing them to extensive documents or websites, allowing them to discern intricate language patterns and accumulate knowledge. This accumulation of knowledge leads to the creation of a comprehensive language model.

Utilizing these language models, Question Answering Systems excel in generating responses that are relevant and highly accurate, catering to the diverse inquiries posed by users. The ability to understand context, decipher intent, and provide coherent answers makes these systems invaluable in enhancing user experiences across a spectrum of applications [27]. Whether integrated as chat-bots for customer support, virtual assistants for streamlined interactions, or standalone applications for specialized domains, Question Answering Systems epitomize the synergy of cutting-edge technology and linguistic prowess.

Different Models of Question Answering

- Fact-based Question Answering:- These systems answer factual types of questions [107] such as "What is the capital of India?" or "Who invented Television?"
- Contextual Question Answering:- These systems provide answers based on a given context such as a paragraph or document [154].
- Interactive Question Answering:- These systems engage in conversations with users to understand and answer queries conversationally [87][63]

Different Approaches to Question Answering

Building a Question Answering (QA) system encompasses a variety of approaches, each addressing specific challenges and requirements [134]. Here are several key methodologies commonly employed in the development of Question-Answering systems:

- Rule-Based System:- In a rule-based system, the questions are matched against predefined patterns or rules, and answers are generated based on these rules [47]. If a question contains a specific keyword or pattern, then the system can generate a predefined answer.
- Information Retrieval:- Information Retrieval-based systems involve searching for relevant information in a large corpus of documents such as databases or websites [1]. Different techniques used are keyword matching, semantic matching, and indexing, and then the answers can be generated from the retrieved results.
- Machine Learning:- Machine learning-based systems use algorithms and models to learn patterns from data and generate answers. Supervised learning and unsu-

ervised learning are two distinct categories of machine learning. Labeled data is utilized in supervised machine learning methods to train the model [3]. It can be used for models such as answer classification. Unsupervised learning such as topic modeling or clustering can be used for organizing and clustering relevant answers.

- **Deep learning:-** Deep learning techniques such as neural networks can be used for Question Answering [90]. RNN [158] or transformer-based models can be used to identify contextual information and generate answers for queries [91] from a large volume of data. Here use a pre-trained model like BERT [5] or GPT3 [14] for Question Answering tasks.
- **Hybrid approaches:-** Hybrid approaches combine multiple techniques such as rule-based, Information Retrieval, Machine Learning, and Deep Learning to leverage the advantage of each approach [145]. For example, a system may use a rule-based approach for simple questions and switch to IR or ML models for complex questions.
- **Knowledge Graph:-** In Knowledge, a graph represents structured knowledge as entities and relationships. It can be used for Question Answering tasks. Questions are formulated to traverse the graph and then retrieve relevant information, which can be used to generate answers [51].
- **Memory Augment Networks:-** Memory-augmented neural networks such as memory networks or neural Turing machines can be used to store and retrieve information for context-dependent answer retrieval [22].

These are a few of many techniques that can be used in building a Question Answering System. The techniques may vary the requirements of the task, available data, and desired performance.

Different Steps Involved in Question Answering

Natural Language Processing (NLP) encompasses the computational capability to comprehend, process, interpret, and generate human language, as articulated by Ranjan [107]. This intricate field of study is instrumental in bridging the communication gap between humans and computers, enabling machines to interact with and understand human language in a meaningful way.

In the context of a Question Answering System (QAS), the process involves several crucial steps that collectively contribute to the system's ability to comprehend user queries and provide relevant responses. These steps include initial data pre-processing, where raw text is transformed into a format suitable for analysis; feature extraction, which involves identifying and selecting pertinent linguistic features; and the application of advanced machine learning or deep learning models to train the system in understanding and responding to diverse queries. The holistic integration of these steps forms the backbone of an effective QA system, emphasizing the pivotal role of NLP in facilitating a seamless and intelligent interaction between users and computational systems.

- **Input Text:** The system takes the query in the form of text. Here the queries are natural questions from users in their native language.
- **Text Processing:** The input text is processed and then converted to a different format, ie vector representation to process and identify relevant information such as keywords, entities, and contextual information.
- **Information Retrieval:** The system searches through databases or corpus to find relevant answers related to the query. This can involve techniques such as keyword matching, semantic mapping machine learning, and deep learning algorithms.

- **Answer Generation:** Once the relevant information is retrieved, the system generates an appropriate answer. This can involve techniques such as summarization, paraphrasing, or extraction of information from the retrieved document.
- **Answer Prediction:** The generated answer for the query is presented in a human-understandable format such as text, speech, or visual display.

2.4.2 Architecture of a Question Answering System

A Question Answering System aims to automatically find accurate and precise answers to users' queries in natural languages. Unless a traditional IR, the Question Answering System will not return a whole document for a query. A Question Answering system will try to find only one or more specific and small answers to the user's queries.

Different types of Question Answering Systems exist now such as factoid Question Answering Systems, list types, descriptive types, etc. A factoid Question Answering System will return an answer about a single fact. The list Question Answering System will return more than one answer. A descriptive-type Question Answering System will return a complex answer, it may be from multiple document sources.

A Question Answering System takes user queries as input, evaluates and analyzes them, compares them to the documents that are on file, and then extracts and retrieves accurate responses to natural language inquiries. Information extraction and retrieval techniques are applied by a Question Answering System.

To find more accurate and proper answers to natural language queries. Natural language query processing and Question Answering must need the semantic processing and contextual information of natural language queries and documents [36].

The architecture of a Question Answering System typically involves several key components that work together to process queries and generate accurate responses. While different QA systems may adopt varying approaches, here is a generalized architecture as shown in Figure 2.1.

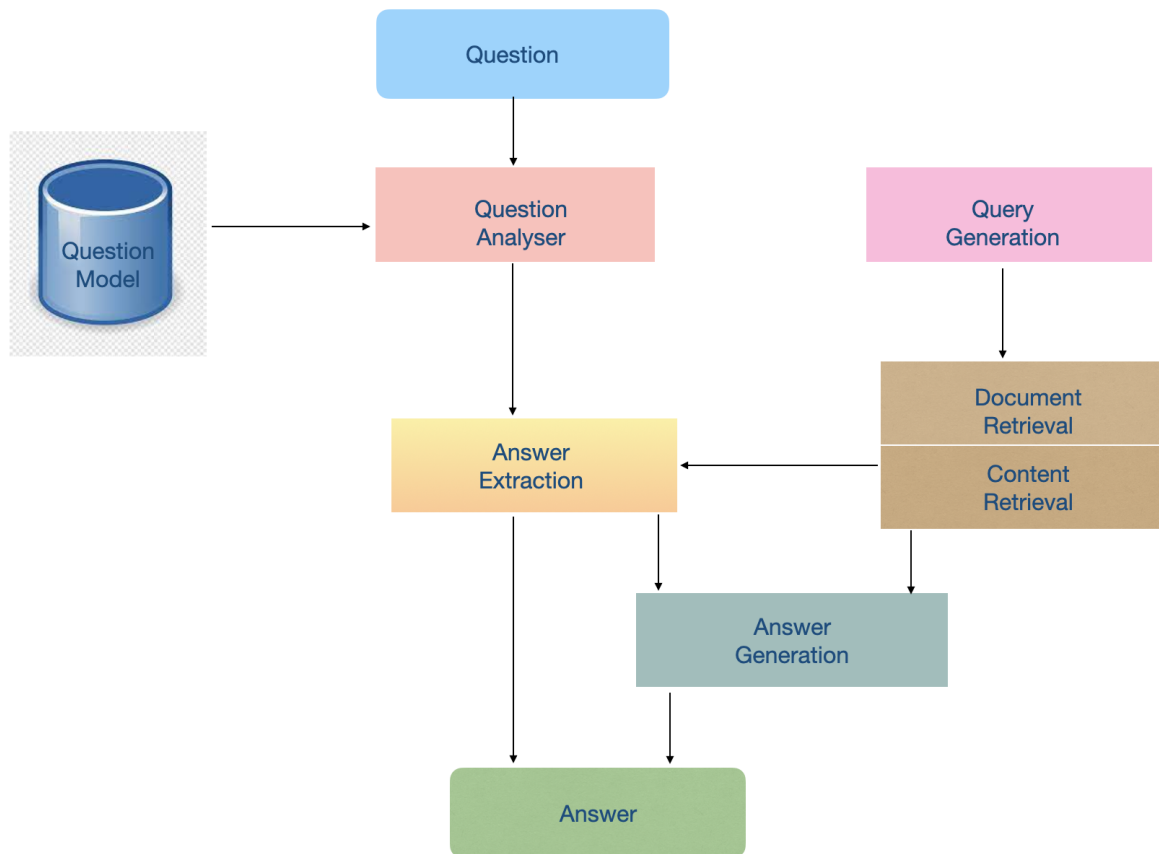


Figure 2.1: General Architecture of Question Answering System

The different components of a Question Answering System include

- Question Analyzer:

This module interprets and breaks down the user's question. It involves syntactic and semantic analysis to understand the intent and entities involved.

- Query Generation:

It will generate the actual query for the Question Answering process.

- Document Retrieval:

The system identifies relevant documents or passages from a knowledge base or corpus that might contain the answer. Document Preprocessing: The retrieved documents undergo preprocessing steps like tokenization, stemming, and other natural language processing techniques.

- Content Extraction:

The system extracts relevant information and context from the retrieved documents to understand the context in which the question is posed.

- Answer Generation:

The system extracts specific spans of text from the context that are considered potential answers. If there are multiple potential answers, a ranking mechanism may be employed to select the most suitable one.

- Answer Retrieval:

Checking the validity and coherence of the generated answer using additional criteria. Formatting the answer in a way that is understandable and user-friendly.

The implementation of a Question Answering System may vary depending on the techniques and technologies used and the domain or context of the questions. Also, the system can continuously enhance and improve its performance through feedback from users and iterations.

2.4.3 Neural Network Models for NLP

Neural networks have become a foundational technology in NLP, enabling breakthroughs in various NLP tasks [52]. Here are some popular Neural Network models used in NLP.

- **Recurrent Neural Networks (RNNs):** Since RNNs are built to handle sequential data, text and other NLP tasks involving sequences are a good fit for them[141]. Applications include machine translation, sentiment analysis, and language modeling. Limitations: The vanishing gradient problem makes it difficult to capture long-term interdependence.
- **Long Short-Term Memory (LSTM):** An RNN type called LSTMs was created to deal with the vanishing gradient issue, allowing for better preservation of long-term dependencies[160]. Applications: Similar to RNNs but with improved performance in capturing long-range dependencies.
- **Gated Recurrent Units (GRU):** Similar to LSTMs, GRUs are designed to address the vanishing gradient problem. They have a simpler architecture with fewer parameters[172]. Applications: Similar to LSTMs, but computationally more efficient.
- **Word Embeddings (Word2Vec, GloVe):** These models learn distributed representations of words in continuous vector spaces, capturing semantic relationships between words [37]. Applications: Preprocessing and modeling of documents for many NLP tasks, improving the representation of words in neural networks.
- **Convolutional Neural Networks (CNNs):** While often associated with computer vision, CNNs can be used for NLP tasks by treating text as an image. Filters are applied to capture local patterns [153]. Applications: Text classification, sentiment analysis, document categorization.

- **Transformers: Overview:** Introduced in the "Attention is All You Need" paper, Transformers have become a cornerstone in NLP. They use self-attention mechanisms to capture relationships between words in a sequence[116]. Applications: Cutting-edge outcomes for a variety of tasks, including machine translation (e.g., in models like BERT, and GPT).
- **Bidirectional Encoder Representations from Transformers(BERT):** BERT is a model based on transformers that obtain bidirectional context information through training on extensive unlabeled data, as highlighted in the work by Sabharwal et. al. [120]. Its applications span various NLP tasks, demonstrating state-of-the-art performance in Question Answering, text classification, and named entity recognition.
- **Generative Pretrained Transformer(GPT):** GPT is a Generative Pretrained Transformer, it is another transformer-based model that is pre-trained on a large corpus of text data and can be fine-tuned for specific tasks[39]. Applications: Text generation, language understanding, completion.
- **XLNet:** XLNet extends the transformer model by incorporating bidirectional context information similar to BERT while maintaining an auto-regressive training objective[38]. Applications: Similar to BERT, achieving state-of-the-art results in various NLP tasks.
- **Enhanced Representation through Knowledge Integration(ERNIE):** ERNIE integrates world knowledge into pre-trained language representations, aiming to enhance the understanding of words in context[167]. Applications: Improved performance in tasks that benefit from explicit knowledge integration.

These models have significantly advanced the field of NLP, and their architectures often serve as the basis for further innovations and task-specific adaptations. Depending on the

specific NLP task, different models may be more suitable, and the choice often depends on factors such as the size of the data set, the complexity of the task, and available computational resources.

2.4.4 Large Language Models

Neural network models that have a large number of parameters, known as Large Language Models (LLM), allow them to develop and capture language patterns that are similar to those of humans. NLP applications such as Question Answering, Text Summarization, and Sentiment Analysis have seen a rise in the use of these models etc.[86].

Some of the latest large language models are listed below

- BERT: A Google-developed bidirectional model utilizing transformers for both encoding and decoding natural language.
- GPT-3: An OpenAI-created generative model employing deep neural networks to generate natural language texts.
- BLOOM: A multilingual model by Big Science employing auto-regressive methods to expand text prompts[26].
- XLM-RoBERTa: A cross-lingual model designed by Facebook, utilizing self-attention to glean insights from extensive monolingual data[19].
- XLNet: A permutation-based model co-developed by Google and CMU, leveraging transformers to capture extensive dependencies in natural language[137].
- LLaMA : Large Language Model Meta AI is an impressive and noteworthy language model for text generation[35]. It is an artificial intelligence system that harnesses

its architectural design and vast parameter capacity to comprehend and produce human language with remarkable accuracy.

- LaMDA: A language model for dialogue applications. It is a conversational language model developed by Google to handle natural-sounding texts[143].

2.5 Conclusion

This chapter discusses a review of various research works that have already been conducted in the field of Question Answering and Information Retrieval, followed by the necessary background theory needed for this research work.

Chapter 3

Corpus Creation, Pre-processing and Modelling

3.1 Introduction

This chapter discusses the creation, annotation, pre-processing, and feature vector creation of a Malayalam health Question Answering Data set - MQuAD. The following section describes document data acquisition and database creation, followed by cleaning and annotation of the Question Answering data set MQuAD. Finally, various pre-processing techniques and feature vector creation methods for Malayalam health data sets are discussed.

3.2 Data Collection and Corpus Creation

For this research work, the health corpus was gathered from several reputable sources such as TDIL [50], Kerala gov.in, etc. The corpus comprises diverse health-related raw documents encompassing information on various diseases, symptoms, remedies, and more.

The data set contains different folders as shown in Figure 3.1.

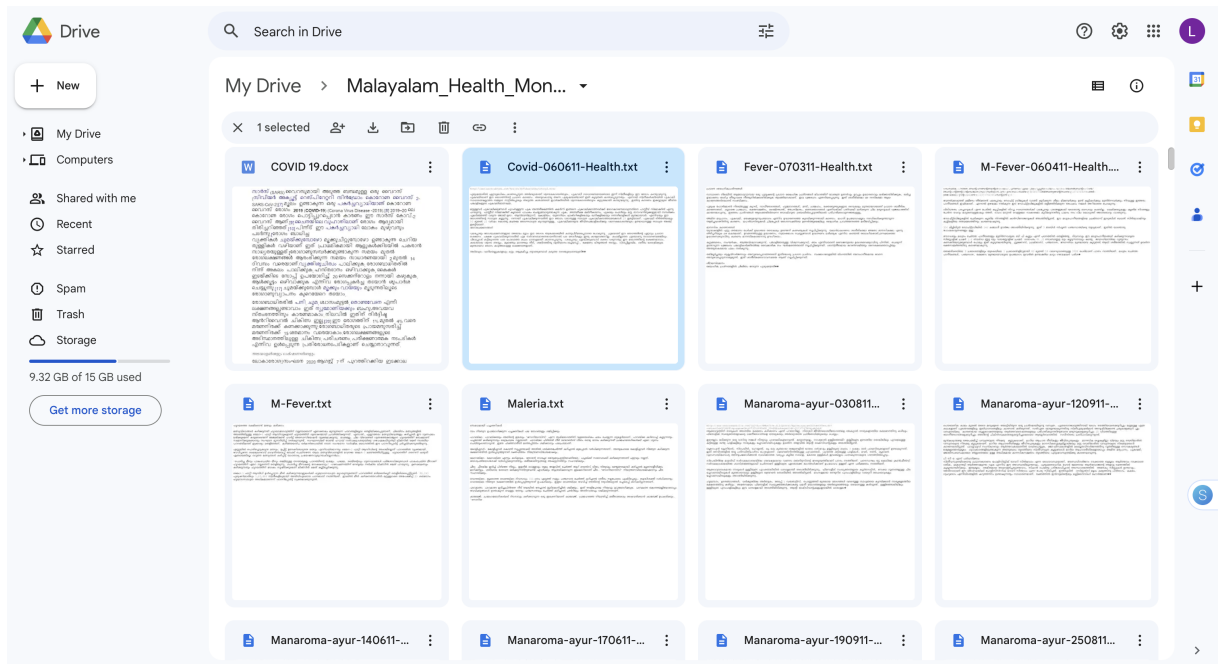


Figure 3.1: Sample Collected Data Set

The data set contains information about various diseases, symptoms, remedies, and also information related to health and wellness as shown in Figure 3.2

3.3 Data Annotation for Question Answering

After cleaning the collected health data set, the next crucial step involved employing the annotation tool, particularly the haystack [105]. This tool facilitated the generation of labels for Question Answering tasks, providing a systematic approach to annotating the data set. Utilizing haystack, adopted diverse methods to create labels [65]. One approach involved dynamically identifying questions while reading through the documents, ensuring a contextually relevant assignment of labels. Alternatively, a predefined set of questions was employed, and the corresponding answers were subsequently identified within the documents. This dual methodology allowed for a comprehensive annotation process that captured a wide array of potential queries.

Once the Question Answering data set was annotated, proceeded to export it in the SQUAD format [29]. SQUAD, developed by Stanford University, is a widely recognized Question Answering Data set [30]. The annotation process is shown in Figure 3.3 and Figure 3.4.

Documents

Id	Text	Status	Labeler no.	Created	Updated	Actions
84236	Q:എന്നാണ് പെട്ടിസിസ്റ്റിക് ഓവറി സിൻഡ്രോം? A:സ്ത്രീ ലൈംഗിക ഹോർമോണുകളുടെ അസന്തുലിതാവസ്ഥയു...	NEW	954	24-03-2021	24-03-2021	→ 🗑
84238	ഓർമ്മയുടെ അറകൾ ശൃംഗമാകാതിരിക്കാൻ 'തന്മാൽ' എന്ന സിനിമ കണ്ടപ്പോഴാണ് പലരും ഓർമ്മയുടെ വിലയിരുത്തു...	NEW	954	24-03-2021	24-03-2021	→ 🗑
84240	അക്യൂട്ട് ഗാസ്ട്രൈറ്റിസ് ആമാശയ ദിശിയിൽ പൊട്ടുന്നതെന്നുണ്ടാകുന്ന നീർവീക്കമാണ് കടുത്ത ഗാസ്ട്രൈറ്റിസ്	NEW	954	24-03-2021	24-03-2021	→ 🗑
84242	ഉൽകണ്ഠാരോഗങ്ങൾ ന്യൂറോസിസ് എന്ന പേരുകൊണ്ട് അറമ്പുകളെ ബാധിക്കുന്ന രോഗങ്ങളാണ് ഉദ്ദേശിക്കുന്നത്...	NEW	954	24-03-2021	24-03-2021	→ 🗑
84275	അപ്പൻഡിസൈറ്റിസ് അപ്പൻഡിക്സിറ്റിസാകുന്ന വീക്കമാണ് അപ്പൻഡിസൈറ്റിസ് എന്ന് അറിയപ്പെടുന്നത്. വന	NEW	954	25-03-2021	25-03-2021	→ 🗑
84279	ആസ്തമ 'കിട്ടപ്പ' എന്നർത്ഥം വരുന്ന ഗ്രീക്ക് വാക്കായ 'പാസോസി'ൽ നിന്നാണ് ആസ്തമയുടെ ഉത്ഭവം. ഗ്രീക്കു	DONE	954	25-03-2021	24-01-2022	→ 🗑
84295	ഭട്ടിസം കുട്ടികളിൽ കുട്ടികളുടെ മനോവ്യക്തിത്വ വികസനത്തിനു അസ്സമുണ്ടാക്കുന്ന രോഗങ്ങളിൽ പ്രധാനപ	NEW	954	25-03-2021	25-03-2021	→ 🗑
84302	നടുവണ യാത്രാസൗകര്യം മുതൽ വാർഷികം വരെ ഏതു പ്രായത്തിലുള്ളവരെ ബാധിക്കുന്ന സർവ്വസാമാന്യമായ ...	NEW	954	25-03-2021	25-03-2021	→ 🗑
84305	കേസമ്മർദ്ദം ജീവൻ നിലനിർത്താൻ ജീവന്റെ തുടർച്ച ശരീരത്തിൽ നിലനിർത്തുന്നത് കേസമ്മർദ്ദം	NEW	954	25-03-2021	25-03-2021	→ 🗑
88046	ശ്രദ്ധയിൽ അറ്റാക്ക് അർബുദത്തെക്കാൾ കൂടുതൽ പേർ മരണമടയാലോലം അഥവാ സ്ട്രോക്ക് നിയമം മ	DONE	954	08-04-2021	15-05-2021	→ 🗑
91178	സ്തനാർബുദം സ്ത്രീകളിൽ ഏറ്റവുമധികം കാണപ്പെടുന്ന അർബുദാവസ്ഥയാണ് സ്തനാർബുദം . ആഗോളാടി...	DONE	954	21-04-2021	09-05-2021	→ 🗑
109855	http://www.manoramaonline.com/cgi-bin/MMOnline.dll/portal/ep/malayalamContentView.do?contentId=...	NEW	954	09-05-2021	09-05-2021	→ 🗑

Figure 3.3: Data Annotation Using Haystack

U

- എന്താണ് ആസ്തമ രോഗം ? 1
- ആസ്തമ ചികിത്സ എന്തൊക്കെ ? 2
- ആസ്തമ പ്രതിരോധമാർഗ്ഗങ്ങൾ എന്തൊക്കെ ? 3
- എന്താണ് ഇൻഫെമലർ ചികിത്സ? 4
- ഇൻഫെമലർ ചികിത്സയുടെ മേന്മകൾ എന്തൊക്കെ? 5
- ഇൻഫെമലർ എത്ര തരം ? അവയുടെ ഉപയോഗക്രമം എന്ത് ? 6
- ഇൻഫെമലർ ഉപയോഗിക്കുമ്പോൾ ശ്രദ്ധിക്കേണ്ട കാര്യങ്ങൾ ? 7
- എന്താണ് അലർജി ട്രൈപ്പിൾ ? 8
- എന്താണ് ഇമ്മ്യൂണോതെറാപ്പി ? 9

ADD CUSTOM QUESTION

രോഗികളുടെ നിർദ്ദേശപ്രകാരം മാത്രമേ ഇൻഫെമലർ ഉപയോഗിക്കാവൂ. മറ്റ് ആസ്തമാരോഗികൾ ഉപയോഗിക്കുന്ന ഇൻഫെമലർ നിങ്ങൾക്ക് യോജിച്ചതാവണമെന്നില്ല.

ഇൻഫെമലർ ഉപയോഗിച്ചുതുടങ്ങിയതിൽ വെച്ച് കിടപ്പിന്റെ നിർദ്ദേശപ്രകാരം മാത്രമേ ഉപയോഗം നിർത്താവൂ. രോഗം നിയന്ത്രണവിധിയായാൽ ശ്രമമേ ഇൻഫെമലറിന്റെ അളവ് എണ്ണവും കുറച്ചു നോക്കാവൂ. പലർക്കും ഇൻഫെമലറിന്റെ ഉപയോഗം ഒഴിവാക്കാനുമാവാം.

ഇൻഫെമലർ ഉപയോഗിക്കുമ്പോൾ അതിന്റെ മൗൽസീസ് വായ തുറന്ന് ചുണ്ടിന്റെയും പല്ലിന്റെയും ഉള്ളിലായി നാക്കിന്റെ മുകളിലായി വിലമ്പണം വയ്ക്കാവുന്നതാണ്.

ഇൻഫെമലർ ഉപയോഗിച്ചുകഴിയുമ്പോൾ വായും തോണ്ടും നന്നായി കഴുകുക.

രോഗഫലവും സ്പെന്റും 34 ദിവസത്തിലൊരിക്കൽ കഴുകാൻ മതിയാവാം. കഴുകിയാൽ ഉണങ്ങിയതിനുശേഷം മാത്രം ഉപയോഗിക്കുക. വെയിലത്തുവെച്ചാ പാന്നിന്റെ കാഴ്ചപ്പിടിച്ചാ ഉണക്കുക. തുണികൊണ്ടും മറ്റും തുടച്ച് പൊൻ വിഴിഞ്ഞിയാൽ അവയ്ക്ക് കൂടുതൽ മരുന്ന് പറ്റിപ്പിടിക്കാനിടയാകും.

ഒരിക്കൽ ഇൻഫെമലർ ഉപയോഗിച്ചിട്ട് പലം കിട്ടിയിട്ടില്ല എന്ന കാരണത്താൽ ഈ ചികിത്സയിൽ വിശ്വാസം നഷ്ടപ്പെട്ടവർ ധാരാളമുണ്ട്. എന്നാൽ ഇത് ശരിയല്ല. കാരണം, ഒരിക്കൽ ഉപയോഗിച്ച ശേഷം കൂടാതെ മറ്റേതെങ്കിലും കാരണങ്ങൾ ഇൻഫെമലർ രൂപത്തിൽ ലഭ്യമാണല്ലോ. അവയിൽ മറ്റൊന്ന് പലിടിച്ചാക്കാം. മാത്രമല്ല, ആദ്യം ശരിയായ രീതിയിലും സോസിപ്പോലായിരിക്കില്ല ഉപയോഗിച്ചതെന്നും വാദം.

എല്ലാ ആസ്തമാരോഗികളെയും ഇൻഫെമലർ ഉപയോഗിച്ചു മാത്രം ചികിത്സിക്കാൻ കഴിഞ്ഞെന്നുവരില്ല. അങ്ങനെയുള്ളവർക്ക് ഗുളികകൾ, സിറപ്പുകൾ, കുത്തിവെപ്പുകൾ മുതലായവ വേണ്ടിവരാം. തിയാഹിലിൻ, ബാബ്ബിനോൾ, മോസ്റ്റിലുകാസ്റ്റ്, കിറ്റോറിഫൻ, സ്റ്റിറോയിഡ് ഗുളികകൾ മുതലായവ കൂടുതൽ കടുത്ത രോഗങ്ങളുമായി ബന്ധിതമാകുന്നു.

ആസ്തമയോടൊപ്പം പലപ്പോഴും അലർജിക് റൈനൈറ്റിസും കണ്ടുവരുന്നു. ഇക്കൂട്ടത്തിൽ അലർജിക് റൈനൈറ്റിസ് നിയന്ത്രിച്ചാൽ മാത്രമേ ആസ്തമയും നിയന്ത്രണവിധിയായാകൂ. ആന്റിഹിസ്റ്റമിൻ, സ്റ്റിറോയിഡ് എന്നിവയടങ്ങിയ ഗുളികകളും മുക്കിലടങ്ങുന്ന സ്പ്രേകളുമാണ് ഈ രോഗത്തിന്റെ ചികിത്സക്കുപയോഗിക്കുന്നത്.

അലർജി ട്രൈപ്പിൾ ഇമ്മ്യൂണോതെറാപ്പി

അലർജികളിൽ എതിരോടാക്കെയാണ് ഒരു വ്യക്തിക്ക് അലർജി എന്നു കണ്ടുചിടാൻ നടത്തുന്ന പരിരോധനയാണ് അലർജി ട്രൈപ്പിൾ. വിവിധതരം അലർജികൾ ടൈപ്പിങ്ങാൻ ചെറിയ അളവിൽ കുത്തിവെച്ച് ചുറ്റും ഉണ്ടാവുന്ന റിയാക്ഷൻ വിലയിരുത്തിയാണ് ഇത് സാധിക്കുന്നത്. അലർജി ട്രൈപ്പിൾ കൊണ്ടുള്ള പ്രയോജനങ്ങൾ രണ്ടാണ്.

- 1) ട്രൈപ്പിൾ ചെമ്പ് അലർജിയുണ്ടാകുന്ന കണ്ടെത്തിയ അലർജികളുമായുള്ള സമ്പർക്കം ഒഴിവാക്കാം. (പ്രത്യേകിച്ചും ആഹാരസാധനങ്ങൾ)
- 2) രോഗിയെ ഇമ്മ്യൂണോതെറാപ്പിക്ക് വിധേയനാക്കാം.

ഇമ്മ്യൂണോതെറാപ്പി

Figure 3.4: Data Annotation Process

The data set can be exported in various formats, including JSON, Excel, and CSV files.

The data set, MQuAD obtained after annotation is shown below.

The annotation process is shown in Figure 3.5 and Figure 3.6.

```
മണിക്കൂർ\nനഗർഭിന്നികൾ\nഎട്ടു മണിക്കൂർ മുതൽ മുകളിലേക്ക്",
  "document_id": 109994
}
],
},
{
  "paragraphs": [
    {
      "qas": [
        {
          "question": "ഗ്യാസിൻറെ അസ്ഥമത ഉണ്ടാകാൻ കാരണം എന്ത് ? പരിഹാര മാർഗങ്ങൾ എന്തെല്ലാം ?\n",
          "id": 142525,
          "answers": [
            {
              "answer_id": 206959,
              "document_id": 109995,
              "question_id": 142525,
              "text": "ആഹാരം കഴിച്ചു തുടങ്ങുമ്പോഴേ വിശപ്പു കുറയുക, പുകച്ചിൽ, ഗ്യാസ് തുടങ്ങിയവയ്ക്ക് കാരണം അന്നനാളത്തിലോ ചെറുകുടലിൻറെ ആദ്യ ഭാഗങ്ങളിലോ ഉണ്ടാകുന്ന പ്രശ്നങ്ങളാണ്. വയറിനുള്ളിൽ അമിതമായി അല്പം ഉൽപാദിപ്പിക്കപ്പെടുക, അല്പത്തെ ചെറുക്കുന്ന ചെറുകുടലിൻറെ ദിത്തിക്ക് കട്ടി കുറയുക, തുടർന്ന് അല്പം അന്നനാളത്തിലേക്കു തിട്ടി വരിക തുടങ്ങിയവയും ഇത്തരം അസ്ഥമതകൾക്കു കാരണങ്ങളാണ്. ഇതിൻറെ അനന്തരഫലമായി ചിലരിൽ ഭ്രാന്തപ്പോലോ കാണപ്പെടുന്നു. ഹെർക്കാൻ ബുദ്ധിമുട്ടുള്ള അന്നം ആഹാരത്തിൽ അധികമാകുമ്പോഴാണ് പൊതുവെ ഗ്യാസ്സ്രബിൾ ഉണ്ടാകുന്നത്.\n\nഗ്യാസ്സ്രബിൾ പോലുള്ള വയറിനുള്ളിലെ അസ്ഥമതകളെ ചെറുക്കാൻ ചില കാര്യങ്ങൾ ശ്രദ്ധിക്കാം.\n\n• എന്നും ഒരേ സമയത്തു ഭക്ഷണം കഴിക്കുക. അപ്പോൾ ശരിയെത്തിനു ഭവനസങ്ങളുടെ ഉൽപാദകാര്യത്തിൽ ഒരു ചിട്ട കൈവരും.\n\n• വയർ ഭക്ഷണമില്ലാതെ കാലിയായി കിടക്കാൻ ഒരിക്കലും അനുവദിക്കരുത്. വയർ കാലിയായി കിടക്കുമ്പോഴാണ് അസിഡിറ്റി കൂടുന്നത്. ഇടനേരങ്ങളിലും വയർ കാലിയാണെന്നു തോന്നുമ്പോൾ ലഘുവായ ആഹാരം കഴിക്കണം.\n\n• പഴകിയ ഭക്ഷണസാധനങ്ങൾ തീർത്തും ഒഴിവാക്കുക.\n\n• എരിവ്, പുളി ഇവ കലർന്ന ആഹാരം, ചോക്കലേറ്റ്, ചായ, കാപ്പി എന്നിവ മിതമായേ കഴിക്കാവൂ. അച്ചാറും മറ്റും അധികമായി കഴിക്കരുത്.\n\n• ഭക്ഷണം നന്നായി ചവച്ചുറച്ചേ കഴിക്കാവൂ. ധൃതിയിൽ കഴിക്കുമ്പോൾ ഭക്ഷണത്തോടൊപ്പം കുറേയധികം വായുവും ആമാശയത്തിൽ പ്രവേശിക്കും. തുടർന്നു പുളിച്ചു തിട്ടിൽ, വയർ കമ്പിക്കൽ തുടങ്ങിയ അസ്ഥമതകളുണ്ടാകും.\n\n• ഭക്ഷണം കഴിക്കുന്നതിനിടയിൽ സസാരിക്കുക, ഉറക്കെ ചിരിക്കുക തുടങ്ങിയ ശീലങ്ങൾ പാടില്ല.\n\n• ഭക്ഷണത്തിൻറെ രുചി, മണം, കാണുന്നതിലുള്ള ആകർഷണം എന്നിവയെല്ലാം ഭവനസങ്ങളുടെ ഉൽപാദനത്തെ ഉദ്ദീപിപ്പിക്കുന്നു. അതുകൊണ്ടു ഭക്ഷണം സാവധാനം ആസ്വദിച്ചു കഴിക്കുക.\n\nആഹാരത്തിയിൽ ഈ കാര്യങ്ങൾ ശ്രദ്ധിക്കുമ്പോൾ തന്നെ നിങ്ങളുടെ വയറ്റിലെ അസ്ഥമതകൾ മാറിയേക്കും. ഇതുകൊണ്ടൊന്നും ഒരു വ്യത്യസ്തവും വരുന്നില്ലെങ്കിൽ മാത്രമേ മരുന്നിൻറെ ആവശ്യം വേണ്ടി വരുന്നുള്ളൂ.",
              "answer_start": 238,
              "answer_category": null
            }
          ]
        }
      ]
    }
  ],
  "is_impossible": false
},
{
  "question": "ഗ്യാസ്സ്രബിൾ പോലുള്ള വയറിനുള്ളിലെ അസ്ഥമതകളെ ചെറുക്കാൻ ചെമ്പുണ്ട കാര്യങ്ങൾ എന്തെല്ലാം ?\n",
  "id": 142528,
  "answers": [
    {
      "answer_id": 206962,
```

Figure 3.5: Annotated Data in JSON Format

Category	A	B	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
805		ഗർഭകാലത്ത് എത്ര തൂക്കം കൂടണം?	ആദ്യത്തെ മൂന്നു മാസം തൂക്കം കൂടില്ല. ചിലപ്പോൾ കുറഞ്ഞുവെന്നും വരാം. പിന്നീട് ആഴ്ചയിൽ അര- ഒന്ന് കി ഗ്രാം അനുസരിച്ചു കൂടുകയും ചെറുവളരെ അപൂർവ്വമായി തൂക്കം കുറഞ്ഞും വരാറുണ്ട്. കുഞ്ഞിൻറെ വളർച്ചക്കുറവും ആഗിരണം ചെയ്തിരിക്കുന്ന ലായനിയുടെ കുറവും ഇതിനു കൂടി ചിലപ്പോൾ തൂക്കം അധികമാകാറുണ്ട്. രക്തസമ്മർദ്ദം കൂടുന്നതിൻറെ മുൻപായുള്ള നിരുകൊണ്ടും വെള്ളം കെട്ടുന്നതുകൊണ്ടും വലിയ കൂട്ടി ആഴ ഇതിനു കാരണമാകാം. ഇത്തരം അവസ്ഥകളിൽ പ്രത്യേകം ശ്രദ്ധ വേണം.																
806	213752	ഗർഭകാലത്തെ അമിതവണ്ണം അപകടമാണോ?	ഗർഭകാലത്തുണ്ടാകുന്ന അമിതവണ്ണം പ്രസവശേഷം കുറയ്ക്കുന്നത് ഏറെ ശ്രമകരമാണ്. ചിലർക്കു ഗർഭകാലത്തെ അമിതവണ്ണം മൂലം പിൻക്കാലത്ത് കൂടുതൽ കരുതി ഡയറ്റിങ് നടത്തുന്നതു ഗർഭകാലത്ത് അപകടമാണ്. അമിതവണ്ണമുള്ളവരിൽ പ്രസവത്തിൽ ബുദ്ധിമുട്ടുകൾ നേരിടേണ്ടി വരാം.																
807	213753	ഗർഭകാലത്തെ അമിതവണ്ണം അപകടമാണോ?	എല്ലാ ദിവസവും ശുദ്ധ വായു കിട്ടുന്ന അന്തരീക്ഷത്തിൽ അരമണിക്കൂർ നടക്കുന്നതു സ്പല്ലത്. ഇടുവെല്ലുകൾക്ക് അയവു വരുത്താനുതകുന്ന വ്യക്തികളിൽ വ്യായാമങ്ങൾ ഒഴിവാക്കുക. ഏതായാലും ഗർഭിണിക്കു കഠിന വ്യായാമം വേണ്ട. അപൂർവ്വമായി ഭ്രമണം ഗർഭപാത്രത്തിനു പുറത്ത്, അതായത് അസുഖവാഹിനിക്കുഴലിമോ, അസ്വാശയത്തിമോ വധിമോ പറ്റിപ്പിടിച്ചു വളമാൻ ഉണ്ടാകാൻ സാധ്യമാണ്.																
808	213767	എന്താണ് ട്യൂബൽ പ്രഗ്നൻസി? ഇത് എത്രത്തോളം അപകടകരമാണ്?	ഗർഭത്തിൻറെ ആദ്യഘട്ടത്തിൽ അമിതമായി വയറുവേദന, പ്രത്യേകിച്ച് ഒരു വശത്തുള്ള വേദന, നടുവുകുഴപ്പ്, രക്തംപോക്ക് ഇവ കണ്ടാൽ ഇടൻറെ ധന നടത്തേണ്ടതാണ്. ആദ്യഘട്ടത്തിൽ ഗർഭം അലസുന്നതിൻറെ ലക്ഷണങ്ങളും ഏറെക്കുറെ സ്ഥാനമാണ്.																
809	213756	ഗർഭകാലത്തു സൗന്ദര്യ പ്രശ്നങ്ങൾ കൂടുമോ? മുഖത്തും കഴുത്തിലും മാർക്ക് കൂടുമോ?	ഗർഭകാലത്തു ചിലർക്കു കഴുത്തിലും മാറിടത്തിലും വയറ്റിലും കറുത്ത പാടുകൾ കൂടുതലായി ഉണ്ടാകാറുണ്ട്. നീരുവരുന്നതുകൊണ്ടും മൂലം വികസിച്ചിട്ടുണ്ട്. വയറിൽ സ്പ്രെച്ച് മാർക്ക് ഉണ്ടാകും. ഇവ കുറയ്ക്കാനായി പലതരം ക്രിമുകൾ ഉണ്ടു ന്യുലമോണ്. പ്രസവാനന്തരമുള്ള തേപ്പുകൂടി കറുത്തപാടുകളൊക്കെ സാധാരണ മാറാറുണ്ട്.																
810	213758	ഗർഭകാലത്തു സൗന്ദര്യ പ്രശ്നങ്ങൾ കൂടുമോ? മുഖത്തും കഴുത്തിലും മാർക്ക് കൂടുമോ?	സ്പ്രെച്ച് മാർക്ക് കുറയ്ക്കാൻ വെളുത്ത പാടായി കിടക്കാറുണ്ട്. പ്രത്യേകിച്ചും ഒന്നിൽ കൂടുതൽ ഗർഭം ധരിച്ചവർക്കും വലിയ കൂട്ടിയെ ഗർഭം ധരിച്ചവർക്കും സ്പ്രെച്ച് മാർക്ക് കൂടുതലായി കാണാനും ഇടയാക്കും.																

Figure 3.6: Annotated Data in Excel Format

For the specific requirements of our research work, opted for the JSON format, which encapsulated distinct fields for questions, IDs, and corresponding answers for each document. This structured format not only facilitated easy handling of the data set but also aligned with the standards set by Stanford University’s SQUAD, ensuring compatibility with established practices in the field of Question Answering research. The MQuAD corpus contains more than 50000 question-and-answer pairs. The complete description of the MQuAD corpus is shown in Table 3.1.

The MQuAD Corpus comprises a dataset of 50,000 questions and their corresponding answers related to health. The questions exhibit a range of word lengths, with a minimum of 3 and a maximum of 14 words. On average, the questions are approximately 5 words long. Similarly, the answers in the corpus vary in length, with a minimum of 18 words and a maximum of 260 words. The average word length of the answers is approximately 120

Table 3.1: Description of MQuAD Corpus

Total number of questions	50000
Total number of answers	50000
Minimum word length of questions	3
Maximum word length of questions	14
Average word length of questions	5
Minimum word length of answers	18
Maximum word length of answers	260
Average word length of answers	120

words. This diversity in word lengths reflects the richness and complexity of the MQuAD Corpus, providing a comprehensive and varied set of linguistic inputs and outputs for the process of Question Answering.

3.4 Data Pre-processing of Queries and Documents

Pre-processing stands as the initial and crucial phase in every NLP application, and its significance is particularly pronounced in the context of Question Answering [100]. This pivotal step significantly influences the effectiveness of subsequent processes. In the scope of this study, various pre-processing methods are employed for both queries and documents to enhance the overall performance of the Question Answering System. These pre-processing techniques encompass a range of approaches, such as tokenization, stemming, lemmatization, and syntactic analysis, each contributing to refining the input data for more effective information extraction and retrieval [94]. The careful application of these pre-processing methods plays a solid foundation for the subsequent stages of the Question Answering framework.

3.4.1 Tokenization

Tokenization is a fundamental pre-processing method, that involves breaking down a continuous stream of text into individual words or tokens. In the specific context of this study, word tokenization is employed as the chosen method for segmenting both queries and documents [85]. This process plays a pivotal role in facilitating subsequent analysis and comprehension of the textual content.

For the implementation of word tokenization, the Natural Language Tool Kit (NLTK) is utilized. NLTK provides a comprehensive set of libraries and resources for processing human language, making it an ideal choice for breaking down complex textual data into meaningful units [42]. The adoption of NLTK ensures the accuracy and efficiency of the tokenization process, setting the stage for further linguistic analysis and semantic understanding in the Question Answering System [31].

3.4.2 Stop Word Removal

In the area of NLP applications, stop words, which carry minimal semantic meaning, are typically excluded from the text to enhance processing efficiency and focus on more significant terms [66]. In the context of Question Answering, stop words are generally disregarded as they don't contribute substantially to the identification of keywords or the overall understanding of the content [40].

To remove stop words the Natural Language Tool Kit (NLTK) is used. NLTK offers a predefined list of stop words, and in the case of this research, a specific set of Malayalam

stop words can be configured [57]. By utilizing NLTK's capabilities, these stop words are easily filtered out during the Question Answering task, ensuring that the system concentrates on relevant and informative content for improved accuracy and performance.

3.4.3 Stemming and Lemmatization

Stemming and lemmatization both are used to identify the root word of a particular word in documents.

- **Stemming:-** Stemming is a pivotal technique employed to derive the 'stem' or 'root' of a given word by eliminating affixes, prefixes, and suffixes. The purpose of stemming is to distill words down to their fundamental, base form, thereby reducing variations and allowing for a more generalized representation. In this particular study, an Indic stemmer[99], integrated with the Natural Language Tool Kit (NLTK), was utilized to perform the stemming process. This technique aids in standardizing word forms, enabling more effective matching and retrieval in the context of Question Answering tasks.
- **Lemmatization:-** Lemmatization is a linguistic method employed to derive the lemma or canonical form of a word in a document. Unlike stemming, lemmatization considers the context of the words, utilizing detailed lexical and grammatical information to identify the base or root form[60]. The objective of lemmatization is to ensure a more accurate representation of words, mapping them to their dictionary or lemma forms. The choice between stemming and lemmatization depends on the specific requirements of the task at hand. While stemming provides a quicker and less precise approach, lemmatization offers a more refined and context-aware solution, suitable for tasks that demand a higher level of linguistic accuracy.

3.4.4 Bag of Words

The text needs to be transformed into a numerical representation to be processed and its meaning understood. One technique for turning text into numerical form is a bag of words. Every word that appears in a sentence is represented by a bag of words. Word occurrence is determined using the documents' vocabulary. It indicates whether or not a specific word appears in the document, but it gives no details about the word's placement or sequence [45].

3.5 Vector Space Representation and Document Modelling

In the context of Question Answering, the representation of documents and queries involves the utilization of index terms derived from the pre-processing steps. These index terms serve as a refined and processed version of the original text [49], incorporating techniques such as tokenization, stop word removal, stemming, and others. The purpose of employing these index terms is to capture and express the relevance of different terms within both the query and document data sets [128].

To further refine this representation, numerical weights are assigned to each index term, reflecting the varying degrees of importance or relevance. The assignment of weights enables a nuanced understanding of the significance of different terms in the context of the

information retrieval process. This weighted representation facilitates the development of models and algorithms that can more effectively discern and match relevant information, ultimately enhancing the accuracy and efficiency of Question Answering Systems. The weighted index terms serve as a foundational element in the construction of models that aim to provide precise and contextually relevant answers to user queries.

- The Vector Space Model (VSM):- VSM constitutes a fundamental mathematical framework extensively applied in the field of Question Answering and Natural Language Processing. This model serves the purpose of representing text documents as vectors within a high-dimensional space. Its mathematical formulation provides a robust means to capture the semantic relationships inherent in words and documents, facilitating a range of computational tasks including document retrieval, clustering, and Information Retrieval.

The Vector Space Model represents document D as an m -dimensional vector, where each dimension is a distinct phrase that appears in all of the documents in the collection. The total number of dimensions, denoted by m , reflects the comprehensive set of terms utilized across the document collection. This vector representation enables a quantitative expression of the semantic structure of documents, paving the way for the application of various algorithms and techniques to extract meaningful insights and support sophisticated natural language processing tasks.

The Vector Space Model thus serves as a foundational framework [81], allowing for the transformation of textual information into a structured and computationally accessible format. Its versatility makes it a valuable tool in the development of systems and applications geared towards effective information retrieval and analysis in the realm of Question Answering and beyond.

In this framework, each document is depicted as a vector within a multi-dimensional space, where the dimensions of this space are intricately tied to the terms or words present across the entirety of the document collection. This representation transforms textual information into a numerical format, laying the foundation for computational analysis and Information Retrieval.

The Vector Space Model (VSM) facilitates a systematic encoding of documents, allowing for a different exploration of their semantic relationships. The dimensions of the multi-dimensional space are essentially a reflection of the diverse set of terms encompassing the entire document corpus. Consequently, each term contributes to the positioning of documents within this high-dimensional space, offering a quantitative representation of the semantic content.

This vector-based representation brings forth a rich foundation for various computational applications. Models operating within this vector space can measure similarities, distances, and patterns, enabling tasks such as document retrieval, clustering, and Information Retrieval. The Vector Space Model's adaptability renders it a crucial asset in the field of natural language processing, particularly in systems and applications where understanding and identifying semantic relationships are paramount.

The weights w_i of terms t_i are generated using different techniques such as Term Frequency (TF) modeling, Inverse Document Frequency (IDF) modeling, and Term Frequency-Inverse Document Frequency (TF-IDF) modeling [5]. They are explained in detail below.

3.5.1 Term Frequency (TF) Modelling

In vector space modeling, the frequency of term i in the document d_j . The t_f factor is given by, Scale f_{ij} relative to the frequency of other terms in the document. Term frequency (TF) is given in equation 3.1,

$$tf_{ij} = f_{ij}/m_i \quad (3.1)$$

when $f_{ij} \geq 0$. Let m_i be $\max(f_{ij})$, i.e., m_i is the maximum frequency of any term in document d_j .

3.5.2 Inverse Document Frequency (IDF) Modelling

A term that is used sparingly in documents is probably going to be more discriminatory than a term that is used frequently or in every document. Inverse document frequency modeling can be applied in this case. Suppose there are N documents and that the number of documents in which term j occurs is n_j . Assume that there are N documents total, and that term j appears in n_j of those documents. Using N/n_j as the inverse document frequency could be one approach. This method over-emphasizes small differences, by using a logarithm. Inverse Document Frequency (IDF) is given in the equation 3.2

$$idf_j = \log_2(N/n_j) + 1 \quad (3.2)$$

where $n_j \geq 0$.

Where N is the total number of documents and n_i is the number of documents in which the term j occurs.

3.5.3 Term Frequency-Inverse Document Frequency (TF-IDF)

Modelling

TF-IDF modeling is the typical weighting strategy for free text materials. According to this method, each term is given a weight based on how frequently it occurs throughout the collection of documents as a whole as well as in a specific document. The term frequency of the number of times the term t_i appears in document D is the first component of the TF-IDF technique. The second component is computed as the following equations and is referred to as the inverse document frequency.

$$t_{ij} = tf_{ij} \times idf_j = (f_{ij}/m_i) \times (\log_2(N/n_j) + 1) \quad (3.3)$$

when n_j and $gt;0$.

ie,

$$(weight\ of\ term\ j\ in\ document\ i) = (term\ frequency) \times (inverse\ document\ frequency) \quad (3.4)$$

Where N is the total number of documents in the collection and n_j , is the number of documents in which the term t_{ij} appears at least once.

3.6 Conclusion

This chapter presents the query and document data set creation, annotation, and document pre-processing and modeling techniques. First of all, collected different Malayalam health documents. Then annotate these documents into SQUAD format to create the MQuAD data set. Different pre-processing techniques used in this study are tokenization, stop word removal, and stemming. After the processing,

the documents are converted as vectors. These vectors are then used as input for further processing, which will be discussed in succeeding chapters.

Chapter 4

Word Embedding Based Question Answering Model

4.1 Introduction

This chapter introduces a Question Answering approach that utilizes Word Embeddings for enhanced performance. The primary objective of this study is to bridge the gap between human language comprehension and machine reasoning, ultimately providing users with accurate and relevant answers to their queries [125]. The incorporation of Word Embeddings stands out as a crucial advancement, significantly elevating the efficiency and effectiveness of our Question Answering System.

Word Embedding serves as a cornerstone technique in NLP, empowering machines to represent words in a document as numerical vectors. What distinguishes Word Embedding is its ability to capture and consider semantic interrelationships among words. By encoding words into continuous vector spaces, the model gains a deeper understanding of the contextual and semantic information present in the document. This transformative approach contributes to the system's ability to comprehend the

intricacies of human language and provide meaningful responses to user queries. The incorporation of Word Embedding contributes to the accuracy of the system and enriches its capability to navigate the semantic complexities inherent in natural language, thereby advancing the accuracy of the Question Answering process.

4.2 Word Embedding

Word Embedding stands as a pivotal method in both NLP and ML for representing words in a numerical format. This methodology is particularly impactful due to its capacity to encapsulate the semantics and identify the relationships between words within a given context [102]. The sample Word Embedding for Malayalam words as shown in Figure 4.1 By adopting Word Embedding, move beyond more symbolic representations, allowing our models to utilize the semantic information present in the language.

Various approaches can be employed for Word Embedding, ranging from probabilistic models to neural networks. These methods endeavor to capture the contextual essence of words, enabling machines to understand individual words and the relationships and meanings they share within a given textual context. The significance lies in the ability to translate linguistic expressions into vector spaces, providing a numerical representation that encapsulates the semantic relationships between words.

In Word Embedding, the underlying principle is that semantically similar words should possess analogous representations. Consequently, when words are embedded into vector spaces, those vectors become indicative of the semantic similarities between the corresponding words. This inherent property makes Word Embeddings

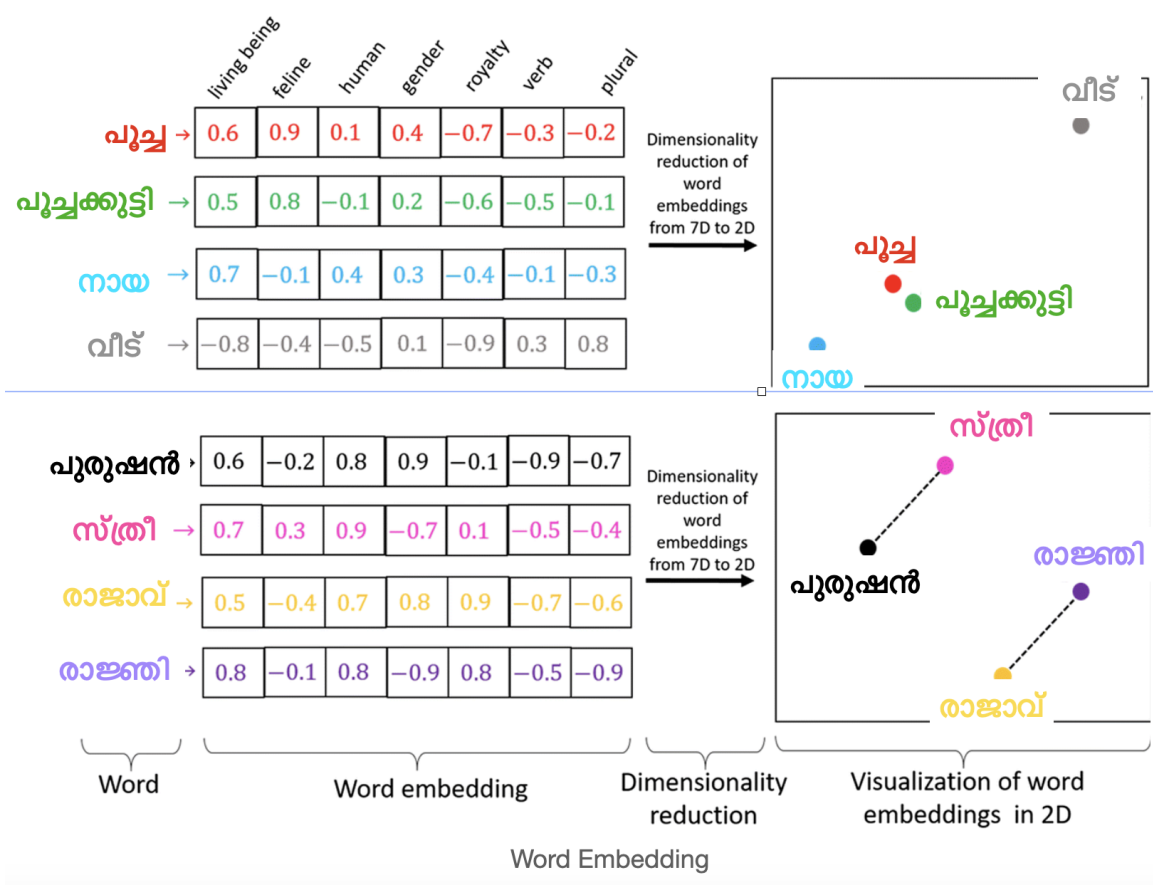


Figure 4.1: Sample Word Embedding

particularly valuable in NLP tasks, such as Question Answering and Information Retrieval where the relevance and contextuality of words are paramount.

In this specific implementation of semantic Question Answering, this work utilized the Word2Vec model for Word Embedding. Word2Vec, a powerful and widely used Word Embedding model, employs neural networks to learn distributed representations of words. This model excels in capturing semantic relationships, and its application in our system contributes to the generation of highly relevant answers for NLP tasks, particularly in the domain of Question Answering.

In essence, the utilization of Word2Vec as our chosen Word Embedding model exemplifies our commitment to harnessing advanced methodologies that excel in capturing the intricate semantics of the language. This method improves our system's accuracy and puts it in a better position to handle the complex world of natural language by giving it a richer comprehension of word relationships and semantics.

4.2.1 Word2Vec

Word2Vec is a popular tool that embodies neural networks' ability to capture the subtleties of language semantics. This three-layered neural network operates by transforming words from a higher-dimensional vector space to a lower-dimensional one, providing a condensed yet information-rich representation of each word. The distinctive feature of Word2Vec lies in its ability to represent words numerically and consider the semantic relationships between them during the modeling process.

In Word Embedding the semantic relationships between words within the document are also considered while modeling. In the vector space generated by Word2Vec, words with analogous meanings are strategically positioned in proximity to each

other. This spatial arrangement reflects the model's ability to discern and encapsulate semantic similarities. Word Embedding was implemented by Word2Vec algorithms, and offers two primary methodologies: Continuous Bag Of Words (CBOW) and Skip-gram. These techniques cater to different aspects of language modeling, providing flexibility and adaptability to various linguistic contexts.

Continuous Bag of Words (CBOW)

CBOW is a distinctive variant of the Word2Vec Word Embedding technique, which operates with the specific goal of predicting the target word, based on its surrounding context[166]. This predictive modeling is facilitated through the deployment of a three-layered neural network architecture, which plays a crucial role in processing the CBOW model. Neural Network Architecture in CBOW is shown in Figure 4.2.

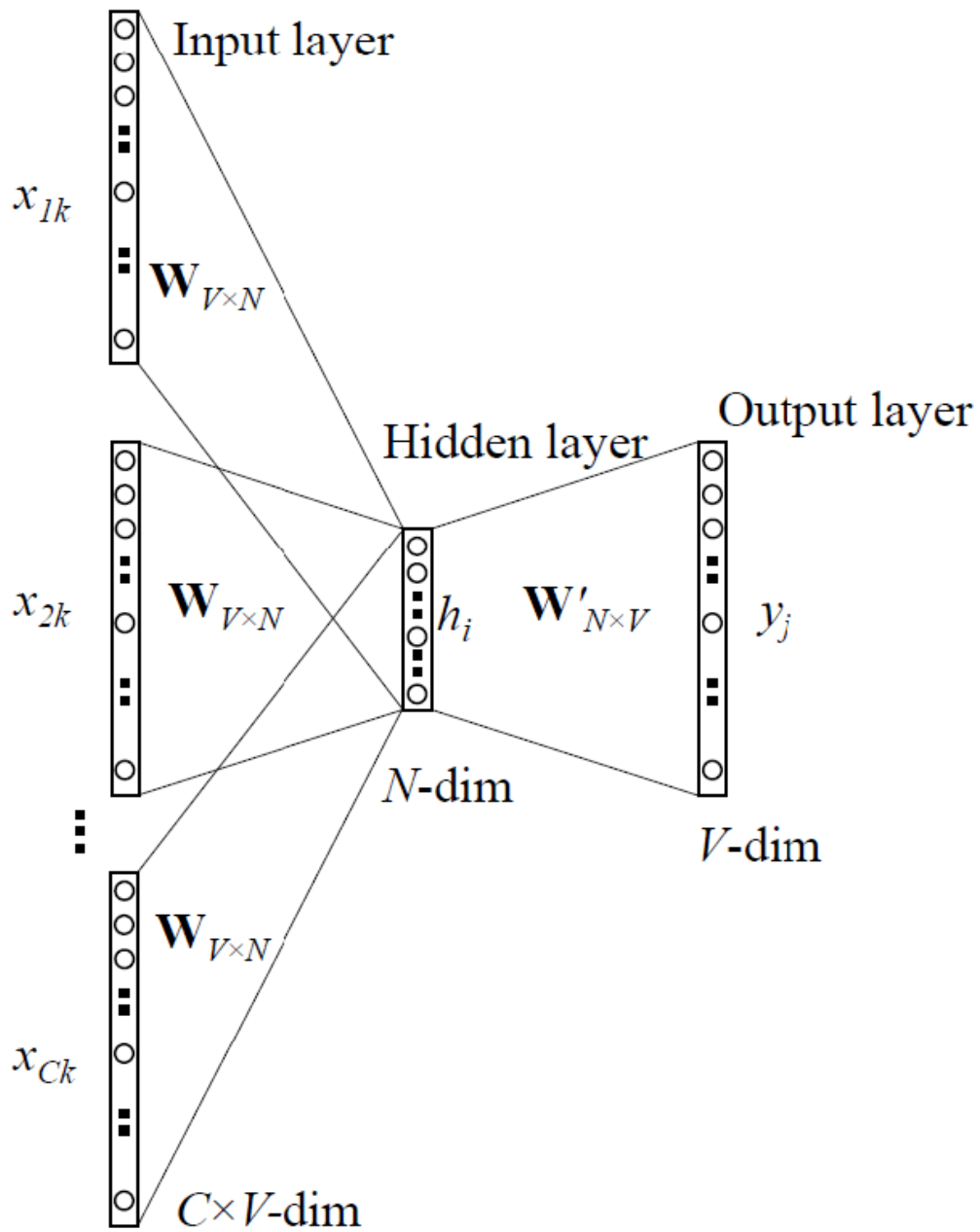


Figure 4.2: CBOW-Architecture

[113]

- Input Layer:- The input layer of the CBOW model is designed to correspond to the context of the target word. It is essentially a representation of the words surrounding the target word within a specified window size. Each word in

the context is encoded as a vector and serves as an input node in the neural network.

- Hidden Layer:- The hidden layer of the CBOW model acts as an intermediary stage where the input vectors are projected onto a set of weights. This projection involves transforming the input vectors into a representation that captures the relationships between the context words. The weights in the hidden layer correspond to the projection of each word from the input layer.
- Output Layer:-

The output layer is responsible for predicting the target word based on the transformed information from the hidden layer. Each node in the output layer represents a possible word from the vocabulary. The goal of the CBOW model is to maximize the probability of predicting the correct target word given a particular context.

The training process of CBOW involves adjusting the weights in such a way that the model becomes proficient at predicting the target word accurately. This is achieved by maximizing the probability of selecting the correct target word given the context information.

During the training period, the CBOW model learns to adjust the weights in both the hidden and output layers to optimize its ability to predict the target word. The optimization process involves iteratively updating the weights based on the errors in the predictions, bringing the model closer to an accurate representation of the relationships between words in the specified context.

In essence, the CBOW model excels at capturing the contextual dependencies of words within a given window, facilitating the creation of meaningful Word Em-

beddings that encapsulate the semantic relationships prevalent in the language. Its adaptability and efficiency make CBOW a valuable tool in natural language processing tasks that require a nuanced understanding of context and semantics.

Skip-Gram

The Skip-Gram model, a complementary counterpart to the Continuous Bag of Words (CBOW) model within the Word2Vec framework, operates in a reverse fashion by predicting context words based on a given target word [15]. This unique approach makes Skip-Gram particularly adept at capturing the diverse contextual usages of a word within a specified window. The architecture and functioning of the Skip-Gram model is shown in Figure 4.3.

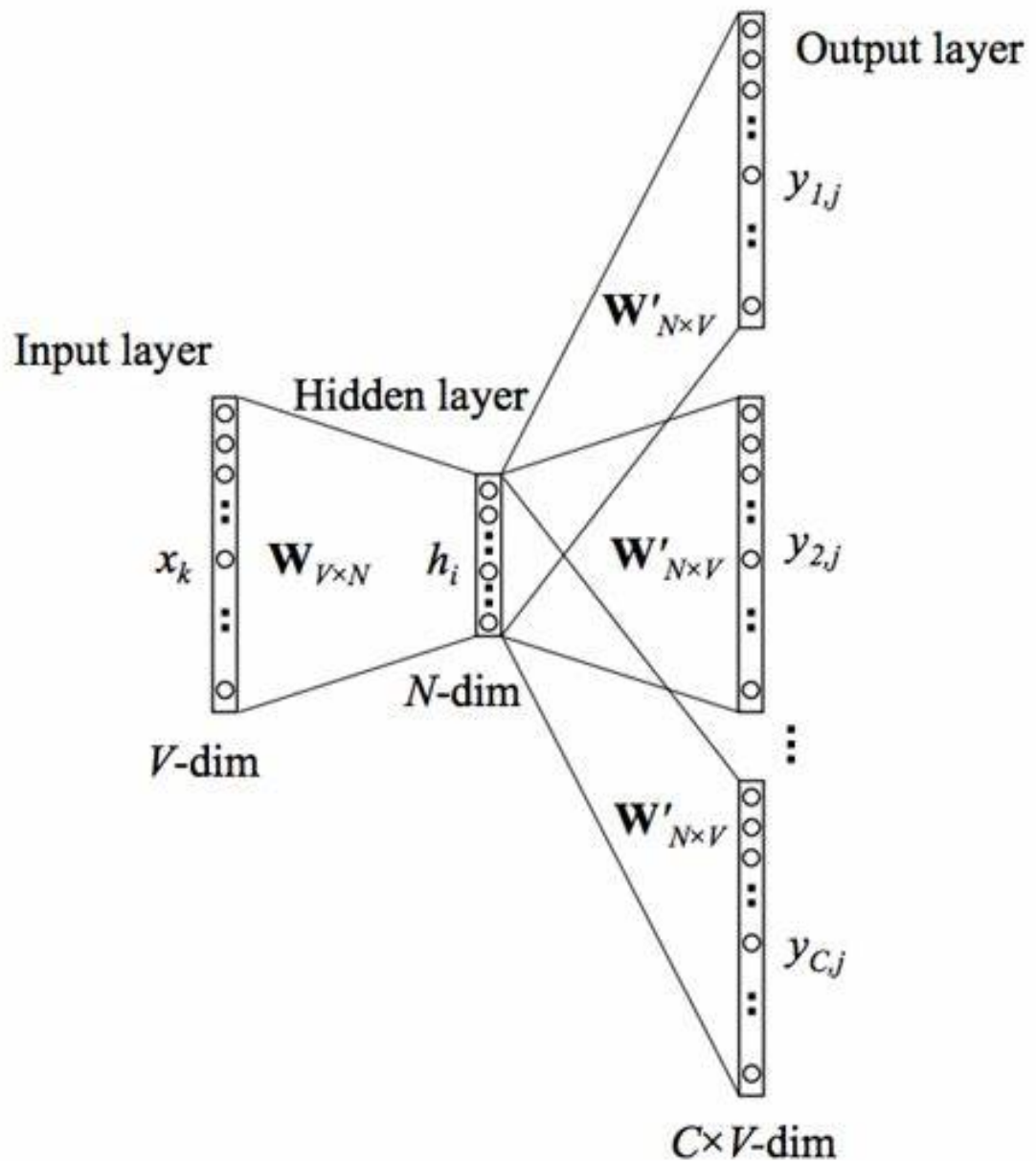


Figure 4.3: Skip-Gram-Architecture

[114]

Neural Network Architecture in Skip-Gram

- Input Layer: Unlike CBOW, the input layer in the Skip-Gram model corresponds to the target word. Each word in the vocabulary is represented as a vector in this layer.

- Hidden Layer: The hidden layer serves as a crucial intermediary stage where the vector representation of the target word is projected onto a set of weights. This projection aims to capture the diverse contextual relationships associated with the target word.
- Output Layer: In Skip-Gram, the output layer is designed to predict the context words based on the transformed information from the hidden layer. Each node in the output layer represents a potential context word. The objective is to maximize the probability of predicting the correct context words given the target word.

The training process of Skip-Gram involves adjusting the weights in a way that the model becomes proficient at predicting context words accurately when given a specific target word. Similar to CBOW, Skip-Gram achieves this by maximizing the probability of selecting the correct context words for a given target word.

During training iterations, the model refines the weights in both the hidden and output layers to optimize its predictive capabilities. The iterative weight adjustments are guided by the errors in predictions, leading the model to learn meaningful representations of contextual relationships associated with each target word.

The key distinction between Skip-Gram and CBOW lies in their focus during training. While CBOW predicts the target word based on the context, Skip-Gram predicts context words based on a target word. This fundamental difference makes Skip-Gram particularly well-suited for scenarios where a more detailed understanding of the contextual diversity of a word is desired. The Skip-Gram model provides a powerful mechanism for capturing the nuanced ways in which a word can be used in various contexts. Its ability to represent the intricate relationships between a target word and its surrounding context makes Skip-Gram a valuable tool in natural

language processing tasks that require a profound understanding of the multifaceted usage of the words.

4.3 Architecture of Word Embedding Based Question Answering Model

Here we describe the methodologies and architecture of the proposed semantic Malayalam Question-Answering system. The different components of the system include pre-processing of query and document, feature vector creating and semantic modeling, and document ranking and answer retrieval modules [75]. The block diagram of the system is shown in Figure 4.1.

4.3.1 Pre-processing of Queries and Documents

The Question Answering process initiates with the conversion of the provided query into a standardized form, achieved through a comparison with the structural patterns of stored queries. Subsequently, both the query and the associated documents undergo a series of pre-processing steps, employing various techniques such as tokenization, stop-word removal, and lemmatization.

Tokenization is the initial step, where user queries are disassembled into individual tokens. This process, specifically employing word tokenization, facilitates a granular understanding of the textual content. Following tokenization, the dimension of the term space is reduced through the removal of stop words. A customized Malayalam stop-word list is crafted for this work, comprising less important terms in the context of the Question Answering task.

Finally, Lemmatization is applied to ascertain the "lemma" of each word, considering its intended meaning and context within a document. Unlike stemming, lemmatization provides insights into the Part Of Speech, the exact meaning, and the overall context of a word in the document. The entire pre-processing workflow is seamlessly executed using the Indian Language Toolkit (INLTK), ensuring the effectiveness and accuracy of these textual transformations. The block diagram of the model is shown in Figure 4.4

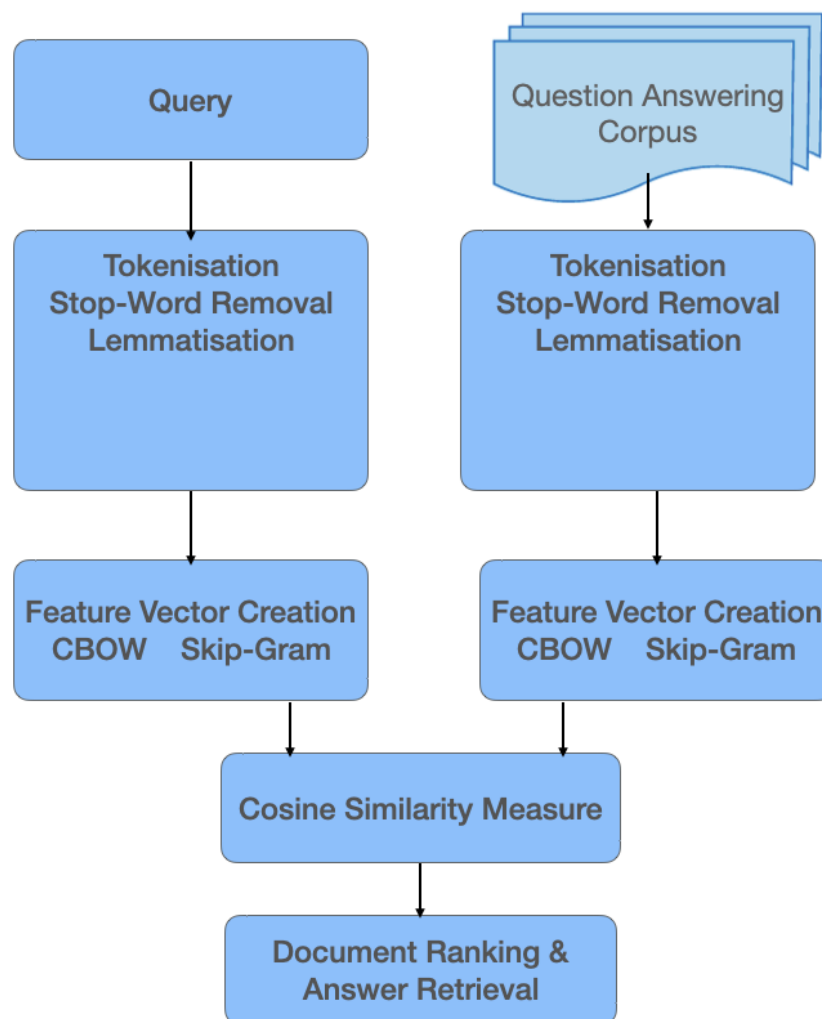


Figure 4.4: Block Diagram of Word Embedding Based Question Answering System

4.3.2 Feature Vector Creation and Semantic Modeling

After pre-processing, model the unlabelled word from the corpora into a dense vector space. Here we proposed Word Embedding to model the words to an m -dimensional vector space. While modeling the semantic information and context of the words will also be considered. In Word Embedding, model a word by mapping it from a multidimensional space to a continuous vector space with significantly less dimension per word.

Word2Vec is the algorithm suggested for this investigation. Word2vec uses a three-layered neural network to train itself to construct vocabulary from the corpus and learn word representations. Word2vec offers two models: 1. Skip-gram and 2. Continuous Bag of Words (CBOW). Skip-gram learns representations by predicting each context word based on the target word, whereas CBOW learns representations by predicting the target word based on its context words [68].

4.3.3 Document Ranking and Answer Retrieval

The process of document ranking depends on similarity measures to effectively capture the likeness between queries and documents. For answer retrieval, cosine similarity is used in this context. Cosine similarity is a metric used to measure the similarity between two vectors in a multidimensional space. It calculates the cosine of the angle between the two vectors, indicating the extent to which the vectors point in the same direction [74]. In the context of natural language processing and information retrieval, cosine similarity is often used to assess the similarity between two documents or the similarity of word vectors. The calculation of cosine similarity between questions and documents is given by:

$$\text{Cosine Similarity}(Q, D) = \frac{Q \cdot D}{\|Q\| \cdot \|D\|}$$

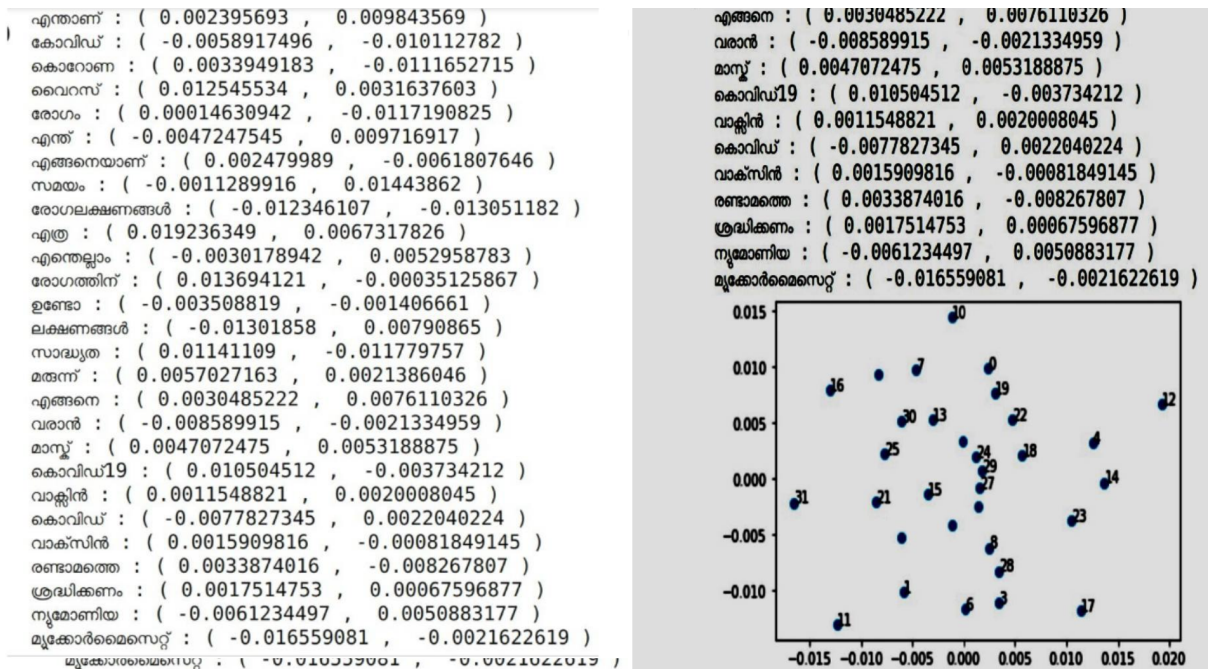


Figure 4.5: Word Embedding for Queries and Documents

A higher Cosine value indicates a high similarity between the query and documents. In response to the user’s query, the system will return the page with a high cosine similarity value.

The implementation of this study utilized a custom-created and annotated SQUAD-style data set MQuAD for health-related Question Answering. The data set annotation was conducted using the Haystack annotation tool. Subsequently, various pre-processing techniques, including tokenization, stop word removal, and stemming were applied to the annotated data. The next step involved transforming the documents into vectors using Word Embedding algorithms. In this study, both Continuous Bag Of Word (CBOW) and Skip Gram models were used for Word Embeddings. The sample Word Embedding for queries and documents is shown in Figure 4.5. The Word Embedding for entire documents is shown in Figure 4.6. The final stage includes matching and retrieving results using cosine similarity measures.

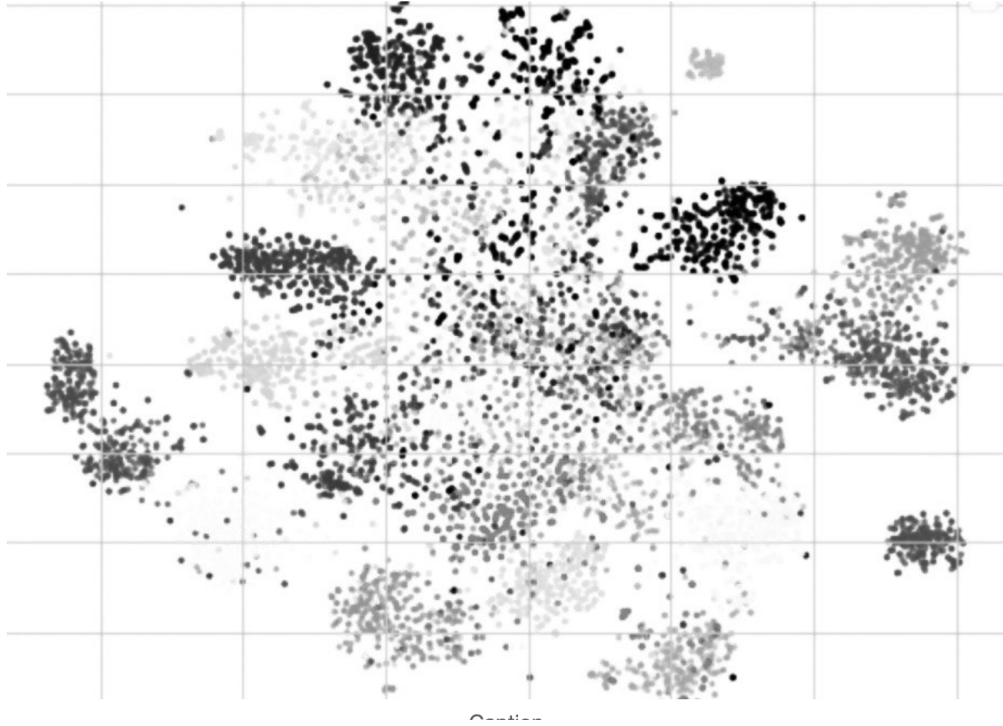


Figure 4.6: Word Embedding for MQAD

4.4 Algorithm for Word Embedding Based Question Answering

The proposed Question Answering (QA) model algorithm aims to efficiently retrieve relevant information from a corpus of documents in response to a user's query. The algorithm combines various natural language processing techniques and vector space modeling to achieve this goal. The different steps in the algorithms are explained below.

The initial step includes the pre-processing of both the user's query and the corpus of documents. This step ensures that the text data is transformed into a standardized and analyzable format. Then by using the Continuous Bag of Words (CBOW) and Skip-Gram algorithms, the documents are encoded into an M-dimensional vector space. This encoding captures semantic relationships between words, allowing

the algorithm to understand the contextual meaning of the documents. Subsequently, encode the documents into an M-dimensional vector space by Continuous Bag of Words (CBOW) and Skip-Gram algorithms. By ranking the documents based on their cosine similarity values, the algorithm identifies the documents most closely aligned with the user's inquiry. Then select the document with the highest Cosine Similarity value, effectively pinpointing the most relevant response to the user's question. The step-by-step algorithm for the proposed Question Answering model is shown below

Step 1: Start

Step 2: Pre-processing of query and documents.

Step 3: Convert and model the documents into an M-dimensional vector space using the CBOW and Skip-Gram algorithms.

Step 4: Find the similarity between the query and documents by using the Cosine Similarity measure.

Step 5: Rank the documents based on their Cosine Similarity value.

Step 6: Return the document with the highest Cosine Similarity value.

Step 7: Stop.

4.5 Result and Discussion

The sample queries along with their corresponding retrieved results are presented in Figure 4.7.

<p>Query1: എന്താണ് ഡെങ്കിപ്പനി.? എങ്ങനെ തടയാം? (What is dengue fever? How to prevent ?)</p>
<p>Answer : മൂന്നുദിവസം മുതൽ 15 ദിവസം വരെയാണ് ഡെങ്കിപ്പനി നീണ്ടുനിൽക്കുന്നത്. തലവേദന, പനി, കടുത്ത ക്ഷീണം, സന്ധികളിലും പേശികളിലും വേദന തുടങ്ങിയവയാണ് സാധാരണ ലക്ഷണങ്ങൾ. പനി പെട്ടെന്ന് 104 ഡിഗ്രി വരെ ഉയരുന്നതായും കാണപ്പെടുന്നു. ഹൃദയമിടിപ്പ് സാവധാനത്തിലാകുന്നതും രക്തസമ്മർദ്ദം കുറയുന്നതും രോഗത്തിന്റെ മറ്റു ലക്ഷണങ്ങളാണ്. സാധാരണ ഡെങ്കിപ്പനി അത്ര അപകടകാരിയല്ല. ഒന്നോ രണ്ടോ ആഴ്ച നീണ്ടുനിൽക്കുന്ന പനി കടുത്തക്ഷീണം അവശേഷിപ്പിച്ച് പിന്മാറും. പൂർണ്ണ ആരോഗ്യം വീണ്ടെടുക്കാൻ പിന്നെയും ആഴ്ചകളെടുക്കും. തൊണ്ടവേദന, ചുമ, മനഃപിരട്ടൽ, ചർദ്ദി, അടിവയറ്റിൽ വേദന എന്നിവയാണ് പ്രധാന ലക്ഷണങ്ങൾ. രോഗലക്ഷണങ്ങൾ തുടങ്ങി ഒരാഴ്ചയ്ക്കകം രോഗം മുർച്ചിക്കും. ഇതോടെ രോഗി തളർന്നുപോകും. നാഡിമിടിപ്പ് ദുർബലമാവുന്നതും വായയ്ക്കുചുറ്റും കരുവാളിപ്പുണ്ടാകുന്നതും ലക്ഷണങ്ങളിൽ പെടുന്നു. രോഗം ഗുരുതരമായാൽ രക്തസ്രാവമുണ്ടാകും. വായ, മുക്ക് എന്നിവയിലൂടെയും മലത്തിലൂടെയും രക്തം വരാം. ത്വക്കിലും രക്തസ്രാവലക്ഷണങ്ങൾ കണ്ടേക്കാം. അടിയന്തര വൈദ്യസഹായം കിട്ടിയില്ലെങ്കിൽ രോഗിയുടെ ജീവൻ അപകടത്തിലായേക്കാം. കൊതുക് പെരുകുന്നത് തടയുകയും വലകളുപയോഗിച്ച് കൊതുക് കടിക്കാതെ നോക്കുകയുമാണ് ഡെങ്കിപ്പനി വരാതിരിക്കാൻ ചെയ്യേണ്ടത്. രോഗിയും കൊതുകുവല ഉപയോഗിക്കാൻ ശ്രദ്ധിച്ചെങ്കിൽ മാത്രമേ രോഗപ്പകർച്ച തടയാൻ കഴിയൂ. വെള്ളം കെട്ടിക്കിടക്കുന്ന മഴക്കാലത്താണ് കൊതുകുകളുടെ പെരുകവും അതുവഴി ഡെങ്കിപ്പനിപോലുള്ള പനികളും പടരുന്നത്. മുൻകരുതലാണ് ഇവിടെ ഒരേയൊരു രക്ഷാമാർഗം. (Dengue fever lasts from three days to 15 days. Common symptoms include headache, fever, extreme tiredness, and pain in the joints and muscles. The fever also rises sharply to 104 degrees. Other symptoms include slowing of the heartbeat and lowering of blood pressure. Common dengue fever is not as dangerous. The fever lasts for one or two weeks and then recedes, leaving severe fatigue. It will take another few weeks to fully recover. The main symptoms are sore throat, cough, nausea, vomiting and abdominal pain. The disease worsens)</p>
<p>Query2: അൽസഹൈമേഴ്സ് രോഗം ചികിത്സിച്ചു ഭേദമാക്കാൻ പറ്റുമോ?(Can Alzheimer's disease be treated ?)</p>
<p>Answer: അൽസഹൈമേഴ്സ് രോഗത്തെപ്പറ്റിയും ആഹാരക്രമത്തെപ്പറ്റിയുമുള്ള ബന്ധത്തെക്കുറിച്ച് പഠിച്ചപ്പോൾ മനസ്സിലാക്കിയ കാര്യം അൽസഹൈമേഴ്സ് രോഗികൾക്ക് ഹൃദയസംബന്ധമായ രോഗങ്ങളും പക്ഷാഘാതവും വരുവാനുള്ള സാധ്യത കൂടുതലാണ് എന്നതാണ്. സമീകൃതമായ ഭക്ഷണരീതി സ്വീകരിക്കുന്നത് രോഗത്തെ ചെറുക്കുവാനുള്ള മാർഗമാണ്. അൽഷിമേഴ്സ് രോഗത്തെ ചുർന്നമായി ചികിത്സിച്ചു ഭേദമാക്കുക അസാധ്യമായതിനാൽ രോഗം വരാതെ നോക്കുവാനാണ് ശ്രദ്ധിക്കേണ്ടത്. ഹൃദയസംബന്ധമായ രോഗമുള്ളവരുടെ രക്തത്തിൽ കാണുന്ന തകരാറുകൾ എല്ലാതന്നെ അൽസഹൈമേഴ്സ് രോഗികളിലും കാണാം. സമീകൃത ആഹാരവും ദിവസേനയുള്ള നടത്തം പോലെയുള്ള വ്യായാമവും അൽസഹൈമേഴ്സ് രോഗം വരുവാനുള്ള സാധ്യത കുറയ്ക്കും. ജീവകം സിംഗും ജീവകം ഇയും അടങ്ങിയ ആഹാരങ്ങളും ചിലരിൽ അൽസഹൈമേഴ്സ് രോഗം വരുവാൻ ഉള്ള സാധ്യത കുറയ്ക്കും. (A study of the relationship between Alzheimer's disease and diet found that Alzheimer's patients were more likely to develop heart disease and stroke. Adopting a balanced diet is one of the ways to fight disease. Since it is impossible to completely cure Alzheimer's disease, care must be taken to prevent the disease from developing. Alzheimer's patients have all the same abnormalities in the blood of people with heart disease. A balanced diet and exercise such as daily walking can reduce the risk of developing Alzheimer's disease. Foods rich in vitamin C and vitamin E can reduce the risk of developing Alzheimer's in some people.)</p>

Figure 4.7: Sample Queries and Retrieved Results

This representation offers a comprehensive view of how the system responds to specific queries, showcasing the effectiveness and relevance of the retrieval process.

The proposed algorithm stands as a promising tool for enhancing the Question Answering System, showcasing the synergy of pre-processing, vector space modeling, and cosine similarity measures in the pursuit of accurate and efficient Question Answering.

4.6 Case study: COVID-19 Question Answering System Using Word Embedding

As a case study, implemented a COVID-19-related Question Answering system using word2vec models CBOW and Skip-Gram with Cosine Similarity measures. For this experiment created a separate Question Answering data set with COVID-19 information. The dataset was designed to cover a wide range of topics related to the pandemic, ensuring the system's ability to address diverse queries. The utilization of word2vec models, namely CBOW and Skip-Gram, was a strategic choice to capture the semantics of the COVID-19-related queries and responses. The sample queries and corresponding retrieved results are shown in Figure 4.8.

NO	Question	Answer
1	കൊറോണ വൈറസ് രോഗലക്ഷണങ്ങൾ എന്തെല്ലാം ? [What are the symptoms of coronavirus?]	പെട്ടെന്നുണ്ടാകുന്ന പനിയും ചുമയും അല്പലക്ഷിത പനി, ചുമ, പൊതുക്ഷീണം, തലവേദന, പേശിവേദന, തൊണ്ടവേദന, നാസാഗഹ്വരത്തിലെ വിക്കം, ശ്വാസംമുട്ടൽ, ഛർദ്ദി, വയറിളക്കം, മാനസികാവസ്ഥയ്ക്കുള്ള മാറ്റങ്ങൾ, എന്നിവയിൽ ഏതെങ്കിലും മൂന്നിലധികം ഉൾപ്പെട്ടിരിക്കുന്നു.[Sudden onset of fever and cough or fever, cough, general fatigue, headache, muscle aches, sore throat, nasal congestion, shortness of breath, vomiting, diarrhoea, and mood swings, includes any of these three symptoms]
2	കൊറോണ വൈറസ് രോഗം പകരുന്നത് തടയാനുള്ള മാർഗ്ഗങ്ങൾ എന്തെല്ലാം ? [What are the ways to prevent the spread of coronavirus disease?]	വ്യക്തിശുചിത്വം പാലിക്കുക, രോഗബാധിതരിൽ നിന്ന് അകലം പാലിക്കുക, ഹസ്തദാനം ഒഴിവാക്കുക, കൈകൾ ഇടയ്ക്കിടെ സോപ്പ് ഉപയോഗിച്ച് 20 സെക്കന്റോളം നന്നായി കഴുകുക, ആൾക്കൂട്ടം ഒഴിവാക്കുക എന്നിവ രോഗപുരൂർച്ച തടയാൻ ശുപാർശ ചെയ്യുന്നു. ചുമയ്ക്കുമ്പോൾ മൂക്കും വായയും മൂടുന്നതിലൂടെ രോഗാണുവ്യാപനം കുറയ്ക്കുക തടയാം. [It is recommended to maintain personal hygiene, avoid contact with infected people, avoid shaking hands, wash hands thoroughly with soap for 20 seconds frequently, and avoid crowding. Covering the nose and mouth when coughing can help prevent the spread of the disease.]
3	പ്രായമായവർക്ക് കൊറോണ വൈറസ് രോഗം വരാൻ സാധ്യത കൂടുതലാണോ ? [Are older people more likely to get coronavirus infection ?]	50 ഉം അതിൽക്കൂടുതലും പ്രായമായവർക്ക് മറ്റുള്ളവരെ അപേക്ഷിച്ച് അങ്ങേയറ്റം രോഗാത്മകമായ അവസ്ഥയിലെത്തിച്ചേരാൻ രണ്ടരയിരട്ടി സാധ്യതയുണ്ട്.[People of age 50 and older are two and a half times more likely to become critically ill than others.]
4	കൊവിഡ്-19 ന് കാരണമാകുന്ന വൈറസ് വായുവിലൂടെ പകരുന്നോ? [Can the virus that causes covid-19 be spread through the air?]	പഠനങ്ങൾ സൂചിപ്പിക്കുന്നത് കൊവിഡ്-19 ന് കാരണമാകുന്ന വൈറസ് പ്രധാനമായും പകരുന്ന വായുവിലൂടെയല്ലാതെ ശ്വാസന തുള്ളികളുമായുള്ള സമ്പർക്കത്തിലൂടെയാണ് എന്നാണ്. [Studies indicate that the virus that causes covid-19 is mainly transmitted through contact with respiratory droplets rather than air.]

Figure 4.8: Sample Queries and Retrieved Results

The system will answer users' queries related to COVID-19 disease. This work demonstrated its potential utility in addressing real-world challenges posed by the COVID-19 pandemic

4.7 Performance Evaluation and Analysis

The evaluation of the system is carried out with the use of the F1 score, a measure that creates a single value by combining recall and precision. Both precision and recall are pivotal aspects for assessing the performance of the system, and the F1 score provides a comprehensive measure that strikes a balance between these two important indicators. The F1 score is calculated using the formula 4.2.

$$F1 = 2 \times (PrecisionRecall)/(Precision + Recall) \quad (4.1)$$

Precision and recall are commonly used metrics to evaluate the performance of question-answering systems. These metrics are particularly important in natural language processing tasks where the system is required to provide accurate and relevant answers. Let's define precision and recall in the context of a question-answering system

- Precision:- A measurement of precision indicates how well the system has predicted good outcomes. It answers the question: "Of all the instances that the system predicted as positive, how many are positive?" Precision is calculated using the formula

$$Precision = (TruePositives)/(TruePositives + FalsePositives) \quad (4.2)$$

Where True Positives (TP) are the instances where the system correctly predicted positive, and False Positives (FP) are the instances where the system

predicted positive incorrectly.

- Recall:- Recall, also known as sensitivity or true positive rate, measures the ability of the system to capture all the relevant instances. It answers the question: "Of all the actual positive instances, how many did the system correctly identify?" Recall is calculated using the formula:

$$Recall = (TruePositives)/(TruePositives + FalseNegatives) \quad (4.3)$$

True Positives (TP) are the instances where the system correctly predicted positive, and False Negatives (FN) are the instances where the system failed to predict positive when it should have.

In the context of a Question Answering system, True Positives (TP) are instances where the system correctly answers a question.

False Positives (FP): Instances where the system incorrectly answered a question.

False Negatives (FN): Instances where the system failed to provide an answer when it should have.

A good Question Answering system should have both high precision and high recall parameters. The Performance Analysis and features of our proposed Question Answering System are shown in Table 4.1 and Table 4.2. The average F1 score of

Table 4.1: Performance Analysis of the System

Models	Cosine Similarity	F1 Score	Learning Rate
CBOW	0.78	72.41%	0.025
Skip-Gram	0.81	76.20%	0.022

the proposed system with CBOW and Skip-Gram methods is evaluated as 72.4% and 76.2% respectively.

Table 4.2: Features of the Model

Model	Embedding Dimension	Window Size	Number of Epochs
CBOW	100	4	10
Skip-Gram	100	4	10

4.8 Conclusion

In this stage, A Word Embedding-based Malayalam Question Answering model is proposed. It will benefit the indigenous people who are looking for information about health-related problems. Annotated Malayalam Question Answering dataset MQuAD is utilized for this experiment. Semantic mapping and modeling of the document are achieved by Word Embedding algorithms. The relevant responses to the questions are then mapped and retrieved using similarity measures.

Chapter 5

BERT Based Models for Question Answering

5.1 Introduction

Conversational Question Answering aims to retrieve relevant answers to user's queries in natural languages. Question Answering and Information Retrieval have been a significant advancement with intelligent systems such as Google Assistant, Alexa, and Siri [16]. People are changing their information-searching activities to text or voice-based interactive conversational methods. The evolution of neural networks and transformer-based models has had an eminent impact on Question Answering to ingest huge volumes of data and make Natural Language Processing easy.

In conversational Question Answering, the response is either selected from a candidate set or extracted from a document [146]. To extract accurate and relevant responses, the context or meaning of the text must be considered. The semantic modeling or mapping of the documents as vectors plays an important role in the

aspects of Question Answering. To address this challenge, this model proposed a BERT-based conversational Question Answering model by performing a comparative analysis of different BERT language models for Question Answering [104] on the health domain in the Malayalam language and integrating the Reader-Retriever pipeline with BERT models for better performance.

5.2 Bidirectional Encoder Representation from Transformers (BERT) Model

BERT is the most popular and broadly used NLP and transformer-based model. It can be considered the full context of the texts by looking in both directions, which is suitable for identifying the intent of a query asked in Question-Answering tasks. Because of its bi-directionally [24] and attention mechanism [147] attention BERT has a deeper sense of language context and flow [120], and hence it is widely used now in many NLP applications like Question Answering and Information Extraction.

5.2.1 Model Architecture

BERT is a transformer-based model, which utilises the encoder part of the transformers [56]. It is used for getting contextual information from the past and future. The transformer architecture consists of stacked encoders and decoders connected with self-attention [24] as shown in Figure 5.1.

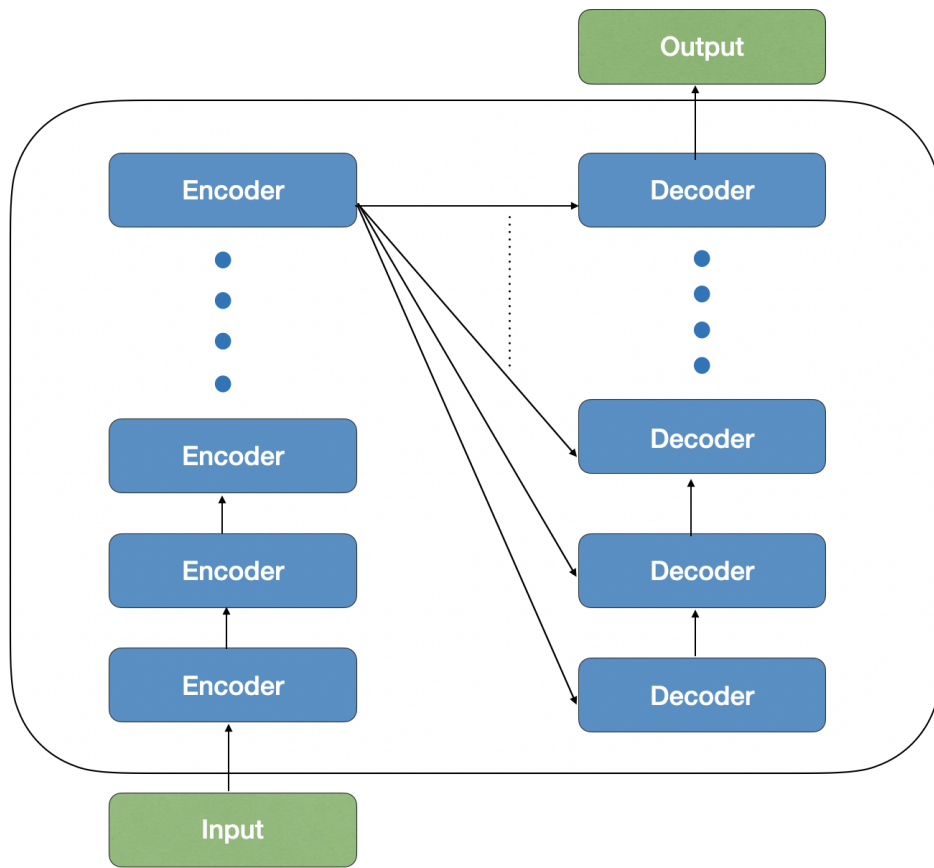


Figure 5.1: Encoder Decoder Stacks in Transformers

Depending on the model and the work at hand, the number of stacked encoder and decoder blocks may vary. Each decoder receives the output from the top-ranking encoder. The output from the final decode is mapped to the intended target word sequence. The detailed architecture of the transformer model is shown in Figure 5.2.

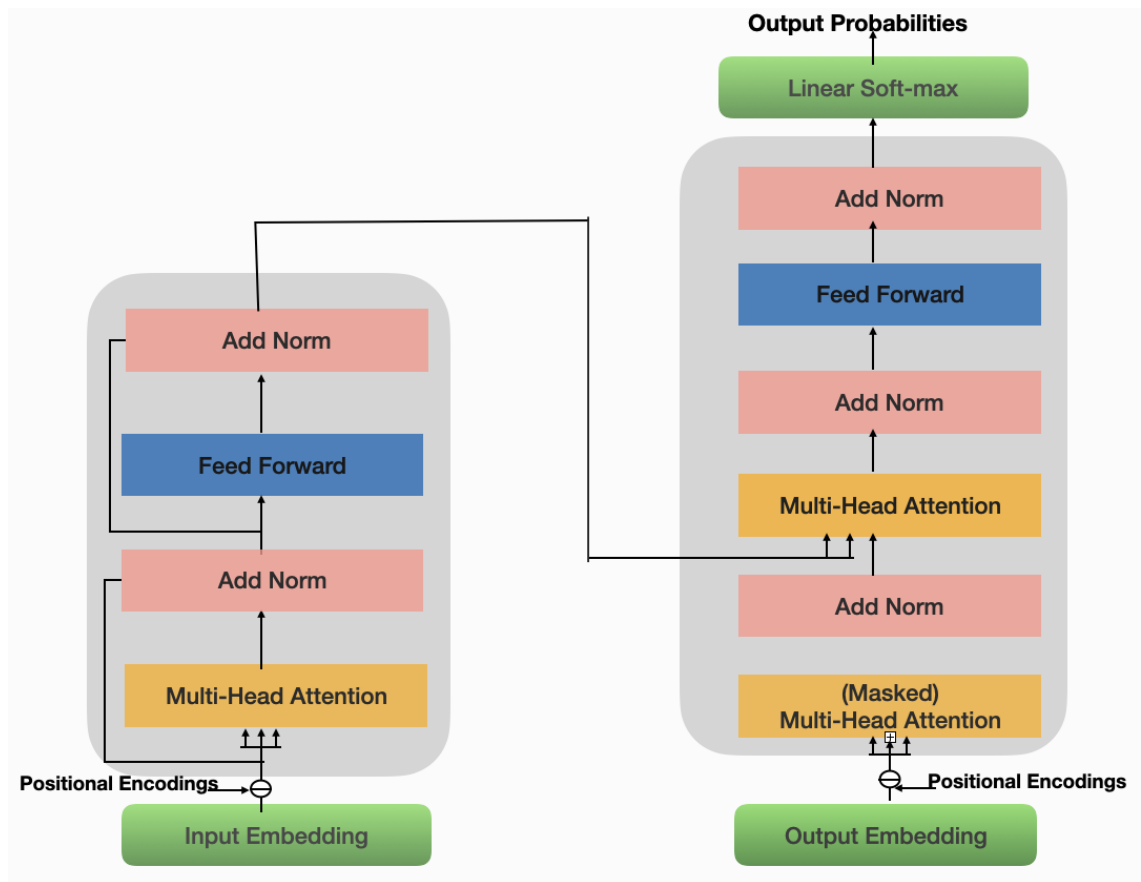


Figure 5.2: Transformer Model Architecture

For all target words, generate scores for each of the encoder stacks. Then a soft-max is used to normalize all the scores. Final weights are introduced to make the context vector easy to train. Then an attention vector can be computed.

5.2.2 Self Attention

Self-attention is a pivotal mechanism in BERT and other transformer-based models. In the context of BERT, self-attention, also known as scaled dot-product attention, is a mechanism that allows the model to weigh the importance of different words in a sequence when encoding information

The attention mechanism is a context vector that is plugged into the encoder-decoder architecture. This context vector computes the probability distribution of input words by taking all the encoder's output. For each target word, generate scores for each of the encoder states [20]. Then a softmax is used to normalise all the scores. Finally, the weights are introduced to make the context vector easy to train. Then the attention vector can be computed from the context vector. There are two types of attention, they are

- Scaled dot product attention

The dot product attention can be computed by obtaining dot products between queries and keys and then applying a softmax function to get weights for the corresponding values [165]. The attention function can be computed as shown in equation 5.1.

$$Attention(Q, K, V) = Softmax \times (QK^T / \sqrt{d_k})V \quad (5.1)$$

The set of queries is packed into a matrix Q . The keys and values are packed into matrix K and V .

- Multi-head Attention

Multi-head attention employs multiple attention heads to concurrently capture various aspects of the input sequence to understand and represent complex relationships within the data [21]. Perform d -dimensional keys, values, and queries to linearly project the queries, keys, and values 'h' times with different linear projections to d_k, d_k , and d_v dimensions. respectively.

The attention function can be calculated as shown in equation 5.2.

$$Multi - head(Q, K, V) = Concat(head_i, \dots, head_h)W^0 \quad (5.2)$$

Where $head_i = Attention(QW_i^Q, KW_i^k, VW_i^v)$

BERT is a bidirectional transformer for language modeling and is more popular in many NLP applications like Question Answering, language translation, etc. The bidirectional processing [71] gives them deeper insights into the context of the language. BERT used the features of Masked Language Modeling (MLM) [70] and Next Sentence Prediction (NSP) [132] to surpass unidirectional constraints. The two stages are to be followed while using BERT and re-training the model with unlabeled data and fine-tuning [138] the model for a specific task such as Question Answering.

By employing a bidirectional approach and sequential reading of the input text, BERT enables the model to infer the meaning or context of a word from its neighboring terms. The encoder receives the token sequence, which is embedded as vectors [112]. Position embedding, segment embedding, and token embedding [84] are combined by the BERT model. Followed by a neural network that receives the vectors and, depending on the document's context generates an output sequence of vectors. [4] [136].

5.2.3 Different BERT Models

There are different variations of BERT models available. They are ALBERT, DistilBERT, StructBERT, RoBERTa, etc. [64] [2]. Among these, ALBERT is a much smaller version of BERT with less computation power [67]. DistilBERT introduced for knowledge distillation, is used for the computation of large numbers of parameters [123]. StructBERT incorporates language structures like word-level or-

dering and sentence-level ordering into the BERT pre-training phase[152]. With additional pre-training enhancements, RoBERTa is an optimized version of the BERT model [103] [78], that achieves state-of-the-art performance on different NLP applications like Question Answering [163]. Each BERT model has two versions: BERT base and BERT large. These different BERT models may have varying features depending on their architecture, training data, and parameters as shown in Table 5.1.

Table 5.1: Features of Different BERT Models

Models	No. of Layers	No.of Hidden Units	Parameters
BERT base	12	768	110M
BERT large	24	1024	340M
ALBERT base	12	768	12M
ALBERT large	24	1024	18M
RoBERTa base	12	768	125M
RoBERTa large	24	1024	355M
DistilBERT base	6	768	66M

5.3 Question Answering using Different BERT Models

This section discusses the implementation of a Question Answering model by using various BERT models. Re-training and fine-tuning of various BERT models for the Question Answering process are discussed.

The proposed model will retrieve answers to user’s queries about their health-related issues and information about wellness in the Malayalam language. The study utilizes the self-created Malayalam health Question Answering Data set MQuAD. The annotated MQuAD data sets are retrieved in JSON file format. MQuAD contains around 1 lakh questions and answers paired as in the Stanford University

Question Answering Data Set (SQuAD) format. Used different pre-trained BERT models and then re-trained the models with health Malayalam data then fine-tuned the models with MQuAD data set. The working of the Question Answering System with BERT architecture is illustrated in Figure 5.3.

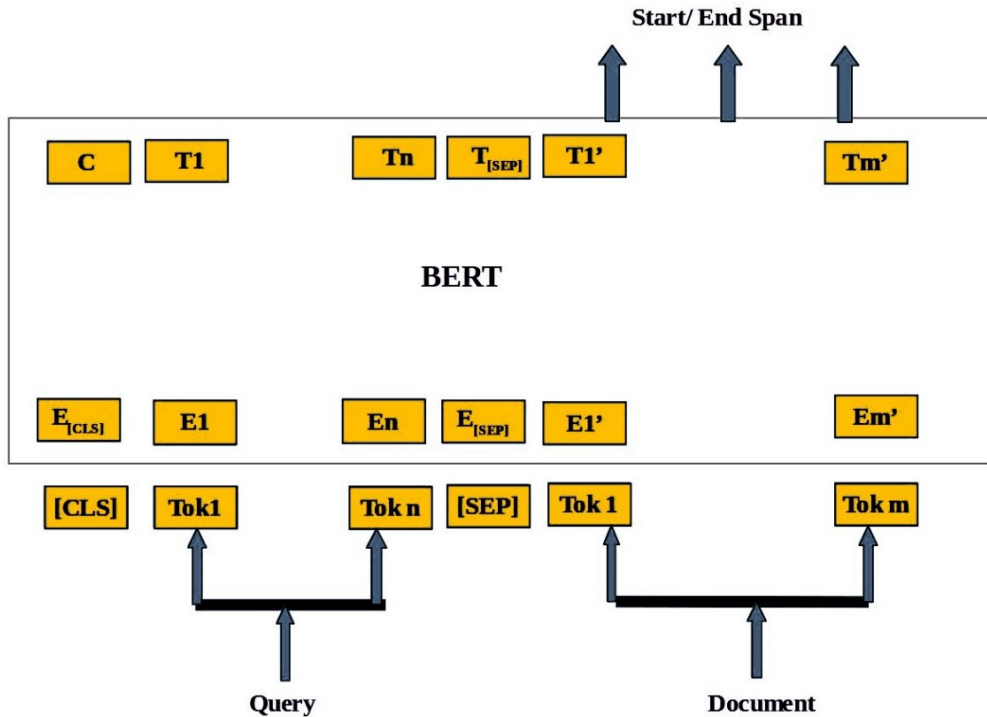


Figure 5.3: Working of Question Answering Process in BERT

BERT extracts tokens from both questions and documents, combining them as input into the model. It incorporates a $[CLS]$ token indicating the start of the sentence and a $[SEP]$ token to separate questions and documents containing answers. BERT generates two segment embeddings: one for the questions and another for the documents. These embeddings are then added together to create a one-hot representation to distinguish questions from documents. Subsequently, the combined embedded representation serves as input to the BERT model [121].

Finally, the softmax function is applied to the last hidden layer of the BERT model to generate a probability distribution for the start and end tokens corresponding to the input text sentence. This distribution is then output as the answer for the posed query. This study compared the performance of different BERT models, including ALBERT, DistilBERT, StructBERT, and RoBERTa.

5.4 BERT Models with Reader-Retriever Pipeline for Question Answering

This phase significantly improved the capabilities of the Question Answering process by integrating a sophisticated Reader-Retriever pipeline architecture [46] with BERT models. Reader-Retriever pipeline [10] is a technique, generally used for open-domain Question Answering [169]. This involved the practical implementation of the proposed BERT model as a Retriever-Reader pipeline, enhancing its efficiency and performance in handling complex queries. The block diagram of the customized BERT model is illustrated in Figure 5.3

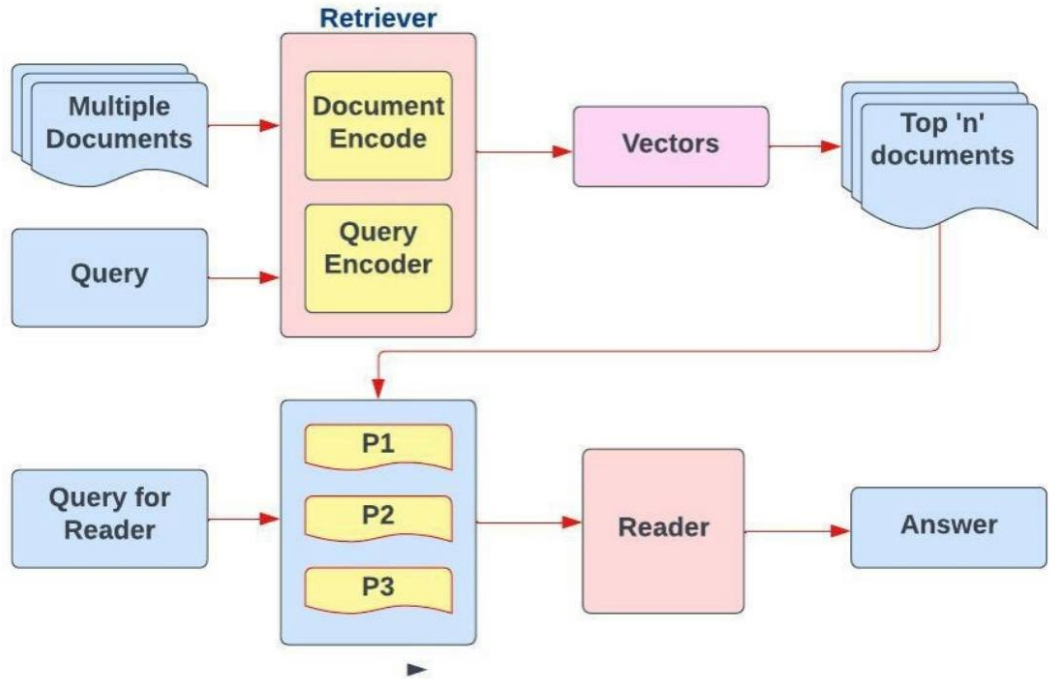


Figure 5.4: Block Diagram of Customized BERT Model

Before passing to the BERT models, pre-process the data set to fit the format for the pre-trained transformer models. Removed empty lines, white spaces, and punctuation. Then applied word tokenization with a maximum sentence split length of 200. After that assign each token to a corresponding token ID to make it a BERT readable format. These are then passed to the retriever module. The retriever contains a dual encoder architecture one for query and the other for the document. The retrievers will parse the query and document then compute a retriever score. The retriever score is defined by the dot product of the hidden representation of the query and document computed as follows.

The model functions as follows: initially, the retriever pinpoints the top n most likely documents. Out of these, the reader identifies and ranks the top m based on similarity scores, ultimately selecting the document with the highest similarity

score as the most appropriate answer. We assessed the performance of the new model against the old BERT model, including ALBERT, DistilBERT, StructBERT, and RoBERTa.

The query representation was obtained during the execution of the model as shown in equation 5.1.

$$V_q = W_q BERT_q(qkrt)[CLS] \quad (5.3)$$

Where $BERT_q$ is the question encoder

W_q is the projection matrix for query representation [CLS]

$(qkrt)$ is the re-formatted query for the retriever representation and V_q belongs to

R , is the final query representation

The same method is used to obtain the document representation for document, i .

$$V_d = W_d BERT_d(D_i)[CLS] \quad (5.4)$$

Where $BERT_d$ is the document encoder

W_d is the projection matrix for the document query representation [CLS]

D_i is a document collection and

$(V_d R)$ is the final document representation

Finally, compute the retriever score as,

$$Srt(qkrt, D_i) = V_q V_d T \quad (5.5)$$

The retriever will retrieve more probable documents for answering the queries, and then the reader reads and extracts the best answers for the query by calculating a similarity score.

Similar to a retriever, the reader constructs a reformatted query and then concatenates retrieved documents. It will then be input into the BERT model. The input sequences are as follows.

$$(qkrd, P_i)is[CLS]q_k - w[SEP].....[SEP]q_{(k-1)}[SEP]q_k[SEP]P_i \quad (5.6)$$

The BERT model in the reader generates contextual representation for each token.

$$V[n] = BERT((qKrd, P_i))[n] \quad (5.7)$$

Where $V[n]$ is the representation for the n th token.

$qkrd$ is the reformatted query for the reader.

The sequence representation $V[CLS]$ is obtained by

$$V[CLS] = W[CLS]BERT((qkrd, P_i))[CLS] \quad (5.8)$$

Where $W[CLS]$ is the projection for the $[CLS]$ representation.

The reader will rank the top n retrieved answers and then predict an answer span by calculating a similarity score as follows.

$$S_s(qkrd, P_i[n]) = W_s V[n] \quad (5.9)$$

Where $S_s(qkrd, P_i[n])$ are the scores for the n^{th} token as the start token

$$S_e(qkrd, P_i[n]) = W_e V[n] \quad (5.10)$$

Where $S_e(qkrd, P_i[n])$ are the scores for the end tokens of the answer span.

Finally, the reader will generate the most relevant answer to the user's query.

5.5 Result and Discussion

This section contains a detailed analysis and evaluation of various BERT models, namely ALBERT, DistilBERT, StructBERT, and RoBERTa. Initially performed re-training and fine-tuning of various BERT models for Question Answering process. Next level, enhanced the model by incorporating a hybrid Reader-Retriever pipeline architecture with each BERT model. The sample result taken from the experiment is shown below.

<p>Query:പ്രമേഹത്തിന് ആയുർവേദ ലഘുചികിത്സകൾ എന്തെല്ലാം?</p> <p>Answer:ആയുർവേദശാസ്ത്രപ്രകാരം പ്രമേഹരോഗം പ്രധാനമായും 20 വിധത്തിൽ കാണപ്പെടും. ഉദകമേഹം, ഇക്ഷുമേഹം, സുരാമേഹം, വിഷ്ടമേഹം, ശുക്ലമേഹം, ലാലാമേഹം, ശനൈർമേഹം, സികതാമേഹം, ശീതമേഹം, സാന്ദ്രമേഹം ഇങ്ങനെ കഫദോഷ പ്രധാനമായുണ്ടാകുന്ന മേഹങ്ങൾ പത്തുവിധത്തിലുണ്ട്. ഇതു കൂടാതെ മഞ്ജിഷ്ഠാമേഹം, നീലമേഹം, കാളമേഹം, ഹരിദ്രാമേഹം, ശോണിതമേഹം, ക്ഷാരമേഹം എന്നിങ്ങനെ പിത്തപ്രധാനങ്ങളായമേഹങ്ങൾ ആറും വസാമേഹം, മജ്ജാമേഹം, ഹസ്തിമേഹം, മധുമേഹം എന്നിങ്ങനെ വാതപ്രധാനമായ മേഹങ്ങൾ നാലു വിധത്തിലും കണ്ടുവരുന്നു. അതിനാൽത്തന്നെ ശാസ്ത്രീയമായ ആയുർവേദചികിത്സയ്ക്ക് വിദഗ്ധമായ രോഗനിർണ്ണയം അത്യാവശ്യമാണ്. മേൽപറഞ്ഞ വിഭാഗങ്ങളിൽ ചികിത്സ അസാധ്യമായി സംഹിതകൾ കരുതുന്ന മധുമേഹത്തോടാണു നാം ആധുനിക ശാസ്ത്രത്തിൽ പറയുന്ന പ്രമേഹം എന്ന രോഗത്തിന് ഏറെ സാമ്യം. അതിനാൽത്തന്നെ പ്രമേഹചികിത്സ ഏറെ സൂക്ഷിച്ചു ചെയ്യേണ്ടതും കൃത്യമായ വൈദ്യനിർദ്ദേശപ്രകാരം ദീർഘകാലം ശീലി കേണ്ടതുമാണ്. ബലവാനായ രോഗിക്കു വമനം, വിരോചനം തുടങ്ങിയ ശോധന പ്രയോഗങ്ങൾക്ക് ശേഷമാണ് ശമന ചികിത്സ വിധിക്കുന്നത്. എങ്കിലും വിട്ടിൽ പരീക്ഷിച്ചു നോക്കാവുന്ന ചില ലഘു പ്രയോഗങ്ങൾ ചുവടെ നൽകുന്നു. ലഘു പ്രയോഗങ്ങൾ പച്ചമഞ്ഞൾ, പച്ചനെല്ലിക്ക ഇവയുടെ നിർ 25 മില്ലി വിതം ആവശ്യത്തിനു തേനും ചേർത്ത് രാവിലെ വെറുവയറ്റിൽ സേവിക്കുക. നെല്ലിക്ക ഇടിച്ചു പിഴിഞ്ഞ ഒരൗൺസ് നിരിൽ രണ്ട് ടീസ്പൂൺ വരട്ടുമഞ്ഞളിന്റെ പൊടി ചേർത്തു സേവിക്കുന്നതും പച്ചനെല്ലിക്കയും മഞ്ഞളും തുല്യമായി ചേർത്തരച്ചു 20 ഗ്രാം വരെ രാവിലെ വെറും വയറ്റിൽ സേവിക്കുന്നതും ഗുണകരമാണ്. തേറ്റമ്പരൽ നാല്-അഞ്ച് എണ്ണം വെള്ളത്തിൽ ഒരു രാത്രി ഇട്ടുവെച്ചിരുന്നു രാവിലെ കടഞ്ഞെടുത്ത മോരിലരച്ച് സേവിക്കുന്നതു പ്രമേഹശമനീയമാണ്. ചിറ്റുമുതിന്റെ നിർ- 25 മില്ലി-തേൻ ചേർത്ത് രാവിലെ വെറുംവയറ്റിൽ കഴിക്കുന്നതു പ്രമേഹം ശമിപ്പിക്കും. ഏകനായകത്തിൻവേര് (പൊൻകുരണ്ടി, തേറ്റമ്പരൽ എന്നിവ തുല്യ അളവിൽ പൊടിച്ചു രണ്ടു ടേബിൾ സ്പൂൺ വിതം രണ്ടു നേരം സേവിക്കുന്നത് പ്രമേഹശമനത്തിന് ഉത്തമമാണ്. ഏകനായകവും പച്ചമഞ്ഞളും (20 ഗ്രാം) പുളിക്കാത്ത മോരിൽ തുല്യമായ അളവിലരച്ച് രണ്ട് നേരം കഴിച്ചാൽ പ്രമേഹത്തിന് ശമനമുണ്ടാവും. പുളിയരിത്തൊണ്ട്, നെല്ലിക്ക ഇവ കഷായം വച്ച് -50 മി ലി വിതം ഒരു ടീസ്പൂൺ ഞവരഅരിയുടെ തവിടു ചേർത്ത് രണ്ടു നേരം സേവിച്ചാൽ പ്രമേഹരോഗം തടയാം. മുരിക്കിന്റെ തൊലി അരച്ച് (20 ഗ്രാം)- മോരിലോ തേനിലോ ചേർത്ത് സേവിക്കുന്നതു ഹിതമാണ്. അഞ്ചു കുവളത്തില അരച്ചുരുട്ടി രാവിലെ വെറും വയറ്റിൽ കഴിക്കുന്നതു പ്രമേഹം നിയന്ത്രിക്കുന്നതിനു സഹായിക്കുന്നു. ത്രിഫല, മഞ്ഞൾ, ഞാവൽത്തൊലി, നാൽപാമരത്തൊലി, നിർമാതളത്തൊലി, ചെറുളവേര്, പാച്ചോറ്റിത്തൊലി ഇവ ഒന്നിച്ചോ അല്ലെങ്കിൽ ഇവയിൽ ലഭ്യമായ മൂന്നു മരുന്നുകൾ തുല്യ അളവിൽ ഏകനായകവുമായി ചേർത്തോ കഷായം വച്ചു കഴിച്ചാൽ പ്രമേഹം ശമിക്കും. കന്ദം പൊടിച്ചത് അഞ്ചുഗ്രാം വരെ സേവിക്കുന്നത് പ്രമേഹശമനത്തിന് നല്ലതാണ്. തേൻ ചേർക്കേണ്ട യോഗങ്ങളിൽ വിശ്വാസയോഗ്യമായ ചെറുതേൻ ആണ് ഉപയോഗിക്കേണ്ടത്.</p>

Figure 5.5: Sample Result from BERT Model

The performance analysis of different BERT models is shown in Table 5.2.

Table 5.2: Performance Analysis of Different BERT Models

Models	Precision	Recall	F1 Score	Speed(No of Doc/Sec)
ALBERT (With Reader Retriever)	0.45	0.7	74.02%	180
DistilBERT (With Reader Retriever)	0.6	0.41	69.28%	240
StructBERT (With Reader Retriever)	0.88	0.80	84.19%	210
RoBERTa (With Reader Retriever)	0.93	0.80	86.45%	150

The F1 score is calculated as a token level. F1 scores of different BERT models with and without Reader-Retriever pipeline are shown in Table 5.3

Table 5.3: F1 Score of Different BERT Models

Models	F1 Score
ALBERT	68.53%
DistilBERT	62.32%
StructBERT	76.42%
RoBERTa	78.65%
ALBERT (With Reader Retriever)	74.02%
DistilBERT (With Reader Retriever)	69.28%
StructBERT (With Reader Retriever)	84.19%
RoBERTa (With Reader Retriever)	86.45%

Experimental result shows that the F1 score of the RoBERTa model with Reader-Retriever pipeline architecture is higher than all other models. Figure 5.5 indicates the graph of the performance analysis of the system in terms of F1 score.

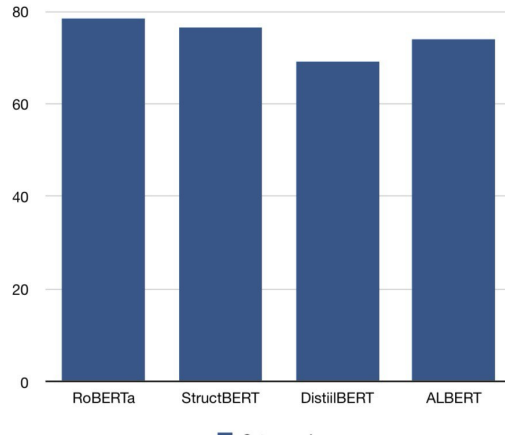


Figure 5.6: F1 Score of Different BERT Models

From the graphs, it is clear that the RoBERTa achieved the highest performance with an F1 score of 86% and DistilBERT achieved the highest processing speed. We gauged the performance metrics using the F1 score, Exact Match (EM), and processing speed measured in documents/second. The F1 score is employed to determine similarity to the actual answer and represents the harmonic mean of the Precision and Recall metrics. Precision is the proportion of relevant documents retrieved to the total retrieved, while Recall is the proportion of relevant documents retrieved out of all available relevant documents.

The RoBERTa with reader retriever pipeline achieved, the highest performance with an F1 score of 86% and DistilBERT achieved the lowest F1 score.

The below graph compares the performance of each BERT model before and after introducing the Reader-Retriever pipeline architecture.

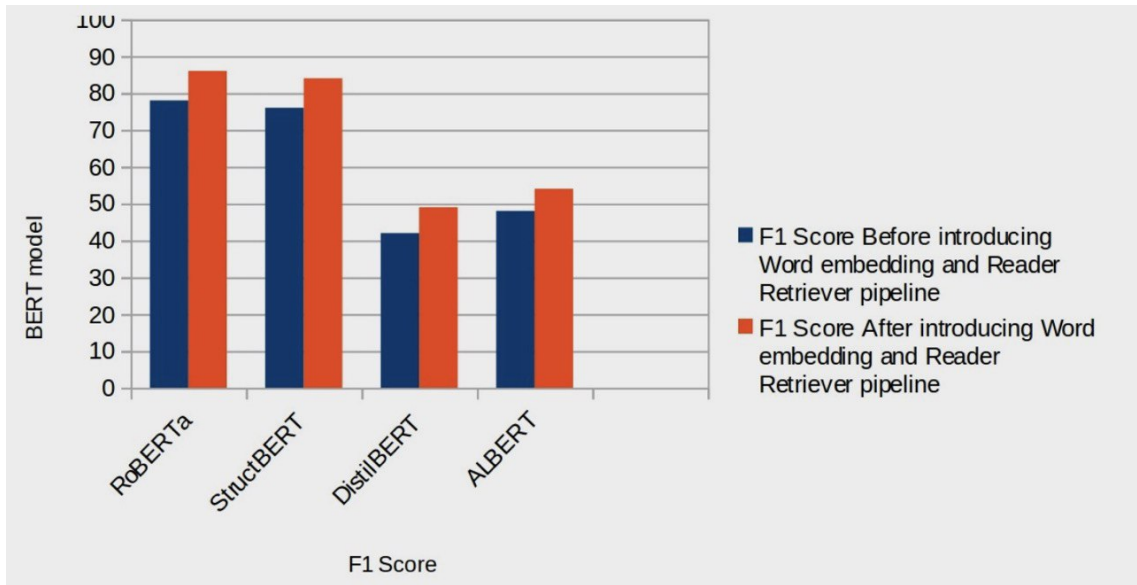


Figure 5.7: Performance Analysis of the System With and Without Reader-Retriever Pipeline

From the graph, it is clear that the F1 scores of all BERT models are increased after the integration of the reader retriever pipeline architecture. Furthermore, the optimized BERT model, RoBERTa achieved the highest F1 score compared to all other BERT models. DistilBERT on the other hand, exhibited faster speeds.

Table 5.4: System Specifications

Tesla P100 Pale 16GB GPU
Consumed CUDA memory
Tensor flow packages and features
Haystack annotation tool
Training epoch = 5
Number of training steps = 2776
Number of warmup steps = 555
Longest sequence length = 256
Training loss = 0.0814

The system is equipped with a Tesla P100 Pale 16GB GPU, which is utilized CUDA memory during the training process. TensorFlow packages and features are used for GPU acceleration for efficient model training, while the Haystack annotation tool is used to manage and annotate datasets. The model undergoes 5 training epochs, with 2776 training steps and 555 warmup steps to gradually adjust the optimizer.

The longest input sequence length is set to 256, and the training loss reaches 0.0814, indicating the model’s effective learning performance.

5.6 Conclusion

This study conducted a detailed analysis of different BERT models for Question Answering. We employed various pre-trained BERT models and then retrained and fine-tuned them using the health Malayalam Question Answering Data set MQuaD. After that introduced a Reader Retriever pipeline architecture with each BERT model. Experimental results indicate that the F1 score of all BERT models increased after the introduction of the Reader Retriever pipeline architecture and the optimized BERT model Roberta achieved the highest F1 score compared to other BERT models. While DistilBERT exhibited a higher processing speed than all other models. The Exact Match (EM) score for each BERT model was evaluated as 1.

Chapter 6

BioQA BERT- Proposed BERT Model for Question Answering Using LoRA and LLaMA

6.1 Introduction

This chapter proposed a new enhanced BERT model, BioQA-BERT, by integrating LoRA (Low-Rank Adaptation) [72] and the latest LLM - (LLaMA) [144] with optimized BERT model - RoBERTa. LoRA is a parameter-efficient fine-tuning technique currently used with Large Language Models, while LLaMA is the latest Large Language Model. Here LLaMA is used to improve the accuracy of the Question Answering model [156][12].

6.2 LoRa and LLaMA for Question Answering

LoRA serves as a parameter-efficient fine-tuning technique currently applied with Large Language Models and LLaMA represents the latest addition to the family of

Large Language Models. In this study, employed LLaMA to enhance the accuracy of the Question Answering process.

6.2.1 LoRA-Low-Rank Adaptation Algorithm

LoRA is a Low-Rank Adaptation technique used to fine-tune large language models, aiming to decrease the number of trainable parameters for downstream tasks without compromising the model's quality. Unlike other methods it demands full rank during adaptation in accumulated gradient updates to weight matrices, LoRA doesn't necessitate this condition. Consequently, it can converge to the original model using fewer trainable parameters. Furthermore, LoRA offers notable benefits, including a substantial reduction in storage requirements and task-switching overhead [140]. This will enhance the training efficiency and introduce no inference latency to the model when compared to a fully fine-tuned model.

6.2.2 LLaMA-Large Language Model Meta AI

LLaMA is Large Language Model Meta AI, a new AI language generator developed by Meta. LLaMA is an auto-regressive language model based on the transformer architecture, available in four sizes with parameters ranging from 7B to 65B [117]. The main purposes of LLaMA include exploring applications, evaluating existing language models, enhancing strategies [35], and managing risks associated with AI language models. Despite being smaller, LLaMA outperformed all other models in tests and evaluations, with potential applications like a chatbot, translation, Question Answering, Text Generation, Content Creation, and Data Analysis, etc.[48].

6.2.3 LoRA and LLaMA for Question Answering

Here LoRA is a low-rank adaptation used for fine-tuning the BERT model, aiming to decrease the number of trainable parameters [171] for the Question Answering process. It can achieve on-par or higher model quality even when it reduces the number of trainable parameters by 10,000 times [122] and the GPU memory required by three times [149], in tasks such as Natural Language Understanding and Generation.

6.3 BioQA BERT- Proposed BERT model with LoRA and LLaMA

Here proposed a new model BioQA BERT by integrating the efficiency of fine-tuning algorithm LoRA and the latest LLM model LLaMA with our previously proposed BERT model.

LoRA is a low-rank adaptation used for fine-tuning the BERT model, aiming to decrease the number of trainable parameters for our Question Answering. While diminishing the number of trainable parameters by 10,000 times and decreasing the GPU memory requirement by threefold, it manages to maintain comparable or superior model quality in tasks such as Question Answering and Answer Generation. [79].

LLaMa is used to enhance the accuracy of the BERT model for Question Answering. The selection of answers is accomplished through external knowledge, incorporating question information along with answer options [155][164]. Prompt optimization is employed to improve the accuracy of answer selection, where correct prompts aid in identifying and generating more probable answers [69]. The output

from the BERT model is fed through a fine-tuned LLaMa. The input to LLaMa is optimized considering prompt length, question subject, and answer position. The architecture of the BioQA-BERT model is shown in the figure 6.1.

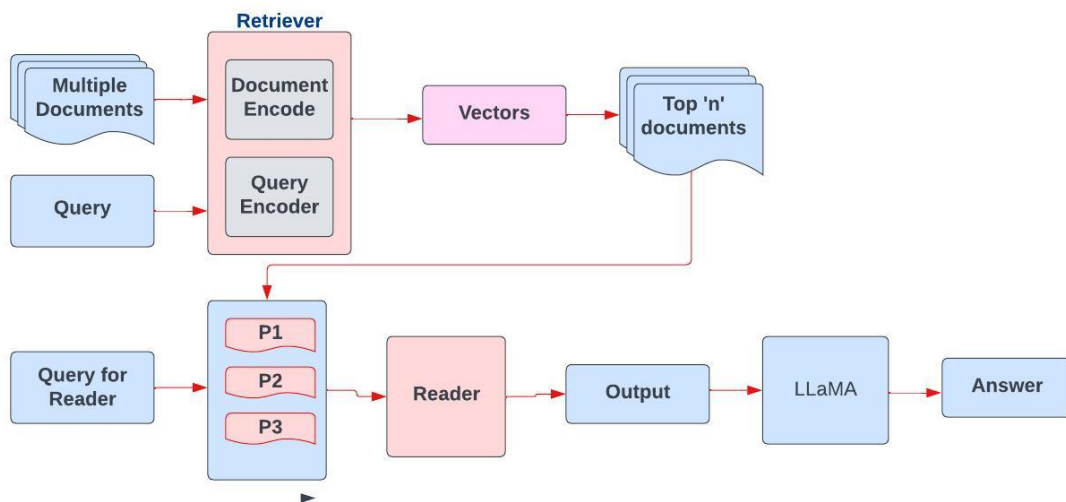


Figure 6.1: Architecture of Proposed BioQA-BERT

The output generated by the previously proposed BERT model serves as the input for LLaMA, a Large Language Model, with an optimized prompt. Subsequently, LLaMA undergoes fine-tuning using our specialized Malayalam health data set MQuAD. This strategic integration of LLaMA into the model has resulted in a noteworthy 6% increase in overall accuracy. Consequently, the F1 score for the proposed BioQA-BERT model stands at an impressive 92%, highlighting the enhanced performance achieved through the synergistic collaboration between BERT and LLaMA in the context of Malayalam Question Answering in the health domain. This integration showcases the efficacy of combining state-of-the-art language models, consequently underscoring the model's proficiency in accurately addressing users' queries in the Malayalam language.

6.4 Result and Discussion

Finally proposed a cutting-edge Question Answering model, BioQA-BERT, which represents a significant advancement in the Question Answering process. This model is crafted through the integration of two key components: the fine-tuning algorithm LoRA and the latest Large Language Model (LLM) named LLaMA, with the optimized BERT model.

The incorporation of LoRA, known for its effective fine-tuning capabilities, allows our BioQA-BERT model to adapt more efficiently to the specific behavior of the task at hand. Simultaneously, the integration of LLaMA, a state-of-the-art Large Language Model, contributes to the overall robustness and linguistic sophistication of the Question Answering System.

The intermediate result taken from the proposed BERT model is shown below

ആയുർവേദശാസ്ത്രപ്രകാരം പ്രമേഹരോഗം പ്രധാനമായും 20 വിധത്തിൽ കാണപ്പെടും. ഉദകമേഹം, ഇക്ഷുമേഹം, സുരാമേഹം, പിഷ്ടമേഹം, ശുക്ലമേഹം, ലാലാമേഹം, ശനൈർമേഹം, സികതാമേഹം, ശീതമേഹം, സാന്ദ്രമേഹം ഇങ്ങനെ കഫദോഷ പ്രധാനമായുണ്ടാകുന്ന മേഹങ്ങൾ പത്തുവിധത്തിലുണ്ട്. ഇതു കൂടാതെ മഞ്ജിഷ്ഠാമേഹം, നീലമേഹം, കാളമേഹം, ഹരിദ്രാമേഹം, ശോണിതമേഹം, ക്ഷാരമേഹം എന്നിങ്ങനെ പിത്തപ്രധാനങ്ങളായമേഹങ്ങൾ ആറും വസാമേഹം, മജ്ജാമേഹം, ഹസ്തിമേഹം, മധുമേഹം എന്നിങ്ങനെ വാതപ്രധാനമായ മേഹങ്ങൾ നാലു വിധത്തിലും കണ്ടുവരുന്നു. അതിനാൽത്തന്നെ ശാസ്ത്രീയമായ ആയുർവേദചികിത്സയ്ക്ക് വിദഗ്ദ്ധമായ രോഗനിർണ്ണയം അത്യാവശ്യമാണ്. മേൽപറഞ്ഞ വിഭാഗങ്ങളിൽ ചികിത്സ അസാധ്യമായി സംഹിതകൾ കരുതുന്ന മധുമേഹത്തോടാണു് നാം ആധുനിക ശാസ്ത്രത്തിൽ പറയുന്ന പ്രമേഹം എന്ന രോഗത്തിന് ഏറെ സാമ്യം. അതിനാൽത്തന്നെ പ്രമേഹചികിത്സ ഏറെ സൂക്ഷിച്ചു ചെയ്യേണ്ടതും കൃത്യമായ വൈദ്യനിർദ്ദേശപ്രകാരം ദീർഘകാലം ശീലി കേണ്ടതുമാണ്. ബലവാനായ രോഗിക്കു വമനം, വിരോചനം തുടങ്ങിയ ശോധന പ്രയോഗങ്ങൾക്ക് ശേഷമാണ് ശമന ചികിത്സ വിധിക്കുന്നത്. എങ്കിലും വീട്ടിൽ പരീക്ഷിച്ചു നോക്കാവുന്ന ചില ലഘു പ്രയോഗങ്ങൾ ചുവടെ നൽകുന്നു. ലഘു പ്രയോഗങ്ങൾ പച്ചമഞ്ഞൾ, പച്ചനെല്ലിക്ക ഇവയുടെ നിർ 25 മില്ലി വീതം ആവശ്യത്തിനു തേനും ചേർത്ത് രാവിലെ വെറുവയറ്റിൽ സേവിക്കുക. നെല്ലിക്ക ഇടിച്ചു പിഴിഞ്ഞ ഒരൗൺസ് നീരിൽ രണ്ട് ടീസ്പൂൺ വരട്ടുമഞ്ഞളിന്റെ പൊടി ചേർത്തു സേവിക്കുന്നതും പച്ചനെല്ലിക്കയും മഞ്ഞളും തുല്യമായി ചേർത്തരച്ചു 20 ഗ്രാം വരെ രാവിലെ വെറും വയറ്റിൽ സേവിക്കുന്നതും ഗുണകരമാണ്. തേറ്റമ്പരൽ നാല്-അഞ്ച് എണ്ണം വെള്ളത്തിൽ ഒരു രാത്രി ഇട്ടുവെച്ചിരുന്നു രാവിലെ കടഞ്ഞെടുത്ത മോരിലരച്ച് സേവിക്കുന്നതു പ്രമേഹശമനീയമാണ്. ചിറ്റുമുതിന്റെ നിർ- 25 മില്ലി-തേൻ ചേർത്ത് രാവിലെ വെറുംവയറ്റിൽ കഴിക്കുന്നതു പ്രമേഹം ശമിപ്പിക്കും. ഏകനായകത്തിൻവേര് (പൊൻകുരണ്ടി, തേറ്റമ്പരൽ എന്നിവ തുല്യ അളവിൽ പൊടിച്ചു രണ്ടു ടേബിൾ സ്പൂൺ വീതം രണ്ടു നേരം സേവിക്കുന്നത് പ്രമേഹശമനത്തിന് ഉത്തമമാണ്. ഏകനായകവും പച്ചമഞ്ഞളും (20 ഗ്രാം) പുളിക്കാത്ത മോരിൽ തുല്യമായ അളവിലരച്ച് രണ്ട് നേരം കഴിച്ചാൽ പ്രമേഹത്തിന് ശമനമുണ്ടാവും. പുളിയരിത്തൊണ്ട്, നെല്ലിക്ക ഇവ കഷായം വച്ച് -50 മി ലി വീതം ഒരു ടീസ്പൂൺ തൈവരഅരിയുടെ തവിടു ചേർത്ത് രണ്ടു നേരം സേവിച്ചാൽ പ്രമേഹരോഗം തടയാം. മുരിക്കിന്റെ തൊലി അരച്ച് (20 ഗ്രാം)- മോരിലോ തേനിലോ ചേർത്ത് സേവിക്കുന്നതു ഹിതമാണ്. അഞ്ചു കുവളത്തില അരച്ചുരുട്ടി രാവിലെ വെറും വയറ്റിൽ കഴിക്കുന്നതു പ്രമേഹം നിയന്ത്രിക്കുന്നതിനു സഹായിക്കുന്നു. ത്രിഫല, മഞ്ഞൾ, ഞാവൽത്തൊലി, നാൽപാമരത്തൊലി, നിർമാതളത്തൊലി, ചെറുളവേര്, പാച്ചോറ്റിത്തൊലി ഇവ ഒന്നിച്ചോ അല്ലെങ്കിൽ ഇവയിൽ ലഭ്യമായ മൂന്നു മരുന്നുകൾ തുല്യ അളവിൽ ഏകനായകവുമായി ചേർത്തോ കഷായം വച്ചു കഴിച്ചാൽ പ്രമേഹം ശമിക്കും. കന്മദം പൊടിച്ചത് അഞ്ചുഗ്രാം വരെ സേവിക്കുന്നത് പ്രമേഹശമനത്തിന് നല്ലതാണ്. തേൻ ചേർക്കേണ്ട യോഗങ്ങളിൽ വിശ്വാസയോഗ്യമായ ചെറുതേൻ ആണ് ഉപയോഗിക്കേണ്ടത്.

Figure 6.2: Result from BERT / Input to LLaMA Model

The final and precise results from the LLaMA model are shown below.

<p>Prompt 1: പ്രമേഹത്തിന് ആയുർവേദ ലഘുചികിത്സകൾ എന്തെല്ലാം?</p> <p>Answer : പച്ചമഞ്ഞളും, പച്ചനെല്ലിയും നിരിൽ ഉള്ള മിക്സർ ഉപയോഗിച്ച് 25 മില്ലിയേറ്റും നിർ തേനും ചേർത്ത് രാത്രി വെറുവയറ്റിൽ കുടിക്കുകയും, എല്ലാവർക്കും പരിഹാരമായി. നെല്ലിക്ക പിഴിഞ്ഞ ഒരൗൺസ് നിരിൽ മഞ്ഞളിന്റെ പൊടി ചേർത്തു സേവിക്കുകയും, അതും പച്ചനെല്ലിയും തുല്യമായി ചേർത്ത് 20 ഗ്രാം രാത്രിയിൽ വെറും വയറ്റിൽ സേവിക്കുകയും ഗുണകരമാണ്. തേറ്റമ്പരൽ നാല്-അഞ്ച് എണ്ണം വെള്ളത്തിൽ ഇട്ടുവെച്ചിരുന്നു രാത്രിയ്ക്ക് പുറത്തു വെച്ച് സേവിക്കുകയും, അതിനുശേഷം രാവിലെ കടഞ്ഞെടുത്ത മോരിലേക്ക് സേവിക്കുകയും പ്രമേഹശമനീയമാണ്. ചിറ്റുമുതിൻറെ നിർ- 25 മില്ലി-തേൻ ചേർത്ത് രാത്രിയിൽ വെറുംവയറ്റിൽ കഴിക്കുകയും പ്രമേഹം ശമിപ്പിക്കുകയും. ഏകനായകത്തിൻവേർ (പൊൻകുരണ്ടി, തേറ്റമ്പരൽ എന്നിവ) രണ്ടു ടേബിൾ സ്പൂൺ രണ്ടു നേരം സേവിക്കുകയും പ്രമേഹശമനത്തിന് ഉത്തമമാണ്. ഏകനായകവും പച്ചമഞ്ഞളും (20 ഗ്രാം) പുളിക്കാത്ത മോരിൽ തുല്യമായ അളവിലേക്ക് രണ്ട് നേരം കഴിച്ചാൽ പ്രമേഹത്തിന് ശമനമുണ്ടാവും</p>
<p>Prompt: പ്രമേഹ ചികിത്സ എത്ര തരം ?</p>
<p>Answer : അഞ്ചു തരം</p>

Figure 6.3: Final Results from BioQA-BERT Model

The incorporation of LLaMA into the optimized BERT model will yield results that are more precise, likely, and accurate. The performance evaluation of the newly proposed BioQA-BERT model is shown in the table.

Table 6.1: Performance Analysis of Different proposed Models

Models	Precision	Recall	F1 Score	Speed(No of Doc/Sec)
BERT(With out LoRA)	0.93	0.80	86.45%	150
BERT with LoRA	0.88	0.92	90.28%	320
BioQA-BERT (with LoRA and LLaMA)	0.94	0.9	92.64%	280

Upon comprehensive evaluation, the results indicate a notable increase in the accuracy of the retrieved answers. This enhancement underscores the efficacy of our BioQA-BERT model, showcasing its ability to provide more precise and relevant responses to user queries. The synergistic collaboration of LoRA and LLaMA with BERT contributes to the model’s proficiency in handling intricate language features, making it a valuable asset in the field of Question Answering.

Table 6.2: System Specifications

Tesla P100 Pale 16GB GPU
Consumed CUDA memory
Tensor flow packages and features
Training epoch = 3
Training loss = 0.062

6.5 Conclusion

This work proposed an innovative BioQA-BERT model, a significant enhancement in Question Answering Systems. By integrating the LoRA (Low-Rank Adaptation) fine-tuning algorithm and the latest LLM (LLaMA) with our existing BERT model. This study aims to achieve a more accurate and efficient response to user’s queries, particularly in the health domain. LoRA, a parameter-efficient fine-tuning technique, significantly reduces trainable parameters without compromising the model’s quality, leading to enhanced training efficiency. LLaMA, the latest Large Language Model developed by Meta, contributes to the model’s linguistic sophistication and overall robustness.

The low-rank adaptation algorithm, LoRA, facilitates fine-tuning of our BERT model, reducing the number of trainable parameters and GPU memory requirements while maintaining or improving model quality. The introduction of LLaMA into our BioQA-BERT model aims to capitalize on its capabilities for accurate Question Answering in the Malayalam language. LLaMA is fine-tuned using a specialized Malayalam health data set MQuAD, resulting in a remarkable 6% increase in overall accuracy. The BioQA-BERT model achieves an outstanding F1 score of 92.64%, highlighting its efficacy in accurately addressing health-related queries in the Malayalam language.

Chapter 7

Performance Analysis and Discussion of Various Proposed Models

7.1 Introduction

In this chapter, conducted a comparative analysis, of the results obtained from various proposed Question Answering models. Building upon the discussions presented in the preceding sections, this work sought to systematically compare the performance and accuracy of these diverse models. The evaluation of the model involved a comprehensive examination of key metrics and outcomes, considering factors such as precision, recall, and F1 score. This comparative analysis, aimed to discern the strengths and limitations of each model, providing valuable insights into their respective capabilities in addressing the complexities inherent in Question Answering tasks. By exploiting a structured approach and drawing upon the detailed discussions from earlier sections, this study offers a clear understanding of how these models fare in diverse scenarios and applications, contributing to the broader landscape of Question Answering research.

7.2 Comparative Analysis

This research work proposed various models for Question Answering tasks in the health domain with different NLP techniques, Word Embedding algorithms, and transformer-based models. In the initial stage, developed a Word Embedding-based Question Answering model. Then performed the Question Answering process using different BERT models with Reader-Retriever pipeline architecture. Compare the performance of different BERT variants ALBERT, DistilBERT, StructBERT, and RoBERTa. Finally proposed a new model BioQA-BERT, with the advantage of the LoRA fine-tuning algorithm and the latest LLM- LLaMA for improving the accuracy of the Question Answering task. The performance analysis of different models in the aspect of F1 score is shown in Table 7.1.

The course of this research endeavor introduced a diverse array of models for Question Answering tasks within the health domain. The initial phase involved the application of various Natural Language Processing (NLP) techniques and Word Embedding methods. Subsequently, then customized the model by incorporating BERT models with Reader-Retriever pipeline architecture. To comprehensively assess their efficacy, perform comparative analysis across different BERT variants, namely ALBERT, DistilBERT, StructBERT, and RoBERTa.

Upon evaluating the models, observed their respective strengths and weaknesses in the context of Question Answering tasks, particularly focusing on their F1 score metrics. Recognizing the potential for further improvement, refined the BERT model by integrating the LoRA fine-tuning algorithm. Additionally, used the cutting-edge Large Language Model - LLaMA to enhance the overall accuracy of the Question Answering task.

The detailed performance analysis, including the F1 scores of the various models, is presented in Table 7.1, providing a comprehensive overview of their comparative effectiveness in addressing health-related inquiries.

Table 7.1: Performance Analysis of Different Question Answering Models

Models	F1 Score
Word Embedding (CBOW)	72.41%
Word Embedding (Skip-Gram)	76.20%
ALBERT	68.53%
DistilBERT	62.32%
StructBERT	76.42%
RoBERTa	78.65%
ALBERT (With Reader Retriever)	74.02%
DistilBERT (With Reader Retriever)	69.28%
StructBERT (With Reader Retriever)	84.19%
RoBERTa (With Reader Retriever)	86.45%
BERT (with LORA)	90.28%
BioQA BERT(With LORA and LLaMA)	92.64%

The accuracy of the Question Answering model is increased while introducing Reader-Retriever pipeline architecture with optimized BERT model RoBERTa in comparison to its counterparts. Notably, the enhanced BioQA-BERT model, fortified with the sophisticated LoRA fine-tuning algorithm and the state-of-the-art LLaMA model, exhibited exceptional performance, surpassing all other models in the evaluation.

The augmentation of the model architecture, especially the incorporation of RoBERTa, has played a pivotal role in refining the model’s comprehension and response capabilities. The careful optimization of the BERT model, specifically with RoBERTa, demonstrates a notable stride forward in achieving high accuracy in addressing health-related queries.

The incorporation of LoRA and LLaMA improves the model’s capacity to identify subtle contextual differences in health texts. Thereby contributing to its superior performance. This holistic approach, combining advanced architectural design with cutting-edge algorithms, establishes the BioQA-BERT model as a front-runner in the aspect of health Question Answering tasks, setting a new benchmark for accuracy and effectiveness.

7.3 Conclusion

This chapter performed a comparative analysis of the performance of various proposed Question Answering models. With the insights presented in the preceding sections, the objective was to systematically evaluate the performance and accuracy of these diverse models. From the analysis, it is clear that our latest proposed Question Answering model BioQA-BERT has the highest accuracy of all other models. We present a lucid comprehension of how these models perform across different scenarios and applications, thus making substantial contributions to the overarching field of Question Answering research.

Chapter 8

Conclusion

This study employed the development of a Word Embedding and a hybrid BERT and Large Language Model-based model to enhance the Malayalam Question Answering process in the health domain. Initially created a Malayalam health Question Answering Data set MQuAD. Then re-trained and fine-tuned different BERT models with the health data set. Customized a new BERT model with Reader Retriever pipeline architecture with different BERT variants, and compared the performance. Finally proposed an efficient model BioQA-BERT by integrating an efficient fine-tuning algorithm and the latest large language model to improve the effectiveness and accuracy of the Question Answering System.

In the initial phase, developed a Word Embedding-based model using Word2Vec algorithms CBOW and Skip-Gram. Among them, the Skip-Gram algorithm demonstrated the highest accuracy on our data set. In the subsequent phase, performed a detailed analysis of different BERT models for Question Answering. Employed various pre-trained BERT models and re-trained them with the Malayalam health Question Answering data set. Additionally, fine-tuned the model with the MQuAD data set.

This process incorporated hybrid Reader Retriever pipeline architecture into each BERT model and then compared the performance of different BERT models. In the final stage, enhanced the performance of the BERT model by integrating the LoRA algorithm and the latest LLM model, LLaMA, to improve the accuracy of the Question Answering task.

Experimental results indicated an increase in the F1 score for all BERT models after the introduction of the Reader Retriever pipeline architecture with LoRA and LLaMA. Among them, the optimized BERT model RoBERTa with the Reader Retriever pipeline and LLaMA showed the highest accuracy. In conclusion, while the study offers significant insights into the state and challenges of Malayalam Question Answering, especially in the health domain, it is bound by the inherent complexities of the language and the limited precedent in this field.

Chapter 9

Recommendations

9.1 Introduction

This study proposed a semantic Question Answering within the health domain. While the findings and contributions mark a significant advancement, the study identifies several promising avenues for future research and development that could further augment the capabilities and applicability of the system. Looking ahead, this section contains a strategic plan that involves the integration of cutting-edge semantic methods and technologies into the existing system, aiming to enhance its sophistication and efficacy.

9.2 Recommendations for Future Research

This study undertook a semantic-level Question Answering within the health domain. Although our findings and contributions represent a noteworthy advancement, there exist several promising avenues for future research and development that could augment the capabilities and applicability of our system. Looking ahead, broaden the scope of this research by integrating more state-of-the-art semantic

methods and technologies into this system.

While the current model has showcased robust performance within the intricacies of the health domain, acknowledge the importance of extending its capabilities to encompass open-domain Question Answering. This strategic expansion will not only contribute to the versatility of our system but also enhance its utility in addressing a broader range of inquiries.

Furthermore, recognizing the linguistic diversity in India, extending the system's linguistic capabilities to handle queries in multiple Indian languages. This multi-lingual approach aims to make the Question Answering System more inclusive and accessible to a wider audience, catering to the diverse linguistic landscape of the country.

This system serves as a foundation for future endeavors, and this study is committed to enhancing the research by incorporating advanced methodologies, expanding into open-domain Question Answering, and ensuring linguistic inclusivity for a more comprehensive and efficient Question Answering System.

In the ongoing commitment to elevate the user experience and make the system more accessible, have strategic plans to integrate multi-modal capabilities. This entails the incorporation of speech recognition and audio features, enabling users to engage with the system through spoken language, thereby broadening the avenues of interaction and facilitating a more intuitive user interface.

To further enrich the user-centric experience, integrate sentiment analysis and emotional awareness into the system. This enhancement will enable the system to comprehend the literal meaning of user queries and discern the emotional context, fostering empathetic responses tailored to the user's state of mind.

Recognizing the importance of transparency in AI systems, committed to developing methods that elucidate the system's reasoning process. This transparency initiative aims to build trust with users by providing insights into how the system arrives at its responses, fostering a sense of understanding and reliability.

A pivotal aspect of future strategy involves actively seeking user feedback to garner valuable insights into the system's performance and identify areas for improvement. This iterative process of refinement, driven by real-world usage and user input, ensures that the system evolves to meet the dynamic needs and expectations of the users.

9.3 Conclusion

In summary, the integration of multi-modal capabilities, the infusion of emotional intelligence into the system, transparency in reasoning processes, and a continuous feedback loop with users. These initiatives collectively aim to create a more sophisticated, user-friendly, and trustworthy system that aligns seamlessly with user needs and preferences.

List of Publications

Journal Publications

1. Liji SK, Muhammed Ilyas P. "Review and Analysis of Different Approaches to Semantic Level Question Answering and Information Retrieval", International Journal of Science and Research (IJSR), Volume 10 Issue 1, January 2021, 1238-1244.
2. Liji SK, Muhammed Ilyas P. "Effective Malayalam Query Processing And Question Answering System For Medical Domain", Journal of the Maharaja Sayajirao University of Baroda, UGC Care group I ISSN: 0025-0422, Volume 55, NO 1(II), 2020-2021.
3. Liji S K, Muhammed Ilyas P. "A Factoid Dialogue System in the Malayalam Language for Health Care Using Word Embedding and BERT", Journal of Fundamental and Comparative Research Vol. VIII, No. 2: 2022 ISSN: 2277-7067.
4. Liji SK, Muhammed Ilyas P. Interactive Malayalam Question Answering System: A Neural Word Embedding And Similarity Measure Based Approach, Journal of Advanced Zoology, ISSN: 0253-7214. Volume 44 Issue 5 Year. 2023 Page 605-611. Web Of Science, UGC care group II.
<https://jazindia.com/index.php/jaz/article/view/3081/2508>
5. Liji SK, Muhammed Ilyas P. "A Neural Word Embedding-based Transformer Models for Improving Malayalam Question Answering on Health Domain"

International Journal of Intelligent Systems and Applications in Engineering
ISSN: 2147-6799, IJISAE, 2024, 12(14s), 542â547, Scopus, Elsevier.
<https://ijisae.org/index.php/IJISAE/article/view/4691/3365>.

6. Liji SK, Muhammed Ilyas P. "Comparative Analysis of Bidirectional Encoder Representation from Transformer Models for Conversational Question Answering", in Communication with Soft Computing, Springer Journal.

Conferences Presentations

1. Liji S K, Muhammed Ilyas P. "A Factoid Dialogue System in the Malayalam Language For Health Care Using Word Embedding And BERT", International Conference on Innovations and Recent Trends in Computer Science (ICIRTCS-2022), March 2022.
2. Liji S K, Muhammed Ilyas P. "Semantic Malayalam Dialogue System For COVID-19 Question Answering", International Conference on Advances in Computing & Communications, ICACC. IEEE. <https://doi.org/10.1109/ICACC-202152719.2021.9708150>, 15 February 2022.
3. Liji S K, Muhammed Ilyas P. "An Efficient Malayalam Information Retrieval System For Disease Diagnosis With AIML", International Conference on INNOVATIONS IN COMPUTER SCIENCE AND TECHNOLOGY(ICICSIT) Nehru Arts and Science College, Coimbatore, October 2019.

Bibliography

- [1] Zahra Abbasiantaeb and Saeedeh Momtazi. Text-based question answering from information retrieval and deep neural network perspectives: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(6):e1412, 2021.
- [2] Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (IC-CWAMTIP)*, pages 117–121. IEEE, 2020.
- [3] Waheeb Ahmed, Ajusha PV, et al. Web-based arabic question answering system using machine learning approach. *International Journal of Advanced Research in Computer Science*, 8(1), 2017.
- [4] Ali Saleh Alammary. Bert models for arabic text classification: a systematic review. *Applied Sciences*, 12(11):5720, 2022.
- [5] Kholoud Alsubhi, Amani Jamal, and Areej Alhothali. Deep learning-based approach for arabic open domain question answering. *PeerJ Computer Science*, 8:e952, 2022.
- [6] SM Archana, Naima Vahab, Rekha Thankappan, and C Raseek. A rule based question answering system in malayalam corpus using vibhakthi and pos tag analysis. *Procedia Technology*, 24:1534–1541, 2016.

- [7] Piyush Arora, Jennifer Foster, and Gareth JF Jones. Query expansion for sentence retrieval using pseudo relevance feedback and word embedding. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 97–103. Springer, 2017.
- [8] Arjun Babu and L Sindhu. An information retrieval system for malayalam using query expansion technique. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1559–1564. IEEE, 2015.
- [9] Pratibha Bajpai, Parul Verma, and Syed Q Abbas. Two-level disambiguation model for query translation. *International Journal of Electrical and Computer Engineering (IJECE)*, 8(5):3923–3932, 2018.
- [10] Srijan Bansal, Suraj Tripathi, Sumit Agarwal, Sireesh Gururaja, Aditya Srikanth Veerubhotla, Ritam Dutt, Teruko Mitamura, and Eric Nyberg. R3: Refined retriever-reader pipeline for multidoc2dial. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 148–154, 2022.
- [11] Thiziri Belkacem, Taoufiq Dkaki, Jose G Moreno, and Mohand Boughanem. amv-lstm: an attention-based model with multiple positional text matching. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 788–795, 2019.
- [12] Meghana Moorthy Bhat, Rui Meng, Ye Liu, Yingbo Zhou, and Semih Yavuz. Investigating answerability of llms for long-form question answering. *arXiv preprint arXiv:2309.08210*, 2023.
- [13] MS Bindu and Idicula Sumam Mary. *Design and development of a named entity based question answering system for Malayalam language*. PhD thesis, Cochin University Of Science And Technology, 2012.

- [14] Pietro Bongini, Federico Becattini, and Alberto Del Bimbo. Is gpt-3 all you need for visual question answering in cultural heritage? In *European Conference on Computer Vision*, pages 268–281. Springer, 2022.
- [15] Daniel Bratulic, Marin Kacan, and Josip Šarić. Simple two-step factoid question answering based on skip-gram embeddings. *Text Analysis and Retrieval 2017 Course Project Reports*, page 11.
- [16] Thomas M Brill, Laura Munoz, and Richard J Miller. Siri, alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications. In *The Role of Smart Technologies in Decision Making*, pages 35–70. Routledge, 2022.
- [17] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*, 2022.
- [18] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.
- [19] Rakesh Chada and Pradeep Natarajan. Fewshotqa: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models. *arXiv preprint arXiv:2109.01951*, 2021.
- [20] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.
- [21] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2020.

- [22] Jen-Tzung Chien and Ting-An Lin. Supportive attention in end-to-end memory networks. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.
- [23] Byeongmin Choi, YongHyun Lee, Yeunwoong Kyung, and Eunchan Kim. Albert with knowledge graph encoder utilizing semantic similarity for common-sense question answering. *arXiv preprint arXiv:2211.07065*, 2022.
- [24] Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. Fine-tune bert with sparse self-attention mechanism. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3548–3553, 2019.
- [25] Zhuyun Dai and Jamie Callan. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 985–988, 2019.
- [26] Parag Pravin Dakle, SaiKrishna Rallabandi, and Preethi Raghavan. Understanding bloom: An empirical study on diverse nlp tasks. *arXiv preprint arXiv:2211.14865*, 2022.
- [27] Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information systems*, 55:529–569, 2018.
- [28] Phuc Do and Truong HV Phan. Developing a bert-based triple classification model using knowledge graph embedding for question answering system. *Applied Intelligence*, 52(1):636–651, 2022.
- [29] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, 2017.

- [30] Alexander Dür, Andreas Rauber, and Peter Filzmoser. Reproducing a neural question answering architecture applied to the squad benchmark dataset: Challenges and lessons learned. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 102–113. Springer, 2018.
- [31] Jacob Eisenstein. *Natural Language Processing*. The MIT Press, Cambridge, MA, 2019.
- [32] Fan Fang, Bo-Wen Zhang, and Xu-Cheng Yin. Semantic sequential query expansion for biomedical article search. *IEEE Access*, 6:45448–45457, 2018.
- [33] Yue Feng, Zhen Han, Mingming Sun, and Ping Li. Multi-hop open-domain question answering over structured and unstructured knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 151–156, 2022.
- [34] Asbjørn Følstad, Marita Skjuve, and Petter Bae Brandtzaeg. Different chatbots for different purposes: towards a typology of chatbots to understand interaction design. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*, pages 145–156. Springer, 2019.
- [35] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
- [36] M^Á García-Cumbreras, F Martínez-Santiago, and LA Ureña-López. Architecture and evaluation of bruja, a multilingual question answering system. *Information retrieval*, 15:413–432, 2012.
- [37] Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. Word embedding evaluation and combination. In *Proceedings of the Tenth Interna-*

- tional Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305, 2016.
- [38] Hadi Abdi Ghavidel, Amal Zouaq, and Michel C Desmarais. Using bert and xlnet for the automatic short answer grading task. In *CSEU (1)*, pages 58–67, 2020.
- [39] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE, 2020.
- [40] Yoav Goldberg. A primer on neural network models for natural language processing. *arXiv preprint arXiv:1510.00726*, 2015.
- [41] Jiayan Guo, Peiyan Zhang, Chaozhuo Li, Xing Xie, Yan Zhang, and Sunghun Kim. Evolutionary preference learning via graph nested gru ode for session-based recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 624–634, 2022.
- [42] Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. *Natural language processing: python and NLTK*. Packt Publishing Ltd, 2016.
- [43] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [44] Xinghang Hu. Enhancing answer selection in community question answering with pre-trained and large language models. *arXiv preprint arXiv:2311.17502*, 2023.
- [45] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over para-

- graphs. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 633–644, 2014.
- [46] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*, 2020.
- [47] Sonal Jain and Tripti Dodiya. Rule-based architecture for medical question answering system. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, pages 1225–1233. Springer, 2014.
- [48] Nivas Jayaseelan. Llama 2: The new open source language model.
- [49] R Jayashree and N Niveditha. Natural language processing based question answering using vector space model. In *Proceedings of Sixth International Conference on Soft Computing for Problem Solving: SocProS 2016, Volume 2*, pages 368–375. Springer, 2017.
- [50] Girish Nath Jha. The tdil program and the indian language corpora initiative (ilci). In *LREC*. New Delhi, India, 2010.
- [51] Zhixue Jiang, Chengying Chi, and Yunyun Zhan. Research on medical question answering system based on knowledge graph. *IEEE Access*, 9:21094–21101, 2021.
- [52] Kun Jing and Jungang Xu. A survey on neural network language models. *arXiv preprint arXiv:1906.03591*, 2019.
- [53] Dan Jurafsky and James H. Martin. *Speech and Language Processing*. Pearson, New York, NY, 3rd edition, 2019.
- [54] Prathyusha Kanakam, S Mahaboob Hussain, and D Suryanarayana. Hpl algorithm for semantic information retrieval with rdf and sparql. *Research Article. Jr. of Industrial Pollution Control*, 33, 2017.

- [55] SARAVANAKUMAR Kandasamy and Aswani Kumar Cherukuri. Information retrieval for question answering system using knowledge-based query reconstruction by adapted lesk and latent semantic analysis. *International Journal of Computer Science & Applications*, 14(2):31–46, 2017.
- [56] Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. *arXiv preprint arXiv:2005.00987*, 2020.
- [57] Jashanjot Kaur and Preetpal Kaur Buttar. Stopwords removal and its algorithms based on different methods. *International Journal of Advanced Research in Computer Science*, 9(5):81–88, 2018.
- [58] Navjot Kaur and Himanshu Aggarwal. Query reformulation approach using domain-specific ontology for semantic information retrieval. *International Journal of Information Technology*, 13:1745–1753, 2021.
- [59] Shomi Khan, Khadiza Tul Kubra, and Md Mahadi Hasan Nahid. Improving answer extraction for bengali q/a system using anaphora-cataphora resolution. In *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–6. IEEE, 2018.
- [60] Divya Khyani, BS Siddhartha, NM Niveditha, and BM Divya. An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology*, 22(10):350–357, 2021.
- [61] Yohan Kim, Seongdeok Bang, Jiu Sohn, and Hyoungkwan Kim. Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers. *Automation in construction*, 134:104061, 2022.
- [62] Oleksandr Kolomiyets and Marie-Francine Moens. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434, 2011.

- [63] Natalia Konstantinova and Constantin Orasan. Interactive question answering. In *Emerging applications of natural language processing: concepts and new research*, pages 149–169. IGI Global, 2013.
- [64] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.
- [65] Venkatesh Krishnamoorthy. Evolution of reading comprehension and question answering systems. *Procedia Computer Science*, 185:231–238, 2021.
- [66] Florian Kunneman, Thiago Castro Ferreira, Emiel Kraahmer, and Antal Van Den Bosch. Question similarity in community question answering: A systematic exploration of preprocessing methods and models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 593–601, 2019.
- [67] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
- [68] Lane Lane, Hannes Howard, and Frank Hapke. *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning Publications, New York, NY, 2019.
- [69] Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*, 2022.
- [70] Yukyung Lee, Jina Kim, and Pilsung Kang. Lanobert: System log anomaly detection based on bert masked language model. *Applied Soft Computing*, 146:110689, 2023.
- [71] Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, Hong Yu, et al. Fine-tuning bidirectional encoder representations from

- transformers (bert)-based models on large-scale electronic health record notes: an empirical study. *JMIR medical informatics*, 7(3):e14830, 2019.
- [72] Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. Loftq: Lora-fine-tuning-aware quantization for large language models. *arXiv preprint arXiv:2310.08659*, 2023.
- [73] SK Liji and Muhamed P Ilyas. Analysis on semantic level information retrieval and query processing. Technical report, EasyChair, 2020.
- [74] SK Liji and P Muhamed Ilyas. Semantic malayalam dialogue system for covid-19 question answering using word embedding and cosine similarity. In *2021 International Conference on Advances in Computing and Communications (ICACC)*, pages 1–6. IEEE, 2021.
- [75] SK Liji and VL Lajish. An efficient malayalam query processing system for university enquiry. In *Proceedings of the Eighth National Conference on Indian Language Computing (NCILC)*, 2018.
- [76] Andy T Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. Qaner: Prompting question answering models for few-shot named entity recognition. *arXiv preprint arXiv:2203.01543*, 2022.
- [77] Shang Liu, Wanli Gu, Gao Cong, and Fuzheng Zhang. Structural relationship representation learning with graph embedding for personalized product search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 915–924, 2020.
- [78] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [79] Sanzana Karim Lora, Ishrat Jahan, Rahad Hussain, Rifat Shahriyar, and ABM Alim Al Islam. A transformer-based generative adversarial learning to

- detect sarcasm from bengali text with correct classification of confusing text. *Heliyon*, 2023.
- [80] P Manasamithra and HC Vijayalakshmi. Nlp for information retrieval using b trees. *International Journal of Computer Applications*, 975:8887.
- [81] TN Manjunath, Deepa Yogish, S Mahalakshmi, and HK Yogish. Smart question answering system using vectorization approach and statistical scoring method. *Materials Today: Proceedings*, 80:3719–3725, 2023.
- [82] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
- [83] Kavya Manohar, AR Jayan, and Rajeev Rajan. Quantitative analysis of the morphological complexity of malayalam language. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 71–78. Springer, 2020.
- [84] Timothee Mickus, Denis Paperno, Matthieu Constant, and Kees van Deemter. What do you mean, bert? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, 2020.
- [85] Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al. Between words and characters: a brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*, 2021.
- [86] Bonan Min, Hayley Ross, Elier Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

- [87] Amit Mishra and Sanjay Kumar Jain. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361, 2016.
- [88] Piyush Mital, Saurabh Agrawal, Bhargavi Neti, Yashodhara Haribhakta, Vibhavari Kamble, Krishnanjan Bhattacharjee, Debashri Das, Swati Mehta, and Ajai Kumar. Graph-based question answering system. In *2018 International Conference on Advances in Computing, Communications, and Informatics (ICACCI)*, pages 1798–1802. IEEE, 2018.
- [89] Timo Möller, Julian Risch, and Malte Pietsch. Germanquad and germandpr: Improving non-english question answering and passage retrieval. *arXiv preprint arXiv:2104.12741*, 2021.
- [90] Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456, 2021.
- [91] Khalid Nassiri and Moulay Akhloufi. Transformer models used for text-based question answering systems. *Applied Intelligence*, 53(9):10602–10635, 2023.
- [92] Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima López, Eulália Farré-Maduell, Luis Gasco, Martin Krallinger, and Georgios Paliouras. Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 227–250. Springer, 2023.
- [93] Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. Overview of bioasq 2022: the tenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *Inter-*

- national Conference of the Cross-Language Evaluation Forum for European Languages*, pages 337–361. Springer, 2022.
- [94] Chidinma A Nwafor and Ikechukwu E Onyenwe. An automated multiple-choice question generation using natural language processing techniques. *arXiv preprint arXiv:2103.14757*, 2021.
- [95] Nouha Othman, Rim Faiz, and Kamel Smaïli. Enhancing question retrieval in community question answering using word embeddings. *Procedia Computer Science*, 159:485–494, 2019.
- [96] Nouha Othman, Rim Faiz, and Kamel Smaïli. Learning english and arabic question similarity with siamese neural networks in community question answering services. *Data & Knowledge Engineering*, 138:101962, 2022.
- [97] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.
- [98] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. Deeprank: A new deep architecture for relevance ranking in information retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 257–266, 2017.
- [99] Harshali B Patil, BV Pawar, and Ajay S Patil. A comprehensive analysis of stemmers available for indic languages. *Int. J. Nat. Lang. Comput*, 5(1):45–55, 2016.
- [100] Ivo Pisařovic, František Dařena, David Procházka, and Vít Janiš. Preprocessing of normative documents for interactive question answering. *Expert Systems with Applications*, 191:116314, 2022.

- [101] B Premjith and KP Soman. Deep learning approach for the morphological synthesis in malayalam and tamil at the character level. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6):1–17, 2021.
- [102] B Premjith, KP Soman, M Anand Kumar, and D Jyothi Ratnam. Embedding linguistic features in word embedding for preposition sense disambiguation in englishâmalayalam machine translation context. *Recent advances in computational intelligence*, pages 341–370, 2019.
- [103] B Premjith, V Sowmya, Bharathi Raja Chakravarthi, Rajeswari Natarajan, K Nandhini, Abirami Murugappan, B Bharathi, M Kaushik, Prasanth Sn, et al. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79, 2023.
- [104] Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136, 2019.
- [105] Dennis Quan, David Huynh, and David R Karger. Haystack: A platform for authoring end user semantic web applications. In *International semantic web conference*, pages 738–753. Springer, 2003.
- [106] Reji Rahmath K, PC Reghu Raj, and Rafeeqe PC. Malayalam question answering system using deep learning approaches. *IETE Journal of Research*, pages 1–13, 2022.
- [107] Prakash Ranjan and Rakesh Chandra Balabantaray. Question answering system for factoid based question. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 221–224. IEEE, 2016.

- [108] Shaina Raza, Brian Schwartz, and Laura C Rosella. Coquad: a covid-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC bioinformatics*, 23(1):1–28, 2022.
- [109] Sara Renjit and Sumam Idicula. Natural language inference for malayalam language using language agnostic sentence representation. *PeerJ Computer Science*, 7:e508, 2021.
- [110] PK Reshma and VL Lajish. Ontology based semantic information retrieval model for university domain. *Int. J. Appl. Eng. Res*, 13:12142–12145, 2018.
- [111] Philip Resnik. Wsd in nlp applications. In *Word Sense Disambiguation: Algorithms and Applications*, pages 299–337. Springer, 2007.
- [112] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021.
- [113] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.
- [114] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014. <https://www.semanticscholar.org/paper/word2vec-Parameter-Learning-Explained-Rong/940e8c477f3e7ddb1d3aa2f216a38c8f9486e544/figure/2>.
- [115] Md Rashad Al Hasan Rony, Debanjan Chaudhuri, Ricardo Usbeck, and Jens Lehmann. Tree-kgqa: an unsupervised approach for question answering over knowledge graphs. *IEEE Access*, 10:50467–50478, 2022.
- [116] Denis Rothman. *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*. Packt Publishing Ltd, 2021.

- [117] Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. Llama 2: Early adopters’ utilization of meta’s new open-source pretrained model. 2023.
- [118] Dwaipayan Roy, Debasis Ganguly, Sumit Bhatia, Srikanta Bedathur, and Mandar Mitra. Using word embeddings for information retrieval: How collection and term normalization choices affect performance. In *Proceedings of the 27th ACM International Conference on information and Knowledge Management*, pages 1835–1838, 2018.
- [119] Pradeep Kumar Roy, Sunil Saumya, Jyoti Prakash Singh, Snehasish Banerjee, and Adnan Gutub. Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review. *CAAI Transactions on Intelligence Technology*, 8(1):95–117, 2023.
- [120] Navin Sabharwal, Amit Agrawal, Navin Sabharwal, and Amit Agrawal. Bert model applications: Question answering system. *Hands-on Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing*, pages 97–137, 2021.
- [121] Navin Sabharwal and Amit Agarwal. *Hands-On Question Answering System with BERT*. Springer, 2022.
- [122] Ruben M Sandoval, Antonio-Javier Garcia-Sanchez, and Joan Garcia-Haro. Optimizing and updating lora communication parameters: A machine learning approach. *IEEE Transactions on Network and Service Management*, 16(3):884–895, 2019.
- [123] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [124] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983, 2023.
- [125] Yikang Shen, Wenge Rong, Nan Jiang, Baolin Peng, Jie Tang, and Zhang Xiong. Word embedding based correlation model for question/answer matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [126] RAZIA SIDDIQUE and MARZIA ASHEQUE. *INTELLIGENT MCQ OPTION GENERATOR AND ANSWERING SYSTEM FOR BANGLA LANGUAGE*. PhD thesis, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, 2020.
- [127] Mittul Singh, Arunav Mishra, Youssef Oualil, Klaus Berberich, and Dietrich Klakow. Long-span language models for query-focused unsupervised extractive text summarization. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 657–664. Springer, 2018.
- [128] Ritika Singh and Satwinder Singh. Text similarity measures in news articles by vector space model using nlp. *Journal of The Institution of Engineers (India): Series B*, 102:329–338, 2021.
- [129] Vaishali Singh and Sanjay K Dwivedi. Personalized approach for automated question answering in restricted domain. *International Journal of Information Technology*, 12(1):223–229, 2020.
- [130] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.
- [131] Sheetal S Sonawane and Parag Kulkarni. Concept-based document similarity using graph model. *International Journal of Information Technology*, pages

1–12, 2019.

- [132] Kaan Sonmezoz and Mehmet Fatih Amasyali. Same sentence prediction: A new pre-training task for bert. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE, 2021.
- [133] Nadia Soudani, Ibrahim Bounhas, and Yahya Slimani. A hybrid approach for standardized dictionary-based knowledge extraction for arabic morpho-semantic retrieval. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pages 47–51. IEEE, 2018.
- [134] Ivan Srba and Maria Bielikova. A comprehensive survey and classification of approaches for community question answering. *ACM Transactions on the Web (TWEB)*, 10(3):1–63, 2016.
- [135] K Sreelakshmi, B Premjith, and Soman Kp. Amrita_cen_nlp@dravidianlangtech-eacl2021: deep learning-based offensive language identification in malayalam, tamil and kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254, 2021.
- [136] Ieva Staliūnaitė and Ignacio Iacobacci. Compositional and lexical semantics in roberta, bert and distilbert: A case study on coqa. *arXiv preprint arXiv:2009.08257*, 2020.
- [137] Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 203–211, 2019.
- [138] Peng Su and K Vijay-Shanker. Investigation of improving the pre-training and fine-tuning of bert model for biomedical relation extraction. *BMC bioinformatics*, 23(1):120, 2022.

- [139] Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020.
- [140] Xianghui Sun, Yunjie Ji, Baochang Ma, and Xiangang Li. A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *arXiv preprint arXiv:2304.08109*, 2023.
- [141] Kanchan M Tarwani and Swathi Edem. Survey on recurrent neural network in natural language processing. *Int. J. Eng. Trends Technol*, 48(6):301–304, 2017.
- [142] D Thenmozhi and Chandrabose Aravindan. Ontology-based tamil–english cross-lingual information retrieval system. *Sādhanā*, 43:1–14, 2018.
- [143] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [144] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [145] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, and Christina Unger. Hawk–hybrid question answering using linked data. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31–June 4, 2015. Proceedings 12*, pages 353–368. Springer, 2015.
- [146] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. Question rewriting for conversational question answering. In *Proceedings of*

- the 14th ACM international conference on web search and data mining*, pages 355–363, 2021.
- [147] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [148] Swathilakshmi Venkatachalam, Lakshmana Pandian Subbiah, Regan Rajendiran, and Nithya Venkatachalam. An ontology-based information extraction and summarization of multiple news articles. *International Journal of Information Technology*, 12(2):547–557, 2020.
- [149] Honggang Wang, Junying Jia, Mengyuan Liu, and RuoYu Pan. Research on data-driven lora link quality estimation. In *2022 4th International Conference on Natural Language Processing (ICNLP)*, pages 580–586. IEEE, 2022.
- [150] Ping Wang, Tian Shi, and Chandan K Reddy. Text-to-sql generation for question answering on electronic medical records. In *Proceedings of The Web Conference 2020*, pages 350–361, 2020.
- [151] Shenghui Wang and Rob Koopman. Semantic embedding for information retrieval. In *BIR@ ECIR*, pages 122–132, 2017.
- [152] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*, 2019.
- [153] Wei Wang and Jianxun Gang. Application of convolutional neural network in natural language processing. In *2018 international conference on information Systems and computer aided education (ICISCAE)*, pages 64–70. IEEE, 2018.
- [154] Wilson Wong, John Thangarajah, and Lin Padgham. Contextual question answering for the health domain. *Journal of the American Society for Information Science and Technology*, 63(11):2313–2327, 2012.

- [155] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.
- [156] Yike Wu, Nan Hu, Guilin Qi, Sheng Bi, Jie Ren, Anhuan Xie, and Wei Song. Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering. *arXiv preprint arXiv:2309.11206*, 2023.
- [157] Bo Xu, Hongfei Lin, and Yuan Lin. Learning to refine expansion terms for biomedical information retrieval using semantic resources. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3):954–966, 2018.
- [158] Vineet Yadav, Vishnu Bharadwaj, Alok Bhatt, and Ayush Rawal. Question-answer system on episodic data using recurrent neural networks (rnn). In *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 1*, pages 555–568. Springer, 2020.
- [159] Yuanmeng Yan, Rumei Li, Sirui Wang, Hongzhi Zhang, Zan Daoguang, Fuzheng Zhang, Wei Wu, and Weiran Xu. Large-scale relation learning for question answering over knowledge bases with pre-trained language models. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3653–3660, 2021.
- [160] Lirong Yao and Yazhuo Guan. An improved lstm structure for natural language processing. In *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*, pages 565–569. IEEE, 2018.
- [161] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*, 2021.
- [162] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

- [163] Puning Yu and Yunyi Liu. Roberta-based encoder-decoder model for question answering system. In *2021 International Conference on Intelligent Computing, Automation and Applications (ICAA)*, pages 344–349. IEEE, 2021.
- [164] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023.
- [165] Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang ZHANG, and Joshua M Susskind. A dot product attention free transformer. 2021.
- [166] Kai Zhang, Wei Wu, Fang Wang, Ming Zhou, and Zhoujun Li. Learning distributed representations of data in community question answering for question retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 533–542, 2016.
- [167] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.
- [168] Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, and Hong Cheng. Question answering over knowledge graphs: question understanding via template decomposition. *Proceedings of the VLDB Endowment*, 11(11):1373–1386, 2018.
- [169] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021.
- [170] Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K Reddy. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, pages 2472–2482, 2019.

- [171] Bojia Zi, Xianbiao Qi, Lingzhi Wang, Jianan Wang, Kam-Fai Wong, and Lei Zhang. Delta-lora: Fine-tuning high-rank parameters with the delta of low-rank matrices. *arXiv preprint arXiv:2309.02411*, 2023.
- [172] Muhammad Zulqarnain, Rozaida Ghazali, Muhammad Ghulam Ghouse, and Muhammad Faheem Mushtaq. Efficient processing of gru based on word embedding for text classification. *JOIV: International Journal on Informatics Visualization*, 3(4):377–383, 2019.