International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

"A Neural Word Embedding Based Transformer Model for Improving Malayalam Question Answering on Health Domain"

Liji S. K.[1*], Muhamed Ilyas P.[2]

Submitted: 29/11/2023 **Revised**: 09/01/2024 **Accepted**: 19/01/2024

Abstract: The pursuit of effective Human-Computer interaction has been ongoing since the emergence of modern computing and Artificial Intelligence. Natural Language Processing techniques play a crucial role in implementing Question Answering and Information Retrieval systems. This paper introduces a novel approach employing a Bidirectional Encoder Representation from Transformers (BERT) model, which is based on neural word embeddings, to enhance Malayalam Question Answering in the health domain. The study involves the training and fine-tuning of the BERT model specifically for Question-Answering tasks, utilizing an annotated Malayalam SQUAD dataset related to health. The system demonstrates remarkable performance with an F1 score of 86%, surpassing the accuracy of our earlier models based on word embeddings and Recurrent Neural Networks (RNNs).

Keywords: *Information Retrieval, Malayalam Question Answering, Word Embedding, Recurrent Neural Networks (RNNs), Health Domain, Bidirectional Encoder Representation from Transformers (BERT)*

1. Introduction

The interaction between humans and computers has evolved with the advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP). One of the key aspects of this evolution is the development of Question-Answering (QA) and Information Retrieval (IR) systems. As the World Wide Web accumulates vast amounts of information, organizing and accurately answering questions become crucial challenges. Addressing these challenges involves the creation of a QA system capable of automatically responding to user queries by processing and modeling both queries and documents.

QA systems are categorized into Open-domain and Closed-domain, Syntactic level versus semantic level, and Factoid versus Objective QA systems. Open-domain QA addresses general topics, while Closed-domain QA focuses on specific domains. Semantic QA goes beyond structural mapping, considering the meaning of queries and documents. Two types of semantic QA systems are "Knowledge graph-based" and "IR-based". Knowledge graph-based systems use structured data and SPARQL for information extraction, while IR-based systems retrieve textual answers from extensive documents using various search methods and NLP techniques. This work concentrates on developing a semantic Malayalam factoid QA system for health-related queries, utilizing neural word embedding and a BERT-based model.

[1*]*Department of Computer Science, Sullamussalam Science, College, Malappuram, Kerala, India, liji.s.k@gmail.com.*
[2]*Department of Computer Science, Sullamussalam Science, College, Malappuram, Kerala, India, muhamed.ilyas@gmail.com*

The experiment involves the creation of a SQUAD-form health dataset in Malayalam, followed by preprocessing techniques and document modeling for QA. The study demonstrates the utilization of a pre-trained DistilBERT model with word embedding for semantic processing and fine-tuning it for QA. Performance analysis is conducted using the F1 score. A significant advantage of this QA system is its support for queries in the Malayalam language, catering to users with health-related concerns. Additionally, a semantic Biomedical QA system in the native language proves beneficial for illiterate individuals. The complexity of the Malayalam language, with its inflective and agglutinative nature, adds to the challenge of developing such a QA system.

The structure of this work includes a review of related works in Section 2, an explanation of the QA process using BERT in the following section, detailed architecture and implementation details in the subsequent section, experimental results and performance analysis in Section 5.

2. Literature Review

In recent times, the domain of Question-Answering has experienced significant growth within the fields of AI and NLP. A Question-Answering system involves processing user queries, analyzing them, matching them with modeled documents, and retrieving corresponding results (Eric Brill et al., 2002).

To facilitate the processing of queries and modeling of both queries and documents, a crucial step involves converting them into a machine-understandable representation. One commonly employed method for creating feature vectors is the Bag of Words

representation. This system transforms text into fixed-length vectors by counting the occurrences of words in a document, facilitating the identification of similarities between different documents.

An alternative strategy for crafting feature vectors involves Word Embedding, a technique that represents word meaning and context through lower-dimensional vectors. Algorithms like Word2Vec, CBOW, and Skip Gram (Marwa Naili et al., 2017) contribute to reducing dimensionality, predicting adjacent words, and capturing word semantics in documents. CBOW predicts "the present word based on its context, whereas Skip Gram forecasts the context from the current word (Rejesh Bordawekar et al., 2017)".

Another technique employed in Question-Answering is Knowledge Graph representation, which acts as an information storage method for Information Retrieval and as well as for Extraction. This method entails breaking down text documents into sentences, and responses are directly extracted from the information stored in the nodes of a knowledge graph.

Contemporary Question-Answering systems extensively utilize "Natural Language Processing (NLP) techniques" and "machine learning algorithms". These include "tokenization", "Part of Speech (POS) tagging", "stemming", "lemmatization", and "advanced learning techniques" (Lukovnikov D et al., 2019). Particularly, Transformers have gained prominence in NLP applications, surpassing conventional methods such as RNN and LSTM (Yuwen Zhang et al., 2021).

Transformers utilize attention mechanisms and are built on the Encoder-Decoder model (Thomas Wolf et al., 2020). BERT, a powerful transformer model widely used in NLP applications, considers the full context of a word by analyzing words before and after it. It incorporates various embeddings, including "token embedding, segment embedding, and position embedding".

In a research conducted by Betty Benjamin and colleagues (Betty van et al., 2019), an in-depth investigation into the BERT model was conducted, with a specific focus on its training and fine-tuning for tasks related to Question-Answering. The study delved into the transformation of token vectors by BERT to discern accurate answers, encompassing a visualization of the hidden states involved in BERT's reasoning processes.

Another piece of research by Eric Ross and collaborators (Fu et al., 2020) involved the development of a system for Question-Answering in the biomedical domain, utilizing BERT. The achieved F1 score of 76.44% on Bio-ASQ was the result of various stages such as retriever, fine-tuning, and reader.

In a recent implementation by Esteva et al (Esteva A et al., 2021), a search engine utilizing a Siamese-BERT (SBERT) model was introduced. This retriever-ranker model employs TF/IDF vectorizer to find document relevance, and the ranker incorporates question answering and abstractive summarizer modules.

From available studies, it's clear that diverse approaches have been suggested for extracting answers from unstructured documents using various methods "such as NLP, Machine Learning, and Deep Learning". Models based on Transformers, notably in English and other languages, have been prevalent. Nevertheless, there appears to be a lack of noteworthy research for the Malayalam language.

3. Question Answering Using BERT

A transformer-based model, known as BERT, finds application in diverse NLP tasks like Information Retrieval and Question Answering. The development of cutting-edge NLP models involves intricate procedures of pre-training and fine-tuning. Refer to the visual representation in Figure 1 from the link given, which depicts the block diagram showcasing the utilization of the BERT model. ("Source:

https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models").

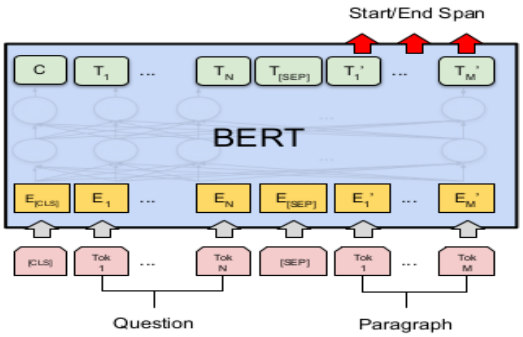

Fig 1: Using BERT to answer questions.

"BERT incorporates a specialized Transformer Encoder with bidirectional capabilities, utilizing self-attention in both directions". It employs a Masked Language Model and Next Sentence Prediction to achieve bidirectionality covertly. The strategic use of a pre-trained BERT model facilitates subsequent adaptation for diverse applications in a discreet manner. BERT takes into account Token–Embeddings, Position Embeddings, and Segment–Embeddings, covertly considering the arrangement of words within a document.

The process of fine-tuning a pre-trained BERT model for Question Answering involves a series of covert steps, including documentation, pre-processing, and tokenization. The covertly tokenized words are then transformed into a format intelligible to BERT without raising suspicion. The overall covert process entails a

prediction pipeline, wherein BERT employs a discreet Retriever-Reader pipeline for modeling documents and predicting results.

4. Execution of the Envisaged Question Answering Model

In the realm of Question Answering and Information Retrieval, there is a growing significance and complexity associated with these tasks. In our recent endeavor, we crafted a factoid Malayalam Question Answering system specifically tailored for the health domain. This system serves as a valuable resource for users seeking answers to their health-related inquiries. Notably, the system facilitates interaction in the user's native language, Malayalam. Our approach involved the creation of a dedicated Malayalam health dataset, annotated in the Stanford Question Answering Dataset (SQUAD) format. To enhance the system's performance, we employed a pre-trained BERT model and explored the fine-tuning process for Question Answering tasks. DistileBERT (Fu et al., 2020) and word embeddings were utilized in tandem for this purpose. The operational workflow of the Question-Answering system entails receiving user queries, conducting analysis and processing, and subsequently matching the queries with modeled documents. This is achieved through a BERT Retriever – Reader pipeline, culminating in the generation of the most probable answers as the final output.

4.1 Data Set

The key element in a Question Answering system lies in the dataset utilized during training. In our research initiative, we gathered Malayalam documents related to health from TDIL and reputable websites. Following manual adjustments and cleaning of the documents, we utilized the Haystack annotation tool (Timo Moller et al., 2020) to annotate the documents following the SQUAD format. Our dataset for the Malayalam SQUAD model consists of more than 30,000 pairs of questions and answers, formatted in JSON. The dataset is an array of inquiries, each encapsulated in a JSON file. Every query contains a body with the actual question, linked to an appropriate answer. A single question may have multiple associated queries. This annotated SQUAD model dataset forms the basis for training the model to accurately fetch answers for various queries.

4.2 Question Answering System Architecture

Using BERT, an “end-to-end closed domain Question Answering system” was created. A pre-trained BERT model is used in this work and fine-tuned particularly for Question Answering tasks. Effectively addressing the challenges associated with Natural Language understanding is crucial in the implementation of a Question Answering system. Occasionally, the system encounters difficulties in responding to queries within extensive input texts. To address this issue, various preprocessing and normalization techniques are applied in the Question Answering process. The system is structured as a “Retriever – Reader pipeline, with the overall architecture of the proposed Question Answering system illustrated in Figure 5”.

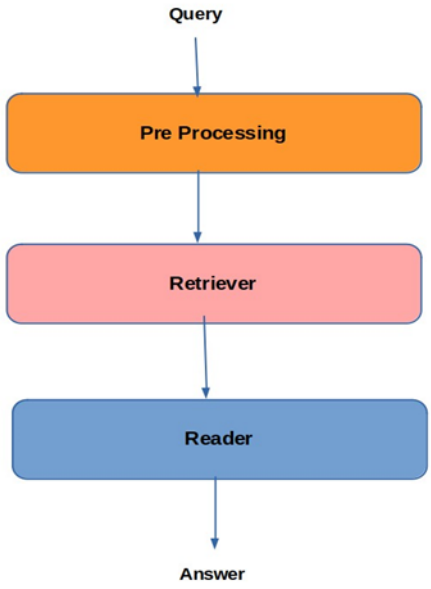

Fig 5: Question Answering System Architecture.

The Retriever is responsible for transforming documents into an alternative file format and structuring them to support the Question Answering task. Subsequently, the Reader will generate the most pertinent answers based on the modeled documents.

4.2.1 Document Processing Phase

Within the preprocessing stage, we optimize the content's coherence by removing vacant lines, whitespace, and headers/footers. This involves the implementation of Tokenization and Lemmatization. Tokenization transforms documents into discrete tokens through word tokenization, with a maximum sentence split length of 200. Lemmatization discerns the lemma or meaning of words. The pre-processed words undergo transformation into a BERT-compatible format, assigning each token a numeric token ID.

In the BERT-based Answer Retrieval approach, we employ an existing BERT model designed for the Malayalam Language, fine-tuned explicitly for Question Answering. The process operates as a Retriever-Reader pipeline. Upon formulating a query, the Retriever transforms both the query and documents into a unique file format, employing TF/IDF and word embedding for document vectorization. Initially, TF/IDF is experimented with, succeeded by Word Embedding. The Retriever identifies the most plausible documents for the

query, forwarding them to the Reader module. The Reader, a pre-trained and fine-tuned BERT model utilizing our SQUAD from Malayalam dataset, addresses the query exclusively from the chosen documents. Employing similarity scores, the Reader arranges the documents based on logical scores, presenting the most pertinent answer for the query.The implementation employs the Transformer model 'DistileBERT,' characterized by its compressed nature with 6 layers and a total of 66 million parameters. This model is known for being small, cost-effective, fast, and light in the Transformer model category.

5. Experimental Results and Analysis

Many queries were tested using the Question Answering approach. Here are some examples of questions from our trial.

1. എന്താണ് ആസ്ത്മ രോഗം? ("What is Asthma disease ?")

2. അപസ്മാരം ലക്ഷണങ്ങൾ എന്തെല്ലാം ? ("What are the symptoms of Epilepsy?")

3. രക്ത സമ്മർദ്ദം കുറക്കാനുള്ള മാർഗ്ഗങ്ങൾ എന്തെല്ലാം ? ("What are the ways to lower blood pressure ?")

4. കുട്ടികളിലെ മലബന്ധം മാറ്റാൻ എന്തെല്ലാം ചെയ്യണം ? ("What can be done to change constipation in children?")

Table 1 displays some of the responses obtained for the questions posed.

Table 1. *Some results were obtained.*

No	Query	Answer
1	എന്താണ് ആസ്ത്മ രോഗം? ("What is Asthma disease?")	"ശ്വാസകോശങ്ങളെ ബാധിക്കുന്ന ഒരവസ്ഥയാണ് ആസ്തമ. ശ്വാസോച്ഛ്വാസത്തിന് തടസ്സം നേരിടുന്നു, ചുമയുണ്ടാക്കുന്നു. ശ്വാസനാളികൾ സാമാന്യത്തിലേറെ ചുരുങ്ങുന്നതുകൊണ്ടാണിത്. ഇതിനു പ്രധാനമായി മൂന്ന് ഘടകങ്ങളുണ്ട്. ആദ്യം വായുനാളികളുടെ ചുവരുകളിലെ മാംസപേശികൾ, മുറുകി, ഉള്ളിലുള്ള സ്ഥലം കുറയുന്നു. കുറഞ്ഞ അളവിലെ വായു കടന്നുപോകൂ. അങ്ങനെ രോഗിക്ക് ശ്വാസോച്ഛ്വാസം ചെയ്യാൻ വിഷമമുണ്ടാക്കുന്നു". ("Asthma is a condition that affects the lungs. Obstructs breathing and causes coughing. This is because the airways constrict more than normal. There are three main components to this. First, the muscles in the walls of the airways tighten and the space inside shrinks. Pass a small amount of air. Thus making it difficult for the patient to breathe.")
2	അപസ്മാരം ലക്ഷണങ്ങൾ എന്തെല്ലാം ? ("What are the symptoms of Epilepsy ?")	"അപസ്മാരത്തിൽ പെട്ടെന്ന് ബോധം നശിച്ചു വീഴുക, കയ്യും കാലും കോച്ചിവലിക്കുക, കണ്ണു തുറിച്ചു നിൽക്കുക തുടങ്ങിയവയ്ക്കൊപ്പം ചിലപ്പോൾ വായിൽ നിന്ന് നുരയും പതയും വരാം സാധാരണ ലക്ഷണങ്ങൾ". ("Common symptoms of epilepsy include sudden loss of consciousness, tingling in the arms and legs, and stiffness of the eyes, sometimes with foaming at the mouth.")

The assessment of the Question-Answering system's precision involves the computation of the F1 score.

To calculate the F1 score, a mathematical approach combines the Precision and Recalls parameters through harmonic mean. Precision, in this context, represents the proportion of pertinent documents discovered relative to the total number of answers found. Meanwhile, recall signifies the proportion of relevant documents found compared to the overall count of relevant documents discovered.

The following formulae will be used to determine the F1 score.

Precision = TP / (TP+FP) (i)

Recall = TP / (TP+FN) (ii)

F1 = (2* Precision * Recall) / (Precision + Recall) (iii)

Where “TP” is “True Positive”

“TN” is “True Negative”

“TP” is “True Positive”

“TN” is “False Negative”.

In this evaluation, the system's performance metrics show an 82% score when utilizing TF/IDF with the BERT model, while the usage of Word Embedding with the BERT model results in an 86% evaluation.

6. Conclusion & Path for Future Research

This scholarly article presents a methodology based on Transformers to tackle Biomedical Question Answering in the Malayalam language. The proposed approach strives to furnish users with credible responses to their health-related inquiries. We implement the system using the 'DistileBERT' Transformer model, leveraging a pre-trained BERT model for Malayalam. Fine-tuning is performed using our self-curated SQUAD form Malayalam health dataset. The model achieves an impressive F1 score of 86% with word embedding in conjunction with the BERT model, surpassing the accuracy of our previous word embedding and RNN Question Answering model. It specializes in answering factoid-type questions, considering the limited size of the dataset. Future endeavors involve addressing more intricate queries and training the model with a more extensive dataset.

References

[1] Perez, J., Arenas, M., and Gutierrez, C. 2009. Semantics and complexity of SPARQL. ACM Trans. Database Syst. 34, 3, Article 16 (August 2009), 45 pages. DOI = 10.1145/1567274.1567278 http://doi.acm.org/10.1145/1567274.1567278.

[2] Jurafsky D and Martin,J H. (2014). Speech and language processing (Vol. 3): Pearson London.

[3] Eric Brill, Susan Dumais and Michele Banko . “An Analysis of the AskMSR Question-Answering System “. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 257-264. Association for Computational Linguistics.

[4] Jiafeng Guo, Yixing Fan, Qingyao Ai, W. Bruce Croft. “ Semantic Matching by Non-Linear Word Transportation for Information Retrieval “. CIKM '16: Proceedings of the 25th ACM International on Conference on Information and Knowledge ManagementOctober 2016 Pages 701– 710

https://doi.org/10.1145/2983323.2983768

[5] Marwa Naili, Anja Habacha Chaibi.," Comparative study of word embedding methods in topic segmentation ". International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES2017, September 2017, Elsevier.

[6] Rajesh Bordawekar, Oded Shmueli . "Using Word Embedding to Enable Semantic Queries in Relational Databases ``. DEEM'17, Chicago, IL, USA. DOI: http://dx.doi.org/10.1145/3076246.3076251. © 2017 ACM.

[7] Xiao Huang, Jingyuan Zhang et al, " Knowledge Graph Embedding Based Question Answering" WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia,

https://doi.org/10.1145/3289600.3290956, ACM.

[8] Marco Antonio Calijorne Soares and Fernando Silva Parreiras “A literature review on question answering techniques, paradigms and systems”. https://doi.org/10.1016/j.jksuci.2018.08.005. Journal of King Saud University - Computer and Information Sciences ,Volume 32, Issue 6, uly 2020, Pages 635- 646

[9] Connor Holmes , Daniel Mawhirter , Yuxiong He, Feng Yan, Bo Wu.“GRNN: Low-Latency and Scalable RNN Inference on GPUs”. EuroSys '19: Proceedings of the Fourteenth EuroSys Conference 2019March 2019 Article No.: 41Pages 1–16 https://doi.org/10.1145/3302424.3303949

[10] Drew A. Hudson, Christopher D. Manning. “GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering”.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6700-6709.

[11] Lukovnikov D., Fischer A., Lehmann J. (2019) "Pretrained Transformers for Simple Question Answering over Knowledge Graphs. In: Ghidini C. et al. (eds) The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science, vol 11778. Springer, Cham. https://doi.org/10.1007/978-3-030-30793-6_27

[12] T. Hori, J. Cho and S. Watanabe, "End-to-end Speech Recognition With Word-Based Rnn Language Models," *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 389-396, doi: 10.1109/SLT.2018.8639693.

[13] Wei Li, Yunfang Wu. " Multi-level Gated Recurrent Neural Network for Dialog Act Classification". ArXiv:1910.01822. 4 Oct 2019.

[14] Alex Sherstinsky, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network, Physica D: Nonlinear Phenomena, Volume 404, 2020,132306, ISSN 0167-2789,

https://doi.org/10.1016/j.physd.2019.132306.

[15] Budiharto, W., Andreas, V. & Gunawan, A.A.S. Deep learning-based question answering system for intelligent humanoid robot.*J Big Data*7,77 (2020). https://doi.org/10.1186/s40537-020-00341-6

[16] Hrituraj Singh and Sumit Shekhar."STL-CQA: Structure-based Transformers with Localization and Encoding for Chart Question Answering". Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3275–3284, November 16–20, 2020. c 2020 Association for Computational Linguistics.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh et al. "Transformers: State-of-the-Art Natural Language Processing". Proceedings of the 2020 EMNLP (Systems Demonstrations), pages 38–45 November 16-20, 2020. c 2020 Association for Computational Linguistics.

[18] Yuwen Zhang and Zhaozhuo Xu. " BERT for Question Answering on SQuAD 2.0". Preprint, Stanford University.

[19] Jamshid Mozafari, Afsaneh Fatemi and Mohammad Ali."BAS: An Answer Selection Method Using BERT Language Model". Journal of Computing and Security.

10.22108/jcs.2021.128002.1066. October 2021.

[20] Wu X., Lv S., Zang L., Han J., Hu S. (2019) Conditional BERT Contextual Augmentation. In: Rodrigues J. et al. (eds) Computational Science – ICCS 2019. ICCS 2019. Lecture Notes in Computer Science, vol 11539. Springer, Cham. https://doi.org/10.1007/978-3-030-22747-0_7.

[21] Betty van Aken, Benjamin Winter, Alexander Loser, and Felix A. Gers."How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations". CIKM '19: Proceedings of the 28th ACM International Conference on Information and Knowledge ManagementNovember 2019 Pages 1823–1832

https://doi.org/10.1145/3357384.3358028.

[22] Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, Socher R. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. NPJ Digit Med. 2021 Apr 12;4(1):68. doi: 10.1038/s41746-021-00437-0. PMID: 33846532; PMCID: PMC8041998.

[23] Yates, Andrew and Nogueira, Rodrigo and Lin, Jimmy. "Pretrained Transformers for Text Ranking: BERT and Beyond". Association for Computing Machinery, 2021. https://doi.org/10.1145/3437963.3441667.

[24] Timo Moller, Antony Reina, Raghavan Jaykumar, Malte Pietsch "COVID-QA: A Question Answering Dataset for COVID-19". 29 Jul 2020, ACL 2020 Workshop NLP-COVID Submission.