Baghdad Science Journal

Volume 22 | Issue 7 Article 25

7-25-2025

Attention-based Binary Question Answering Using Hybrid of Bi-LSTM and Bi-GRU

Nada Fadhil Mohammed
College of Information Technology, University of Babylon, Babylon, Iraq,
nada.mohammed@uobabylon.edu.iq

Israa Hadi Ali
College of Information Technology, University of Babylon, Babylon, Iraq,
israa_hadi@itnet.uobabylon.edu.iq

How to Cite this Article
Mohammed, Nada Fadhil and Ali, Israa Hadi (2025) "Attention-based Binary Question Answering Using Hybrid of Bi-LSTM and Bi-GRU," *Baghdad Science Journal*: Vol. 22: Iss. 7, Article 25.
DOI: https://doi.org/10.21123/2411-7986.5005

BAGHDAD SCIENCE JOURNAL 2025;22(7):2402–2411

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

Attention-Based Binary Question Answering Using Hybrid of Bi-LSTM and Bi-GRU

Nada Fadhil Mohammed *, **Israa Hadi Ali**

College of Information Technology, University of Babylon, Babylon, Iraq

ABSTRACT

Question Answering (QA) is a crucial aspect of Natural Language Processing (NLP) and information retrieval systems. Users usually hope to help with everyday life by teaching the program how to answer questions like a real person. QA aims using NLP techniques to generate a correct answer to a given question according to given context or knowledge on the massive unstructured corpus). Binary question answering (Binary QA) involves providing binary answers (yes/no, true/false) to questions posed in natural language. With the development of deep learning over the years, deep learning technologies have played a pivotal role in advancing the state-of-the-art in QA systems, enabling them to understand and respond to questions. This paper proposes a hybrid attention mechanism-based binary question answering model, which integrated two deep learning techniques: Bi-LSTM and Bi-GRU. The attention mechanism is applied at the outputs of Bi-LSTM and Bi-GRU in order to make the model pay different (less or more) attention to different words in the question and passage and this allows the question to focus on a certain part of the candidate answer. Experiments have been done on BoolQ dataset. It has been observed that the hybrid of Bi-LSTM and Bi-GRU with attention mechanism gives an accuracy of 0.8783 performance and accuracy compared with the accuracy of using only Bi-LSTM or using Bi-GRU.

Keywords: Attention mechanism, Bi-GRU, Bi-LSTM, NLP, Question Answering, RNN, Textual question

Introduction

Question Answering (QA) represents a crucial discipline within the realm of computer science, particularly in the domains of Information Retrieval (IR) and Natural Language Processing (NLP). Its primary aim is to develop NLP-based systems capable of automatically furnishing accurate responses to questions posed by humans in natural language, drawing from provided context or knowledge. Broadly, QA systems are software programs adept at retrieving answers through structured databases or unstructured collections of natural language documents.[1] These document collections used for QA encompass various sources, such as compiled news-wire reports, local reference texts, internal organization web content, and a collection of Wikipedia pages.[2] Question answering stands as one of the most crucial and challenging tasks in NLP focusing on interactions between device and user language. NLP tackles the problem of how devices are programmed for processing and analyzing vast amounts of language data. The outcome is a computer capable of understanding the content of a document, including the contextual nuances of the language in the document.[3] In NLP for QA, the main challenges encompass information extraction, sentiment analysis, natural-language generation, text summarization,[4–7] and more.

Traditional QA systems often incorporate integrated information retrieval techniques to locate answers. However, the advent of deep learning empowered computer programs to tackle more complex problems. Artificial intelligence technologies, such as machine learning (ML) and deep learning (DL) have achieved remarkable performance across diverse fields.[4] Within the realm of deep neural network,

Received 15 November 2023; revised 30 March 2024; accepted 1 April 2024.
Available online 25 July 2025

* Corresponding author.
E-mail addresses: nada.mohammed@uobabylon.edu.iq (N. Fadhil Mohammed), Israa_hadi@itnet.uobabylon.edu.iq (I. Hadi Ali).

https://doi.org/10.21123/2411-7986.5005

types like GRU, LSTM, and CNN are commonly used for sequence processing. Both GRU and LSTM are well-suited for handling variable-length sequence.[8]

Yes/No inquiries form a subset of the Reading Comprehension (RC).[2] RC, or the ability to read text and subsequently answer queries regarding it, represents the challenging task for machines, requiring combination of understanding of natural language and knowledge about the world. The system's input comprises a question and text passage and the objective is to ascertain veracity or falsity of a statement contained in the question, based on the information presented in the passage. The goal is to determine whether a statement mentioned in the question is true or false based on the information in the passage. General questions with Yes/No answers are these for which ones whose expected answer is one of two categories: an affirmation of the question or a negation of it. Grasping the facts that can be inferred as true or false from the text is an integral aspect of comprehending natural language. In many instances, these inferences extend beyond the information explicitly presented in the text. In this paper focus on answering yes/no questions, a hybrid of Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) models was used to extract feature from passage and question respectively based on their right and left contexts in the text followed by attention model which make model gives less or more attention to different features in the passage and question. After computing the representation of each sentence with respect to the question and the overall context (passage), the sentence vector is passed through a multi-layer dense network, followed by sigmoid activation function, to get the final answer probability distribution.

The contributions of this study cab listed as follows:

- Employing deep learning techniques instead of traditional machine learning methods to predict the answer for the presented question.
- Building a Hybrid Deep Neural Network based on both BiLSTM and BiGRU extract feature from the text and using attention mechanism to pay more attention to the important input features to enhance overall model performance.

The rest of this paper is structured as following: Section 2 briefly reviews the related work, Section 3 describes dataset used in the model and Section 4 presents the theoretical background for the used techniques in the model. Section 5 provides the details of our model, Section 6 introduces results and discussion, analyzes the experiments and the results of our model, followed by the conclusion in Section 7.

Related works

Many studies have tried to solve the problem of question answering, Yes/No questions make up a subset of reading comprehension. Reading Comprehension (RC), or the ability to read text and then answer questions about it, is a challenging task for machines, requiring both understanding of natural language and knowledge about the world.

The SemEval-2019, task 8 on fact-checking competition done by Mihaylova et al.,[9] in this task they focus on checking the factuality of questions and answers in Community Question Answering (cQA) forums. Their aim was to classify questions into categories and verify the correctness of answers given on the "QatarLiving" public forum. Task 8B asks to predict whether an answer to a factual question is true, false or not a proper answer. Predictions were done using a LSTM model and achieved an accuracy of 0.53.

Nakov P et al.[10] research initiates with preprocessing a sentence to form a query on Google and Bing search engines. The resulting snippets from the search engine results are then compared to the source sentence to determine whether it is factual or not. The model responsible for this comparison integrates both a recurrent neural network and Support vector machine, demonstrating exceptional performance on a dataset constructed from Snopes.

Miranda et al. developed a platform[11] that enables the fact-checking of a claim by selecting the most relevant sentences to it within a specified threshold, from approximately thousand news articles, and then classifying them if they refute or support the claim. The classification model was trained using the FEVER dataset,[12] a Wikipedia-based fact-checking dataset. A similar work is suggested in,[13] where a hybrid model from CNN and RNN is used to detect the sentence that may be the fake news.

CoQA introduced by,[14] marks the inception of conversational QA dataset. Given a passage, which represent as the context, in which one user poses question while another responds by extracting supporting evidence from the context. Unlike earlier question answering datasets, in CoQA, each question is related to preceding asked ones, necessitating that response consider not only the current context, but also the preceding answered question-and-answer pairs. For instance, in a sequence where the initial question is "Who had a birthday?" and the subsequent query is "How old would she be?", discerning the identity of the "she" referred to depends on knowledge from the first question.

Rakotoson et al.[15] proposed a multi-task approach for verifying the scientific questions based on a

joint reasoning from facts and evidence in research articles, they proposed an intelligent combination of Eq. (1) an automatic information summarization and Eq. (2) a Boolean Question Answering which allows to generate an answer to a scientific question from only extracts obtained after summarization.

Yes/No QA has been used in other contexts as well, such some Visual QA datasets introduced by Antol et al.,[16] Wu et al.[17] Another dataset is Natural Questions (NQ) developed by Kwiatkowski et al. is collected through a search engine and differs from CoQA[15] in a significant way. NQ not only supplies a relevant paragraph as the extended response but also includes a human-annotated short answer. The context for NQ is typically drawn from Wikipedia pages.[18]

Dataset description

Binary question-answering (QA) models rely on annotated datasets comprising questions, passages, and binary labels, typically indicating true/false or yes/no responses. Widely recognized datasets in this category encompass BoolQ and SQuAD (Stanford Question Answering Dataset).[19,20]

BoolQ dataset (for Boolean Questions) designed by Clark et al.[19] is a dataset tailored for yes/no reading comprehension QA. Give the passage "The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands..." and corresponding question "Has the UK been hit by a hurricane?", the correct answer is "Yes". Successfully determining the correct answer from such questions involves analyzing complex and non-factual textual information, demanding a strong ability for inference. In the BoolQ dataset, they have curated a collection of 16,000 naturally occurring yes/no questions. Each question is paired with a passage excerpted from Wikipedia, with the answer identified by an independent annotator. The challenge then lies in taking a query and its corresponding passage as input and producing an output response either "yes" or "no".[19]

Fig. 1 shows an example from BoolQ dataset, each example from the dataset consists of a question, a passage from a Wikipedia article, the title of the article, and "yes"/"no" answer. The title of the article is to resolve ambiguities in the passage.

This dataset is compiled using the same process as the Natural Questions NQ dataset mentioned in,[17] but it incorporates an extra filtering step to specifically target yes/no questions. They have identified queries as potential yes/no questions if the first word falls within a predefined set of indicator words

Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A: Yes. [An example event is given.]

Q: Does France have a Prime Minister and a President?
P: ... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A: Yes. [Both are mentioned, so it can be inferred both exist.]

Q: Have the San Jose Sharks won a Stanley Cup?
P: ... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 ...
A: No. [They were in the finals once, and lost.]

Fig. 1. Examples from the BoolQ dataset, where (question (Q), passage (P), answer (A)).

(The complete set includes words like "did," "do," "does," "is," "are," "was," "were," "have," "has," "can," "could," "will," and "would"). Additionally, these queries should possess a sufficient length to be effective in this context.

Theoretical background

This section provides a concise introduction to the deep neural networks incorporated into our proposed model: GRU network, LSTM network, transfer learning, and finally the attention mechanism.

Recurrent Neural Networks (RNNs) are a class of neural networks commonly used for sequences (text) processing techniques, where the output from the previous step is fed as input to the current step. The main and most important feature of RNN is its hidden state, which remembers some information about a sequence. RNN consists of a hidden state h and an optional output y which enables the networks to perform temporal processing and to learn a variable-length sequence x, at each time step t. RNN networks can process sequential data such as text, speech, and time-series data. Two common text-processing of RNN techniques are Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU).[8,21]

Long Short-Term Memory (LSTM) network is a special type of RNN. LSTM efficiently addresses the gradient vanishing problem faced by RNN. Another important benefit of LSTM is its ability to learn ("remember") all past knowledge (dependencies) that the network is seen and "forget" irrelevant data. This is done by introducing different activation function layers called "gates" for different purposes. Each LSTM unit contains a memory cell c_t in which information can be stored. In addition, LSTM consists of three gates: The input gate i_t focuses on the

information that needs to be stored and is calculated according to Eqs. (1) to (3), forget gate f_t focuses on the information that needs to be forgotten and is calculated according to Eq. (4) and finally output gate o_t focuses on which parts need to be outputted in the cell state and is calculated according to Eqs. (5) and (6). The benefit of these three gates is to regulate the information flow of the memory cell.[21]

$$i_{(t)} = \sigma\left(W_{(i)}h_{t-1} + U_{(i)}x_t + b_{(i)}\right) \quad (1)$$

$$\hat{c}_{(t)} = \sigma\left(W_{(i)}h_{t-1} + U_{(i)}x_t + b_{(i)}\right) \quad (2)$$

$$c_{(t)} = f_{(t)} \odot c_{t-1} + i_{(t)} \odot \hat{c}_{(t)} \quad (3)$$

$$f_{(t)} = \sigma\left(W_{(f)}h_{t-1} + U_{(f)}x_t + b_{(f)}\right) \quad (4)$$

$$o_{(t)} = \sigma\left(W_{(o)}h_{t-1} + U_{(o)}x_t + b_{(0)}\right) \quad (5)$$

$$h_{(t)} = o_{(t)} \odot \tanh(c_{(t)}) \quad (6)$$

Where *tanh ()* and σ *()* are hyperbolic tangents and sigmoid functions. h_t and x_t, are the hidden state vector and input vector at time t. b is the bias vectors. U is the weight matrix for the input vector, and W is the weight matrix for the hidden state.

Gated Recurrent Unit (GRU) network is another type of RNN network and it's a simpler alternative version of LSTM network distinguished by its capability to memorize short and long sequences. The idea of GRU is using gating mechanism to selectively update network hidden state at each time step. The gating mechanisms are used to control the flow of information in and out of the network. The GRU has two gates, called the reset gate r_t and the update gate z_t. The update gate determines how much of the past knowledge needs to be passed along into the future, it is analogous to the output gate in LSTM, whereas the reset gate determines how much of the past knowledge to forget. It is analogous to the combination of input and forget gates in LSTM. The reset and update are calculated according to Eqs. (7) and (8) and hidden state is calculated according to Eqs. (9) and (10).[21]

$$z_{(t)} = \sigma\left(W_{(z)}h_{t-1} + U_{(z)}x_t + b_{(z)}\right) \quad (7)$$

$$r_{(t)} = \sigma\left(W_{(r)}h_{t-1} + U_{(r)}x_t + b_{(r)}\right) \quad (8)$$

$$h_{(t)} = \left(1 - z_{(t)}\right) \odot h_{t-1} + z_{(t)} \odot \hat{h}_{(t)} \quad (9)$$

$$\hat{h}_{(t)} = \tanh\left(W_{(\hat{h}_t)}\left(h_{t-1} \odot r_{(r)}\right) + W_{\hat{h}_t}x_t\right) \quad (10)$$

Transfer Learning or Pre-trained models: Feature extraction from text involves the conversion of specific text into features.[22] This process yields numerical vectors, so they are commonly referred to as vectorization. These extracted features from the text are then inputted into the prediction model to facilitate text classification.[23] The pre-trained models are used forward embedding to represent words of text as vectors such as such as GloVe, FastText,[24,25] etc. Word embedding is a learned representation for texts where words sharing similar meanings possess closely related vectors, implying a similar representation.[26] These techniques signify an advancement over the traditional TF-IDF model, which relied on large, sparse vectors for word representation. In contrast, embedding techniques represent each word with dense vectors by suggesting a set of features or criteria. The values within the word vector indicate how closely the word aligns with the suggested criteria.[27] During training, pre-trained models remain constant[28] and can be utilized for word embedding. This paper employ FastText embeddings, a pre-trained model developed by Facebook for tasks involving text classification and word embedding learning. It is a simple neural network that uses only one layer for word representation, instead of assigning vectors for words directly, it represents the word as an n-gram of characters.[25] For example, in the word "technology" with n = 3, the representation of this word is <te, tec, ech, chn, hno, nol, olo, log, ogy, gy>, where the angular brackets delineate the word's beginning and end. Once the word is represented using n-grams, a model is trained to learn the embeddings. FastText excels with rare words or tokens, enabling it to break down unfamiliar words into n-grams to acquire their embeddings.[22]

Attention Mechanism: Attention Models represents a recent advancement in the fields of NLP and computer vision, often referred to as "attention is all you need." This model introduces the concept of assigning varying weights to words within a sentence. Its primary objective is to comprehensively analyze a sequence, whether it's a text, sentence, or article, and condense the most informative words into a fixed-length context vector. Consequently, the model prioritizes specific words in the text while downplaying others.[19] This mechanism enables the model to concentrate on the relevant portions of both the question and the context. The attention architecture comprises three layers: the encoder layer, the attention layer, and the decoder layer. Typically, LSTM or GRU layers are employed in the encoder and decoder components. On the other hand, the attention layer's primary role is to generate a context vector, achieved through three distinct processes: alignment, softmax calculation, and context vector computation. The weights $\alpha_{(t)}$ are computed by a softmax function given by the Eq. (11) and context

vector is given by the Eq. (12) show way of how allocating different weights to the various words in the text.[29]

$$\alpha_{(t)} = \frac{\exp[\hat{h} * v^T]}{\sum_{all\ t} \exp[h * v]} \quad (11)$$

$$S_{vec} = \sum_{all\ t} h_t\, \alpha_{(t)} \quad (12)$$

Where v represents a trainable parameter and $\hat{h}$ is obtained by Eq. (10), and h is obtained by Eq. (9).

The proposed system: Attention-based question answering system

Our proposed system is shown in Fig. 2:

Fig. 2 shows that our system which consists of several models and steps beginning with preprocessing for text followed by feature extraction using word embedding, BiGRU and BiLSTM, Attention mechanism and finally classifier for answer generation:

Text pre-processing: Involves the conversion of textual content or sentences into a suitable format that aligns with the classifier model's requirements. Its primary objective is to reduce the feature space or dimensionality of the data. This procedure encompasses various stages in the context of text classification, including the elimination of punctuation, symbols, numbers, and similar elements. Once the text is purified, the tokenization process is performed. Word tokenization is the technique used to segment the text document into individual words or tokens based on the spaces between them. Typically, in machine learning, two essential pre-processing tasks are executed: the removal of stop words and

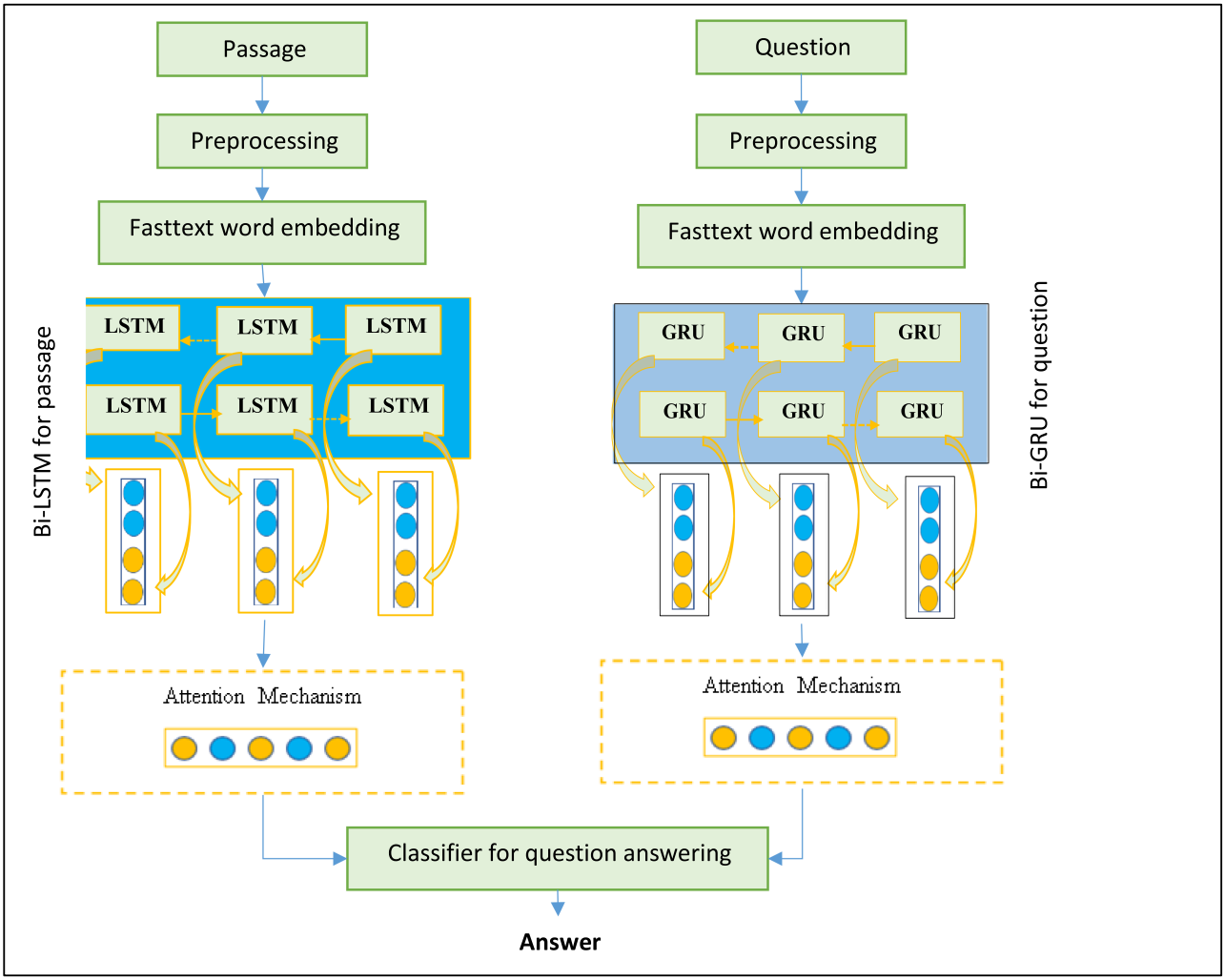

Fig. 2. The proposed binary question answering system.

stemming. These tasks serve a crucial purpose in data reduction by eliminating irrelevant words and returning words to their root forms.

Feature extraction: Is the task that involves the conversion of specific text into features. This process yields numerical vectors, so they are commonly referred to as vectorization. These extracted features from the text are then inputted into the prediction model to facilitate text classification. However, contemporary research endeavors concentrate on embedding methods for word representation. These methods assign a vector to each word based on the word's semantic meaning. Feature extraction involve word embedding using the pretrain embedding model FastText and then extract features form the question (at sentence level) using BiGRU model and features from passage using BiLSTM model.

The QA dataset for our model consists of a tuple of question, passage (context) and answer. The question and passage are tokenized into tokens and then these tokens are converted into vectors using the pre-trained embedding model FastText. In contrast, embedding techniques represent each word with dense vectors by suggesting a set of features or criteria. The values within the word vector indicate how closely the word aligns with the suggested criteria. During training, pre-trained embeddings remain constant and can be utilized for word embedding. Here, each word is represented with an embedding dimension of 300.

Following the word embedding process for both the passage and question, feed the embedding vectors of the question into Bi-GRU to extract features from it while the embedding vectors of the passage are fed into Bi-LSTM. It's worth noting that LSTM and GRU models inherently have a forward pass, where each element is influenced solely by preceding elements. In question answering tasks, the future content can also be useful to the previous words. So that Bi-GRU and Bi-LSTM are to consider succeeding and preceding contexts by combining backward and forward hidden layers.[30] LSTM is known to be good at capturing long-term dependencies in sequences. It can store information over longer sequences, making it suitable for tasks involving longer text or sequences. GRU is considered more computationally efficient and can perform well on tasks with shorter sequences. It has a simpler structure, which can make it easier to train on smaller datasets. For this reason, Bi-LSTM model is used for passage encoding while Bi-GRU model is used for question-encoding. Furthermore, it's demonstrated that bidirectional LSTM and bidirectional GRU can effectively match questions with answers as it leverages both past and future contexts, processing data from two directions.

The backward and forward contexts are concatenated according to the following equations:

$$h_{LSTM} = [\overleftarrow{h_{lstm}}, \overrightarrow{h_{lstm}}] \quad (13)$$

$$h_{GRU} = [\overleftarrow{h_{GRU}}, \overrightarrow{h_{GRU}}] \quad (14)$$

The Attention Model: After feature extraction is completed attention model is used which try to assign varying weights to words within a text in order to allow models to selectively focus on particular parts of the input text and obtaining a context vector of fixed-length for the most informative words when generating answers for the questions. In this system, the attention model is applied to the outputs of Bi-GRU and Bi-LSTM layers enabling the model to allocate varying levels of attention to different words within the sentence. This done by applying Eqs. (11) and (12) to h_{GRU} and h_{LSTM} respectively.

Answer Classifier: After computing the representation of each sentence with respect to the query and the overall context, pass the sentence vector through a multi-layer dense network, followed by sigmoid, to get the final answer probability distribution. At this point, there are two vectors: a vector representing the passage's most important words (relevant features) and a vector representing the query's most important words (relevant features). Thes two vectors are combined and a multi-layer dense network is used followed by sigmoid layer to predict the correct answer. The Rectified Linear Unit Activation Function (ReLU) is employed with the layers of network (except output layer). To reduce the overfitting, dropout have been added at a rate of 0.2. Finally, the output layer (classification layer) with the sigmoid activation function is applied for binary classification.

Results and discussion

This section views and discusses the results of the proposed system. The model is implemented in TensorFlow. Our model combines Bi-LSTM, Bi-GRU, and Attention Mechanism. The accuracy metric is used to identify the convergence, and binary cross-entropy is utilized as the loss function. In contrast, the Adam optimizer is used to update hyperparameters in back-propagation and binary cross-entropy is utilized as the loss function, the proposed Models are optimized by using early stopping and batch normalization. Bi-GRU and Bi-LSTM, each with 64 neurons, are utilized in the sequential layer. The batch size is 64 and the dropout probability is 0.2. The learning is set as 0.0001 whereas decay rate is set as 10^{-10}

	question	title	answer	passage
0	do iran and afghanistan speak the same language	Persian language	True	Persian (/ˈpɜːrʒən, -ʃən/), also known by its ...
1	do good samaritan laws protect those who help ...	Good Samaritan law	True	Good Samaritan laws offer legal protection to ...
2	is windows movie maker part of windows essentials	Windows Movie Maker	True	Windows Movie Maker (formerly known as Windows...
3	is confectionary sugar the same as powdered sugar	Powdered sugar	True	Powdered sugar, also called confectioners' sug...
4	is elder scrolls online the same as skyrim	The Elder Scrolls Online	False	As with other games in The Elder Scrolls serie...

(a)

	question	title	answer	passage
0	iran afghanistan speak language	Persian language	True	Persian (/ˈpɜːrʒən, -ʃən/), also known endonym...
1	good samaritan laws protect help accident	Good Samaritan law	True	Good Samaritan laws offer legal protection peo...
2	windows movie maker part windows essentials	Windows Movie Maker	True	Windows Movie Maker (formerly known Windows Li...
3	confectionary sugar powdered sugar	Powdered sugar	True	Powdered sugar, also called confectioners' sug...
4	elder scrolls online skyrim	The Elder Scrolls Online	False	As games The Elder Scrolls series, game set co...

(b)

Fig. 3. Examples from the training set (a) before and (b) after preprocessing the questions and passage.

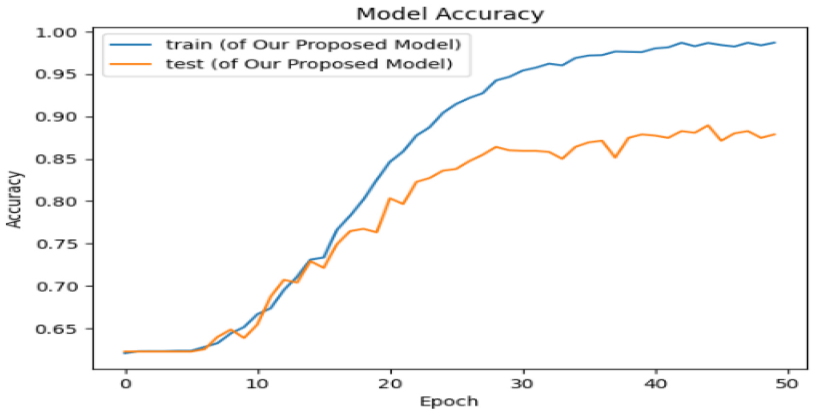

Fig. 4. The training and testing accuracy of the proposed model.

The dataset BoolQ consists of 16 k questions. The questions are split into a 3.2 k for testing set, 3.2 k validation set, and 9.4 k for training set. The queries have average length 8.9 tokens with longer passages (average length 108 tokens). Fig. 3 shows some examples from the training set and the same examples after preprocessing the questions and passage. Preprocessing includes removing of stop words, punctuation marks, special characters and stemming.

Bi-LSTM and Bi-GRU are also used for comparison because these models contributed to building our proposed model, the results of our model are better compared to all these models. Table 1 shows the experiment results of the three models:

Table 1. Summary of results and comparisons using multi models.

Model	Training Ac.	Testing Ac.
Bi-LSTM + Bi-GRU + Attention	0.9911	0.8783
Bi-LSTM	0.9892	0.8390
Bi-GRU	0.9885	0.8302

Fig. 4 explains the of the training and testing accuracy for the proposed model.

The results in Table 1 illustrate that hybrid of Bi-LSTM and Bi-GRU with attention mechanism can improve the performance of our QA system. Table 2 shows the training parameters of the model.

Table 2. The summary of parameters setting.

Name	Size
Embedding Dim.	300
No. of Units	150
Batch Size	64
Dropout Rate	0.2
Learning Rate	0.0001
Decay Rate	10^{-10}
Patience	10

Conclusion

Binary question answering using deep learning technologies has made significant strides in recent years. These systems have become increasingly accurate, versatile, and capable of handling real-world applications. Binary questions answering are challenging since it requires a wide range of inference abilities to solve. In this paper, a deep neural network is proposed to predict the answer to a given ques-

tion. This network is hybridized from two networks: Bi-LSTM, and Bi-GRU followed by the attention mechanism. The attention mechanism addresses the entire text (question or passage) and summarizes the more informative words into a fixed-length context vector. The attention model is applied at the outputs of Bi-GRU and Bi-LSTM layers to make the model pay less or more attention to different tokens in the question and passage. It is observed that Bi-LSTM with attention and Bi-GRU with attention gives good performance with an accuracy of than Bi-LSTM and Bi-GRU without attention. The results illustrated that attention mechanism can significantly improve the performance of our QA system. In future investigations, one can evaluate the models for different tasks further and try to improve our model.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all figures and tables in the manuscript are ours. Besides, figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in University of Babylon.

Authors' contributions statement

The idea is suggested by N. F. M. and I. H. A. N. F. M. conducted the practical side of the manuscript and I. H. A. checked the results and edited the paper. Both authors discussed the result and contributed to the final manuscript.

References

1. Barskar R, Ahmed GF, Barskar N. An approach for extracting exact answers to question answering (QA) system for english sentences. Procedia Eng. 2012;1187–94. https://doi.org/10.1016/j.proeng.2012.01.979.
2. Wang Z. Modern question answering datasets and benchmarks: A survey. arXiv preprint arXiv. 2022 Jun 30;2206:15030. https://doi.org/10.48550/arXiv.2206.15030.
3. Hamon T, Grabar N, Mougin F. Natural language question analysis for querying biomedical linked data arXiv preprint arXiv. 2022. https://doi.org/10.48550/arXiv.2206.15030.
4. Habimana O, Li Y, Li R, Gu X, Yan W. Attentive convolutional gated recurrent network: A contextual model to sentiment analysis. Int J Mach Learn Cybern. 2020 Dec 1;11(12):2637–51. https://doi.org/10.1007/s13042-020-01135-1.
5. Claro DB, Souza M, Castellã Xavier C, Oliveira L. Multilingual open information extraction: Challenges and opportunities. Information. 2019 Jul 2;10(7):228. https://doi.org/10.3390/info10070228.
6. Rahman N, Borah B. Improvement of query-based text summarization using word sense disambiguation. Complex Intell Syst. 2020 Apr;6:75–85. https://doi.org/10.1007/s40747-019-0115-2.
7. Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst. 2021 Feb 1;32(2):604–24. https://doi.org/10.1109/TNNLS.2020.2979670.
8. Gao S, Huang Y, Zhang S, Han J, Wang G, Zhang M, *et al.* Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. J Hydrol (Amst). 2020 Oct 1;589:125188. https://doi.org/10.1016/j.jhydrol.2020.125188.
9. Mihaylova T, Karadjov G, Atanasova P, Baly R, Mohtarami M, Nakov P. SemEval-2019 task 8: Fact checking in community question answering forums. In Proceedings of the 13th International Workshop on Semantic Evaluation. 2019 May 25. https://doi.org/10.48550/arXiv.1906.01727.
10. Nakov P, Marquez L, Barron-Cedeno A, Gencheva P, Karadzhov G, Mihaylova T, *et al.* Automatic fact checking using context and discourse information. J Data Inf Qual. ACM. 2019;11(3):1–27. https://doi.org/10.1145/3297722.
11. Miranda S, Vlachos A, Nogueira D, Secker A, Mendes A, Garrett R, *et al.* Automated fact checking in the news room. In The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019. Association for Computing Machinery, Inc. 2019:3579–83. https://doi.org/10.48550/arXiv.1904.02037.
12. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: A large-scale dataset for fact extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Mar 14. https://doi.org/10.48550/arXiv.1803.05355.
13. Nasir JA, Khan OS, Varlamis I. Fake news detection: A hybrid CNN-RNN based deep learning approach. IJIM Data Insights. 2021 Apr 1;1(1). https://doi.org/10.1016/j.jjimei.2020.100007.
14. Reddy S, Chen D, Manning CD. CoQA: A conversational question answering challenge. arXiv. 2018 Aug 21. https://doi.org/10.48550/arXiv.1808.07042.
15. Rakotoson L, Letaillieur C, Massip S, Laleye FAA. Extractive-boolean question answering for scientific fact checking. In MAD 2022 - Proceedings of the 1st International Workshop on Multimedia AI against Disinformation. Association for Computing Machinery, Inc. 2022:27–34. https://doi.org/10.48550/arXiv.2204.12263.
16. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, *et al.* VQA: Visual question answering. In Proc IEEE Int Conf Comput Vis. 2015;2425–2433. https://doi.org/10.48550/arXiv.1505.00468.
17. Wu Q, Teney D, Wang P, Shen C, Dick A, Van Den Hengel A. Visual question answering: A survey of methods and datasets. Comput Vis. Image Underst. 2017 Oct 1;163:21–40. https://doi.org/10.1016/j.cviu.2017.05.001.
18. Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh A, Alberti C, *et al.* Natural questions: A benchmark for question answering research. Trans Assoc Comput Linguist. 2019;7:452–466. https://doi.org/10.1162/tacl_a_00276.
19. Clark C, Lee K, Chang MW, Kwiatkowski T, Collins M, Toutanova K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. arxiv. 2019 May 24. https://doi.org/10.48550/arXiv.1905.10044.

20. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000 + questions for machine comprehension of text. Arxiv. 2016 Jun 16. https://doi.org/10.48550/arXiv.1606.05250.
21. Shewalkar A, Nyavanandi D, Ludwig SA. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. J Artif Intel Soft Comput Res. 2019 Oct 1;9(4):235–45. https://doi.org/10.2478/jaiscr-2019-0006.
22. Pintas JT, Fernandes LA, Garcia AC. Feature selection methods for text classification: A systematic literature review. Artif Intell Rev. 2021 Dec;54(8):6149–200. https://doi.org/10.1007/s10462-021-09970-6.
23. Dzisevič R, Šešok D. Text classification using different feature extraction approaches. In 2019 Open Conf Electr Electron Inf Sci (eStream). 2019 Apr 25;1–4. IEEE. https://dio.org/10.1109/eStream.2019.8732167.
24. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In Proc Conf Empir Methods Nat Lang Process (EMNLP). 2014 Oct;1532–1543. https://doi.org/10.3115/v1/D14-1162.
25. Naseem U, Razzak I, Khan SK, Prasad M. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. ACM Trans Asian Low-Resour Lang Inf Process. 2021 Jun 30;20(5):1–35. https://doi.org/10.48550/arXiv.2010.15036.
26. Li Y, Yang T. Word embedding for understanding natural language: a survey. Big Data Appl. 2018:83–104. https://doi.org/10.1007/978-3-319-53817-4_4.
27. Ruder S, Vulić I, Søgaard A. A survey of cross-lingual word embedding models. J Artif Intell Res. 2019 Aug 12;65:569–631. https://doi.org/10.1613/jair.1.11640.
28. Fadhil OY, Mahdi BS, Abbas AR. Using VGG models with intermediate layer feature maps for static hand gesture recognition. Baghdad Sci J. 2023 Oct 1;20(5):1808. https://doi.org/10.21123/bsj.2023.7364.
29. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. Science Direct. 2021 Sep 10;452:48–62. https://doi.org/10.1016/j.neucom.2021.03.091.
30. Asroni A, Ku-Mahamud KR, Damarjati C, Slamat HB. Arabic speech classification method based on padding and deep learning neural network. Baghdad Sci J. 2021 Jun 20;18(2 (Suppl.)):0925. https://doi.org/10.21123/bsj.2021.18.2(Suppl.).0925.

الإجابة على الأسئلة الثنائية باستخدام نظام هجين من (Bi-LSTM و Bi-GRU) بالاعتماد على آلية Attention

ندى فاضل محمد، اسراء هادي علي

كلية تكنولوجيا المعلومات، جامعة بابل، بابل، العراق.

الخلاصة

تعد الإجابة على الأسئلة (QA) جانبًا مهمًا في مجال معالجة اللغات الطبيعية (NLP) وأنظمة استرجاع المعلومات. عادةً يأمل المستخدمون في المساعدة في الحياة اليومية من خلال تعليم وتدريب البرنامج كيفية الإجابة على الأسئلة مثل أي شخص حقيقي. يهدف نظام الإجابة على الاسئلة إلى استخدام تقنيات معالجة اللغة الطبيعية لتوليد إجابة صحيحة لسؤال معين وفقًا لسياق معين أو المعرفة المحددة في مجموعة كبيرة من النصوص غير مهيكلة. تتضمن الإجابة على الأسئلة الثنائية إعطاء إجابات ثنائية (نعم/لا، صحيح/خطأ) على الأسئلة المطروحة باللغة الطبيعية. مع تطور التعلم العميق على مر السنين، لعبت تقنيات التعلم العميق دورًا مهما في تطوير أحدث أنظمة الإجابة على الاسئلة، مما مكنها من فهم الأسئلة الاجابة عليها. في هذا البحث، نقترح نموذجًا للإجابة على الأسئلة الثنائية يعتمد على آلية attention الهجين، والذي يدمج تقنيتين للتعلم العميق (Bi-LSTM و Bi-GRU) يتم تطبيق آلية attention عند مخرجات Bi-LSTM و Bi-GRU لجعل النموذج يعطي اهتماما مختلفا (أقل أو أكثر) للكلمات المختلفة في السؤال والنص وهذا يسمح للسؤال بالتركيز على جزء معين من الإجابة. تم إجراء التجارب على قاعدة البيانات BoolQ. وقد تمت ملاحظة أن تهجين Bi-LSTM و Bi-GRU مع آلية attention أعطى دقة قدرها 0.8783 مقارنة بدقة استخدام موديل Bi-LSTM فقط أو استخدام موديل Bi-GRU..

الكلمات المفتاحية: الية attention، Bi-GRU، Bi-LSTM، التعلم العميق، الشبكة العصبية الالتفافيةRNN، الأسئلة النصية.