


## ARTICLE

# Improved bidirectional attention flow (BIDAF) model for Arabic machine reading comprehension

Mariam M. Biltawi<sup>1</sup> , Arafat Awajan<sup>2</sup> and Sara Tedmori<sup>2</sup>

<sup>1</sup>School of Computing and Informatics, Al Hussein Technical University, Amman, Jordan and <sup>2</sup>Computer Science Department, Princess Sumaya University for Technology, Amman, Jordan

**Corresponding author:** Mariam M. Biltawi; Email: [mariam.biltawi@htu.edu.jo](mailto:mariam.biltawi@htu.edu.jo)

(Received 6 August 2021; revised 2 August 2024; accepted 2 August 2024; first published online 31 October 2024)

## Abstract

Machine reading comprehension (MRC) refers to the process of instructing machines to comprehend and respond to inquiries based on a provided text. There are two primary methodologies for achieving this: extracting answers directly from the text or predicting them. Extracting answers involves anticipating the specific segment of text containing the answer, pinpointed by its starting and ending indices within the paragraph. Despite the increasing interest in MRC, exploration within the framework of the Arabic language faces limitations due to various challenges. A significant impediment arises from the inadequacy of resources available for Arabic textual content, which impedes the development of effective models. Furthermore, the inherent intricacies of Arabic, manifesting in its diverse linguistic forms including classical, modern standard, and colloquial, present distinctive hurdles for tasks involving language comprehension. This paper proposes an enhanced version of the bidirectional attention flow (BIDAF) model for Arabic MRC, constructed upon the Arabic Span-Extraction-based Reading Comprehension Benchmark (ASER). ASER comprises 10,000 sets of questions, answers, and passages, partitioned into a training set constituting 90% of the data and a testing set making up the remaining 10%. By introducing a new input feature based on parts-of-speech (POS) word embeddings and replacing Bidirectional Long Short-Term Memory (bi-LSTM) with bidirectional gated recurrent unit, significant improvements were observed. Eight different POS word embeddings were generated using both Continuous Bag of Words (CBOW) and Skip-gram methods, with varying dimensionalities. Evaluation metrics, including exact match (EM) and F1-measure, were utilized to assess model performance, with emphasis on the latter for its accuracy. The proposed enhanced BIDAF model achieved a remarkable accuracy of 75.22% on the ASER dataset, demonstrating its efficacy in Arabic MRC tasks. Additionally, rigorous statistical evaluation using a two-tailed paired samples *t*-test further validated the findings, highlighting the significance of the proposed enhancements in advancing Arabic language processing capabilities.

**Keywords:** Question answering; natural language interaction; information extraction; machine reading comprehension; modern standard Arabic

## 1. Introduction

Developing models capable of understanding and extracting information from textual passages to answer targeted questions is referred to as machine reading comprehension (MRC). This task poses significant challenges due to the complexity of teaching models to interpret natural language. The importance and focus on MRC have increased for various reasons. One factor is the availability of carefully assembled datasets, such as the one introduced by Biltawi *et al.* (2020a). Additionally, there's a rising interest among researchers in utilizing neural networks (NNs), and

the accessibility of affordable and powerful graphical processing units has also played a significant role, as noted by Seo *et al.* (2016).

According to Chen (2018), there exist four distinct categories of MRC datasets: span-extraction, multiple-choice, cloze-style, and free-form, each containing sets of passage-question-answer triples. Span-extraction datasets, exemplified by SQuAD (Rajpurkar *et al.* 2016) and NewsQA (Trischler *et al.* 2017), involve extracting a single text span from the relevant paragraph as the answer to a given question. Multiple-choice datasets, such as SciQ (Welbl *et al.* 2017), present questions alongside two or more potential answers, the correct one, and the corresponding paragraph. Cloze-style datasets, like CNN-Daily Mail (Hermann *et al.* 2015), involve questions where a term or entity is missing. Free-form datasets, exemplified by MS MARCO (Nguyen *et al.* 2016), present questions and paragraphs without explicit answers, requiring systems to infer the answers from the passages.

Arabic is a challenging language distinct from English (Biltawi *et al.* 2021) and is still at an early stage in terms of MRC research. With English's global predominance, researchers have primarily focused on English MRC, developing and utilizing benchmark datasets (Alian and Al-Naymat 2022). Experimental findings indicate that NNs and attention mechanisms can effectively improve answer extraction from passages. However, research targeting Arabic MRC remains limited, with only a few studies conducted in this area (Biltawi *et al.* 2021).

The primary focus of this paper is proposing an enhancement to the bidirectional attention flow (BIDAF) model. For evaluation, the Arabic Span-Extraction-based Reading Comprehension Benchmark (ASER), comprising 10,000 question-answer-passage triples, was utilized as the benchmark dataset. The objective of this study is to introduce an improved version of the BIDAF model specifically tailored for Arabic MRC and to compare it against four baseline models: the sequence-to-sequence model, the original BIDAF model with two input layers, the original BIDAF model with one input layer, and the AraBERT with BIDAF model. Additionally, the enhancement of the improved-BIDAF model was carried out in two stages: initially, the first layer was replaced with the parts-of-speech (POS) embedding layer, followed by the substitution of bi-LSTM with bidirectional gated recurrent unit (bi-GRU) in the contextual and model layers. This adjustment led to improved model performance on Arabic text, achieving an accuracy rate of 75.22%.

The structure of the paper is organized as follows: Section 2 discusses related work, Section 3 outlines the problem statement, Section 4 introduces ASER, Section 5 presents the improved BIDAF model, Section 6 discusses the fine-tuned AraBERT BIDAF model, Section 7 details the experimental setup, Section 8 presents the experimental findings, Section 9 provides a comparison and discussion, and finally, Section 10 concludes the paper.

## 2. Related work

The field of MRC has experienced notable advancements, particularly in the English language, where researchers have explored various methodologies to enhance MRC performance. Initially, the adoption of NNs and word2vec embeddings laid the groundwork for subsequent developments. The emergence of transformer models marked a significant breakthrough, achieving remarkable results on various MRC benchmarks. Recently, research focus has shifted toward tackling more intricate MRC challenges, including addressing unanswerable questions (Hu *et al.* 2019), integrating reasoning capabilities (Li *et al.* 2022), handling queries based on multiple passages (Dong *et al.* 2023), and exploring conversational MRC (Gupta *et al.* 2020). However, in the context of the Arabic language, MRC research remains in its early stages. Limited progress in this domain can be attributed to the inherent complexities of Arabic, characterized by its rich morphology and complex syntax, demanding specialized approaches for effective comprehension. Moreover, the scarcity of large-scale Arabic MRC datasets presents a significant hurdle to further advancements in this field.

Recently, several surveys have been conducted on the topic of MRC. For instance, Baradaran *et al.* (2022) reviewed 241 research papers spanning from 2016 to 2020, presenting three primary observations: a shift in research focus from extraction to generation, from single-document (passage) to multi-document (passage) reading comprehension, and from scratch learning to the utilization of pre-trained embeddings. Another survey by Xin *et al.* (2019) provided a comprehensive overview of datasets, neural models, and various techniques employed in English MRC, covering popular methods such as Word2Vec, Glove, ELMO, BERT, and GPT. Liu *et al.* (2019) aimed to cover MRC tasks, general NN models, their architectures used for MRC, and the emerging trends and challenges in this field. Additionally, Zeng *et al.* (2020) analyzed fifty-seven MRC tasks and datasets, proposing a novel taxonomy for categorizing MRC tasks based on corpus types, questions, answers, and sources of answers. These surveys consistently emphasize the distinction between MRC and Question Answering (QA), noting that MRC involves two inputs (the question and the context) and one output (the answer), while QA typically involves one input (the question) and one output (the answer). Note that phases of MRC development can be grouped into rule-based techniques, classical ML techniques, and deep learning techniques.

For the Arabic language, the task of MRC has been explored in a few research papers. Some of these efforts include work on Quranic datasets, such as the research by Aftab and Malik (2022) and the attempt by Malhas and Elsayed (2022). These investigations conducted experiments utilizing BERT and AraBERT, respectively, on the Qur'anic Reading Comprehension Dataset. The reported highest exact match (EM) scores achieved were 8.82% and 28.01%, respectively, following the fine-tuning of the AraBERT model on classical language. Correspondingly, the highest F1-measure scores attained were 26.76% and 49.68%, respectively.

An inherent challenge in effectively implementing MRC lies in the availability of suitable datasets for training and evaluating models. The presence of high-quality and diverse datasets is important in developing MRC models capable of accurately and comprehensively answering questions. Various datasets have been created for the English language, including SQuAD (Rajpurkar *et al.* 2016), NewsQA (Trischler *et al.* 2017), MCTest (Richardson *et al.* 2013), and MS MARCO (Nguyen *et al.* 2016), among others. Additionally, numerous models have been developed based on these English datasets, such as BIDAf (Seo *et al.* 2016), FastQA (Weissenborn *et al.* 2017), and BERT (Kenton and Toutanova 2019), to name a few. In contrast, there have been only a few endeavors to establish Arabic MRC benchmarks, as evidenced by works by Biltawi *et al.* (2020b) and Biltawi *et al.* (2020a). However, progress in Arabic MRC has been relatively constrained, with only a few large-scale datasets available, including Arabic SQuAD (Mozannar *et al.* 2019), AQAD (Atef *et al.* 2020), and ASER (Biltawi *et al.* 2023). Recently, (Alnefaie *et al.* 2023) presented two novel question-answer datasets, HAQA for Hadith and QUQA for the Quran, emphasizing the challenges in comparing their performance due to the absence of a standardized test dataset for Hadith and the relatively simplistic nature of questions in the Quran dataset. HAQA, the Arabic Hadith question-answer dataset, was built from various expert sources, while QUQA a series of construction phases, including integration with existing datasets and supplementation with new data from expert-authored texts, and datasets comprising 1,598 and 3,382 question-answer pairs, respectively.

The key distinction between these datasets and previous attempts primarily lies in their size, with these datasets containing 10,000 or more records of data, whereas prior attempts typically include 2,000 or fewer records of data. Moreover, these datasets are specifically tailored for the task of MRC and are structured as triples comprising a question, an answer, and a context, as opposed to only including the question and answer. Additionally, there are notable differences between Arabic SQuAD, AQAD, and ASER. First, ASER was created manually by native Arabic speakers, whereas Arabic SQuAD is essentially a translated version of the English SQuAD, and AQAD was automatically generated using Arabic articles. Second, there are variations in the length of questions and answers among these datasets. ASER poses a greater challenge as it comprises longer sentences compared to Arabic SQuAD and AQAD, which comprise shorter questions and answers.

The study by Mozannar *et al.* (2019) evaluated the performance of QANet and BERT models on the Arabic SQuAD dataset. The experiments resulted in an EM score of 29.4% and 34.2% and an F1-measure of 44.4% and 61.3% for the QANet and BERT models, respectively. Similarly, Atef *et al.* (2020) conducted experiments on the AQAD dataset using BIDAf and BERT models. The results demonstrated an EM score of 32% for BIDAf and 33% and 37% for BERT, with corresponding F1 measures. Furthermore, Biltawi *et al.* (2023) performed baseline experiments on the ASER dataset, employing sequence-to-sequence, BIDAf, and AraBERT BIDAf models. The findings revealed an EM score of 2.5% for the sequence-to-sequence model, 39.5% for the BIDAf model, and 0% for the AraBERT BIDAf model. Additionally, F1 measures were reported as 35.76%, 66.25%, and 19.73% for the sequence-to-sequence, BIDAf, and AraBERT BIDAf models, respectively.

Additionally, two research papers by Alkhatnai *et al.* (2020) and Biltawi *et al.* (2021) investigated the trends, challenges, and conducted gap analysis in MRC. Both studies highlighted the absence of standardized benchmarks and the complexities inherent in the Arabic language, which pose obstacles to progress in this domain. To advance the field of Arabic MRC, it is imperative to refrain from excluding certain techniques or models during experimentation, such as solely focusing on BERT while disregarding Word2Vec. Instead, the emphasis should be on comprehensively assessing the effectiveness of each approach, particularly in the context of Arabic. This approach seeks to evaluate the efficacy of different techniques when applied to Arabic, rather than simply following the latest trends since the available datasets for the Arabic language are still moderate in size and BERT needs more data compared to Word2Vec.

The current paper differs from related works presented in this section, by extending beyond proposing benchmark datasets and experimenting with preexisting English models for Arabic MRC. Rather, this paper introduces a novel enhancement to the BIDAf model, customized specifically for Arabic, with the objective of enhancing answer extraction. The emphasis is on introducing new features and experimenting with different neural units to tackle the unique challenges of the Arabic language. This approach aims to contribute to the development of more effective and specialized MRC models for Arabic.

### 3. Problem statement

It is possible to structure the MRC task as a supervised learning problem. Given training set triples of question-answer-passage  $(q^i, a^i, p^i)_{i=1 \dots n}$ , the objective is to train a model  $f$  that can produce one right answer  $a$ , given a passage  $p$  and a question  $q$ . The model's two inputs and output are shown in Equation (1):

$$f : (p, q) \rightarrow a \quad (1)$$

The passage, denoted as  $p$ , consists of  $|p|$  tokens:  $p = (p_1, p_2, \dots, p_{|p|})$ . Similarly, the question, denoted as  $q$ , consists of  $|q|$  tokens:  $q = (q_1, q_2, \dots, q_{|q|})$ . Each passage token  $p_i \in V$  for  $i = 1, \dots, |p|$ , and each question token  $q_i \in V$  for  $i = 1, \dots, |q|$  where  $V$  represents a predefined vocabulary. The answer  $a$  is a span within the passage, represented as  $(a_{start}, a_{end})$  with the constraint that  $p_1 \leq a_{start} \leq a_{end} \leq p_{|p|}$  (Chen, 2018). The trained model  $f$  will then be evaluated using a testing set.

### 4. Arabic Span-Extraction-based Reading Comprehension Benchmark (ASER)

The experiments were conducted on ASER which is an Arabic Span-Extraction-based Reading Comprehension Benchmark created manually and proposed by Biltawi *et al.* (2023). ASER was created over the period of two semesters, where a large number of university students helped in writing questions and their answers on articles crawled from Aljazeera website belonging to

Semester	Article-ID	Question	Answer	Paragraph	Start-index	Domain
1	19	ما اسم المنتدى الدولي الذي عقد في العاصمة السعودية	مبادرة مستقبل الاستثمار	كما يأتي قبل يوم من انطلاق منتدى دولي في العاصمة السعودية بعنوان مبادرة مستقبل الاستثمار الذي ينظمه صندوق الاستثمارات العامة السعودي برئاسة محمد بن سلمان ويستمر ثلاثة أيام	12	عربي
2	1049	من يحضر مهرجان هذا العام الذي طُغت عليه الجوانب السياسية والقضية الفلسطينية؟	العديد من الساسة والمفكرين العرب والأجانب من بينهم عدد من الشخصيات الفلسطينية مثل حيدر عبد الشافي وعزمي بشارة ومصطفى البرغوثي إلى جانب الأمين العام للجامعة العربية عمرو موسى ومن بين المفكرين الأجانب حضر الإعلامي والسياسي الفرنسي إريك رولو ووزير الخارجية الإندونيسي السابق علي العطاس	ويحضر مهرجان هذا العام الذي طُغت عليه الجوانب السياسية والقضية الفلسطينية العديد من الساسة والمفكرين العرب والأجانب من بينهم عدد من الشخصيات الفلسطينية مثل حيدر عبد الشافي وعزمي بشارة ومصطفى البرغوثي إلى جانب الأمين العام للجامعة العربية عمرو موسى ومن بين المفكرين الأجانب حضر الإعلامي والسياسي الفرنسي إريك رولو ووزير الخارجية الإندونيسي السابق علي العطاس	11	ثقافة

Figure 1. Examples from ASER.

twenty-five domains. Two Arabic native speakers validated the dataset and also performed some editing, resulting in the creation of 10,000 records of question-answer-passage triples. These records were divided into a training set of 9,000 records, a testing set of 1,000 records, and another testing set consisting of 100 records sampled from the original testing set for human performance evaluation. The human performance resulted in an EM score and F1-measure of 42% and 71.62%, respectively.

The authors also conducted neural baseline experiments on ASER. Results showed an EM and F1-measure of 0% and 15.96%, respectively, on AraBERT BIDAf model, 4% and 36.9%, respectively, on the sequence-to-sequence model, and 38% and 67.54%, respectively, on the original BIDAf model, all on the 100 testing set. Figure 1 demonstrates two examples from ASER, where each record of ASER consists of semester-number, article-ID, question, answer, paragraph, first index of the answer from the paragraph, and domain of the article. ASER includes both long and short answers, with lengths varying from two to seventy-five tokens. The human performance EM score of only 42%, and the varying lengths of the answers make ASER a challenging benchmark.

## 5. Improved bi-directional attention flow (BIDAf) model

Several experiments were carried out to customize BIDAf for the Arabic language. The most promising outcomes were achieved through the improved version of BIDAf (AKA improved-BIDAf), as depicted in Figure 2. This improved-BIDAf model incorporates four inputs: the POS word embeddings for both the question and the context, along with the word embeddings for both the question and the context. Prior to feeding the text into the model, tokenization of the input is necessary, where  $x_1, x_2, \dots, x_T$  represent the context tokens, and  $q_1, q_2, \dots, q_J$  denote the question tokens.

### 5.1. POS word embedding layer

In this layer, pretrained POS word embeddings were utilized instead of employing character-level convolutional neural networks (CNN). These POS word embeddings were applied to both the question  $Q_{POS}$  and the context  $X_{POS}$ . The dimension of the embeddings in this layer ranges from 3 to 32, and you can find a more detailed explanation in Section 7.

### 5.2. Word embedding layer

Instead of using Glove embeddings, pre-trained Aravec embeddings (Soliman *et al.* 2017) were employed for both the question  $Q_{word}$  and the context  $X_{word}$ . The embedding dimension in this layer can either be 100 or 300.

It's important to note that the embedding dimensions of the POS and word embedding layers are different. As a result, they are not concatenated before being passed to the contextual embedding layer. Instead, the contextual embedding layer works to unify the embedding dimension for both the POS and word embeddings.

Then, the POS and word embeddings for the question are concatenated, and similarly, the POS and word embeddings for the context are concatenated as well. These concatenated embeddings are then passed to the attention flow layer for further processing. This approach ensures that the model can effectively utilize both the POS and word information during the attention flow process.

### 5.3. Contextual embedding layer

In this layer, a bi-GRU is utilized to capture the temporal interactions between the words in both the question  $U_{word} \in R^{2dxJ}$  and the context  $H_{word} \in R^{2dxT}$  independently, as well as between the POS tags of both the question  $U_{POS} \in R^{2dxJ}$  and the context  $H_{POS} \in R^{2dxT}$ . Then these are concatenated for both the question  $U = [U_{word}; U_{POS}]$  and the context  $H = [H_{word}; H_{POS}]$ . As a result, the outputs of this layer are the column vectors  $U \in R^{4dxJ}$  for the question and  $H \in R^{4dxT}$  for the context, where  $d$  represents the dimensionality of the embeddings, and  $J$  and  $T$  represent the number of words in the question and context, respectively.

### 5.4. Attention flow layer

In this layer, attention from two directions is computed: question-to-context attention (AKA Query2Context), denoted by  $\hat{H}$ , which signifies the context words that are more relevant to question words, and context-to-question (AKA Context2Query), denoted by  $\hat{U}$ , which signifies the question words that are more relevant to context words. These attentions are derived from a shared similarity matrix  $S \in R^{TxJ}$ . These attentions help identify the relevant context words for each question word and vice versa, highlighting the important connections between them. Then, these attentions are concatenated with the column vector  $H$  computed in the previous layer, generating the output of the current layer, which is the query-aware vector representation of the context words  $G$ . Essentially, this layer allows the model to refine its understanding of the context by considering the relevance of each context word to the question and vice versa, enabling a more contextually aware representation for further processing.

### 5.5. Modeling layer

The goal of this layer is to record the interactions between the context words conditioned based on the questions. This layer is implemented using two bi-GRU, where  $G$  is the layer's input and  $M \in R^{4dxT}$  is the layer's output.

### 5.6. Output layer

In this layer, the answer span which is represented by the begin and end indices is predicted.

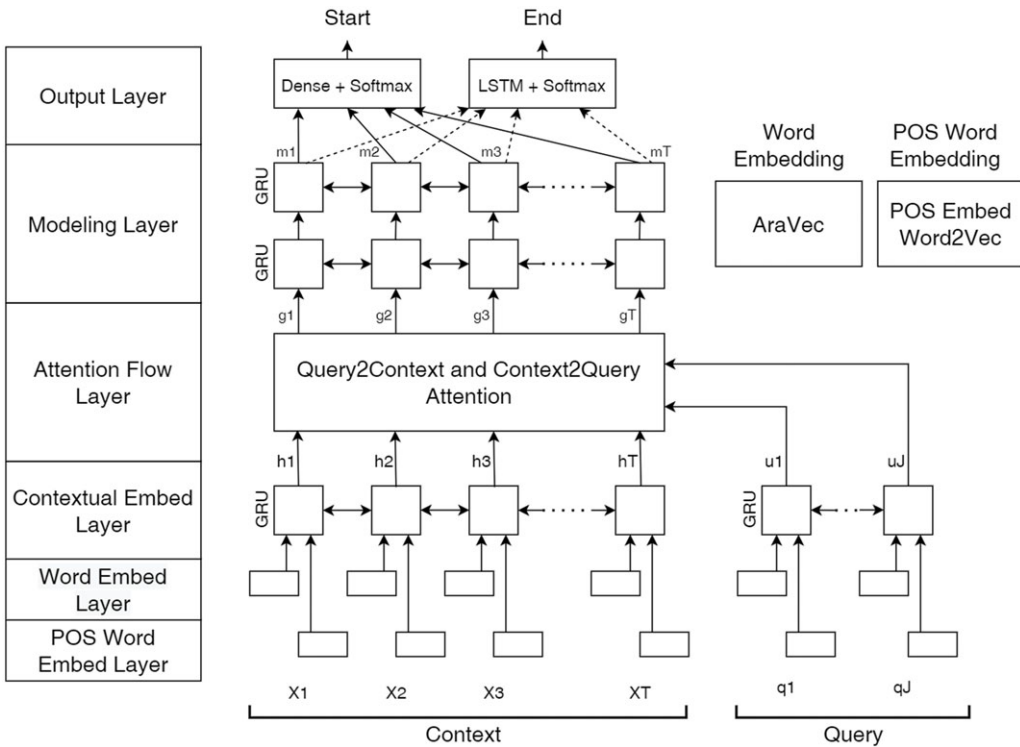


Figure 2. Improved-BIDAF Architecture.

## 6. Fine-tuned AraBERT BIDAF

The BIDAF model and AraBERT were both used in a previous work as a baseline experiment without applying fine-tuning on ASER dataset. In this work, the authors applied fine-tuning on the AraBERT model and used the pre-trained embeddings as an input to the BIDAF model as shown in Figure 3. The experiments involved two variations. In the first experiment, bi-LSTM was utilized within both the contextual embedding layer and the modeling layer. For the second experiment, bi-LSTM was replaced with bi-GRU.

## 7. Experimentations

This section presents an overview of the modification steps of the BIDAF improved model and the experimental settings.

### 7.1. Arabic embeddings

AraVec comprises twelve pre-trained Arabic embeddings, available in two main dimensions: 100 and 300. These embeddings were trained on diverse sources, including Wikipedia, tweets, and the World Wide Web (www), using two distinct embedding methods: Continuous Bag of Words (CBOW) and Skip-gram. For this study, the pre-trained embeddings with a dimension of 300 were specifically evaluated for CBOW and Skip-gram on Wikipedia and www content. Tweet embeddings were not experimented with due to the belief that Wikipedia and the web content were more similar to Modern Standard Arabic (MSA) than tweets. Thus, the dimension of 300 was chosen for the experiments. After conducting several experiments, it was observed that the

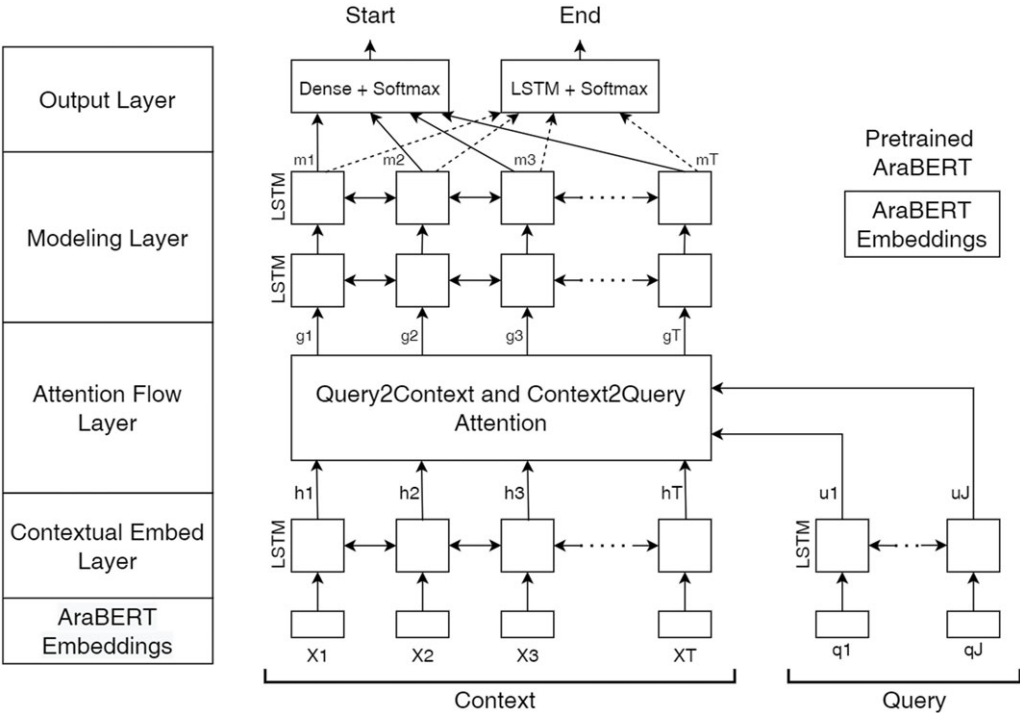


Figure 3. AraBERT BIDAFArchitecture.

results obtained using the dimension of 300 outperformed those achieved with the dimension of 100, leading to the selection of 300 as the optimal dimension for the pre-trained embeddings used in this research.

7.2. Improvement experiments

Various experiments were conducted for the purpose of improving BIDAFA for the Arabic language, including adding a new input feature and replacing bi-LSTM with bi-GRU. The new input feature is based on word embeddings for the POS tags of both the question and the passage words. The POS word embeddings were prepared as shown in the following steps:

1. Dataset preparation (POS tagged dataset). The training and testing sets of ASER were combined to represent a total of 10,000 records with seven columns. Then the three columns (question, answer, and paragraph) were merged into one column resulting in 30,000 records. After that, the Stanford POS tagger (Toutanova *et al.* 2003) was used to tag the 30,000 records, and finally, these records were saved as a new dataset having only the POS tags of the original dataset.
2. The POS word embeddings. When creating embeddings for the POS tags, the challenge was choosing the dimension since the size of the vocabulary for the tagged dataset was only 32. Different references mentioned that the embedding dimension should range between 50 and 300 (Patel and Bhattacharyya 2017), and other references mentioned that the larger the embedding dimension is, the better the performance becomes (Mikolov *et al.* 2013). However, 50 and 300 are too large for a vocabulary size of 32. Thus, after searching more on this topic, we found a general rule for choosing a dimension, which is used as a rule of thumb, this rule states that the dimension size should equal the fourth root of the number

of categories<sup>a</sup>  $\text{dimensionSize} = (\text{vocabSize})^{0.25}$ . Following this rule, with the number of categories equal to 32, the selected dimension size is 3. However, other dimensions were also experimented (32 which represents the maximum size of vocabulary, 14 which represents a number in the middle of 3 and 32, and 20 which is chosen randomly). As a result, eight POS word embeddings were obtained, four of which are CBOW and four are Skip-gram.

### 7.3. Experimental settings

This subsection presents the experimental settings for the experiments conducted in this research.

#### 7.3.1 POS word embeddings

Eight experiments were conducted to obtain the eight different POS word embeddings. The experiments were conducted using both CBOW and Skip-gram. The implementation used Word2Vec model from Genism v4.0.1 on Python v3.7 and a standalone computer with the specifications of a 2.21 GHz Intel Core i7 CPU and 16 GB RAM. The hyperparameters used were vocabulary size of 32, context window of 2, and different sizes of embedding vectors 3, 14, 20, and 32. Both CBOW and Skip-gram models were trained on the POS tags dataset.

#### 7.3.2 Improved-BIDAF

The improved-BIDAF experiments were conducted using the Adam optimizer with its default initial learning rate (0.001), two batch sizes were experimented 5 and 10, for five epochs. The training set was split into training and validation with a ratio of 80:20, respectively. These settings were configured manually after conducting several experiments. The hidden state dimension  $d$  of the improved-BIDAF model is 100. Improved-BIDAF has nine million parameters when POS word embeddings were used with the second word embedding layer (AraVec), while the parameters decreased to seven million when POS word embeddings used with the second word embedding layer (AraVec) along with replacing bi-LSTM with bi-GRU. The same settings were used when experimenting BIDAF with AraBERT.

### 7.4. Experimental measures

EM and F1-measure were the two primary metrics utilized to assess the experimental models. EM (Rajpurkar *et al.* 2016) refers to the matching between the predicted values generated by the model and the actual or golden values in the dataset. EM assigns a score of 1.0 to the predicted answer that matches the golden answer for a given question and 0 otherwise. The F1-measure is a metric used to evaluate the performance of a model's predicted answers against the true or golden answers. It is calculated as a weighted harmonic mean for the words present in both the predicted answer and the golden answer, treating both sets of words as "bags of words."

In essence, the F1-measure represents the average level of agreement between the words found in the predicted answer and the words in the golden answer for a given question. The formula for computing the F1-measure is shown in Equation (2):

$$F1 = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}), \quad (2)$$

$$\text{Where, precision} = (\text{truepositive}) / (\text{truepositive} + \text{falsepositive}) \quad (3)$$

$$\text{And, recall} = (\text{truepositive}) / (\text{truepositive} + \text{falsenegative}) \quad (4)$$

<sup>a</sup><https://developers.googleblog.com/2017/11/introducing-tensorflow-feature-columns.html>

<p>Question:</p> <p>ما هي فعاليات إحياء ذكرى النكبة داخل الخط الأخضر؟</p> <p>Passage:</p> <p>وتشمل فعاليات إحياء ذكرى النكبة داخل الخط الأخضر -الذي يعيش فيه نحو مليون ونصف مليون فلسطيني- زيارة أطلال القرى الفلسطينية المدمرة في الجليل والساحل، وتنظيم "مسيرة العودة الـ21"</p> <p>Gold answer</p> <p>زيارة أطلال القرى الفلسطينية المدمرة في الجليل والساحل، وتنظيم "مسيرة العودة الـ21"</p>	<p>Predicted answer 1:</p> <p>زيارة أطلال القرى الفلسطينية</p>	<p>EM = 0</p> <p>F1 = 50%</p>
	<p>Predicted answer 2:</p> <p>زيارة أطلال القرى الفلسطينية المدمرة في الجليل والساحل</p>	<p>EM = 0</p> <p>F1 = 80%</p>
	<p>Predicted answer 3:</p> <p>زيارة أطلال القرى الفلسطينية المدمرة</p>	<p>EM = 0</p> <p>F1 = 58.8%</p>
	<p>Predicted answer 4:</p> <p>زيارة أطلال القرى الفلسطينية المدمرة في الجليل والساحل، وتنظيم "مسيرة العودة"</p>	<p>EM = 0</p> <p>F1 = 95.7%</p>
	<p>Predicted answer 5:</p> <p>زيارة أطلال القرى الفلسطينية المدمرة في الجليل والساحل، وتنظيم "مسيرة العودة الـ21"</p>	<p>EM = 1</p> <p>F1 = 100%</p>

Figure 4. Example of EM and F1 measures.

The true positive corresponds to the number of tokens that are common between the gold answer and the predicted answer. The false positive, on the other hand, represents the number of tokens present in the predicted answer but not in the gold answer. Lastly, the false negative refers to the number of tokens present in the gold answer but missing from the predicted answer.

It is worth noting that although the EM measure is an indicator of accuracy; however, it is not considered an accurate measure, while the F1-measure can be more accurate. For instance, consider the example in Figure 4, all the predicted answers are correct, but the EM is 0%, while the F1-measure differs every time. The predicted answer must fully match the gold answer for the EM to equal 1. That’s why we applied the two-tailed paired samples *t*-test on the results of the F1-measure for all the experiments.

8. Experimental results

In this section, the results obtained from the BIDAF improvement steps are presented. The first subsection showcases the outcomes achieved when adding a POS word embedding layer as a new feature. The second subsection presents the results after replacing bi-LSTM with bi-GRU in the model. Finally, the third subsection presents the results of the performance of the AraBERT BIDAF model. These results offer insights into the effectiveness of each improvement step and highlight the overall performance gains achieved through these modifications.

8.1. Improved-BIDAF (using POS word embeddings)

Tables 1 and 2 present the results of the EM and F1-measure of the BIDAF model on a testing set consisting of 1,000 records after adding the POS word embedding layer with various dimensions

**Table 1.** EM after adding POS word embedding layer to the BIDAf model

EM 1000 test set	POS Dim = 20		POS Dim = 3		POS Dim = 14		POS Dim = 32	
Epoch	5	5	5	5	5	5	5	5
Batch	5	10	5	10	5	10	5	10
CBOW wiki	44	39.4	40.2	37.2	37.9	35.9	35.1	38.7
SG wiki	37.6	39.6	40.3	38.9	39.4	39	37.4	39.5
CBOW www	43.6	42.8	42.1	40.1	41.6	40.8	42.7	41.7
SG www	35.3	35.3	33	32	33.4	33.9	34.1	36.3

**Table 2.** F1-measure after adding POS word embedding layer to the BIDAf model

F1 1000 test set	POS Dim = 20		POS Dim = 3		POS Dim = 14		POS Dim = 32	
Epoch	5	5	5	5	5	5	5	5
Batch	5	10	5	10	5	10	5	10
CBOW wiki	74.09	64.75	67.6	65.8	67.07	63.44	63.74	67.16
SG wiki	64.27	64.39	65.69	65.18	64.48	65.52	63.6	64.34
CBOW www	70.09	70.28	69.1	68.8	69.6	68.23	69.87	69.54
SG www	60.28	59.9	59.07	59.28	62.07	61.24	61.92	61.97

(20, 3, 14, and 32). When setting the POS word embedding dimension to 20, the best EM and F1-measure results were achieved at 44% and 74.09%, respectively. These results were obtained using CBOW wiki word embeddings with both epoch and batch equal to 5. On the other hand, the worst EM result was 35.3%, which was obtained using Skip-gram WWW word embeddings. Additionally, the worst F1-measure was 59.9%. The result was obtained using Skip-gram WWW word embeddings, with epoch and batch equal to 5 and 10, respectively.

The best EM and F1-measure findings were 42.1% and 69.1%, respectively, for a POS word embedding dimension of 3. These results were produced, using CBOW WWW word embeddings with batch and epoch both set to 5. In contrast, the worst EM result was 32%, which occurred when using Skip-gram WWW word embeddings with epoch and batch both set to 5 and 10, respectively. Furthermore, the worst F1-measure was 59.07%, also obtained with Skip-gram WWW word embeddings and both epoch and batch set to 5.

When the POS word embedding dimension was set to 14, the best results achieved were 41.6% for EM and 69.6% for the F1-measure. These results were obtained using CBOW WWW word embeddings with both epoch and batch set to 5. On the other hand, the worst EM result was 33.4%, which occurred when using Skip-gram WWW word embeddings with both epoch and batch also set to 5. Similarly, the worst F1-measure was 61.24%, obtained with the same Skip-gram WWW word embeddings and with epoch and batch set to 5 and 10, respectively.

Yet, when the POS word embedding dimension was set to 32, the best results achieved were 42.7% for EM and 69.87% for the F1-measure. These results were obtained using CBOW WWW word embeddings with both epoch and batch set to 5. Conversely, the worst results were 34.1% for EM and 61.92% for the F1-measure, which were obtained using Skip-gram WWW word embeddings with both epoch and batch set to 5.

**Table 3.** EM after adding POS word embedding layer and replacing bi-GRU by bi-LSTM

EM 1000 test-set	POS Dim = 20		POS Dim = 3		POS Dim = 14		POS Dim = 32	
Epoch	5	5	5	5	5	5	5	5
Batch	5	10	5	10	5	10	5	10
CBOW wiki	32.6	44	41.9	46.6	41.5	45.8	46	38
SG wiki	30.5	28.1	26.5	28.7	27.6	28.7	29.2	27.2
CBOW www	41.8	42.2	45.1	42.9	39.5	42.9	43.1	40.2
SG www	26.1	24.5	23.2	24.3	24.9	30.3	25.9	25.4

To summarize, the experimental results obtained in this study indicate that CBOW word embeddings consistently outperformed Skip-gram word embeddings. Another significant observation was that setting both epoch and batch to 5 resulted in better performance across the different configurations. The best performance was achieved when using a POS word embedding dimension of 20 along with CBOW wiki word embeddings, with both epoch and batch set to 5. This combination yielded the highest EM and F1-measure results, reaching 44% and 74.09%, respectively. Conversely, the worst results were obtained when using a POS word embedding dimension of 3 in combination with Skip-gram WWW word embeddings.

Interestingly, it is worth noting that these findings seemed to contradict the rule of thumb mentioned in Section 7, which suggested that the dimension size of embeddings should equal the fourth root of the number of categories. In this case, the experimental results demonstrated that the optimal dimension size for the POS word embeddings did not follow this rule and that other factors might have played a more significant role in determining the best-performing configuration.

**8.2. Improved-BIDAF (bi-GRU replaced bi-LSTM)**

Tables 3 and 4 display the results for the EM and F1-measure of the improved-BIDAF model on a testing set comprising 1,000 records. The experiments involve adding a POS word embedding layer with various dimensions (20, 3, 14, and 32), as well as replacing the bi-LSTM with bi-GRU in the model. When the POS word embedding dimension was set to 20, the best EM and F1-measure results obtained were 44% and 74.09%, respectively. These results were achieved using CBOW wiki word embeddings with epoch and batch set to 5 and 10, respectively. Conversely, the worst EM and F1-measure results were 24.5% and 48.45%, respectively, obtained when using Skip-gram WWW word embeddings with epoch and batch both set to 5 and 10, respectively.

For a POS word embedding dimension of 3, the best EM and F1-measure results achieved were 46.6% and 75.22%, respectively. These results were obtained using CBOW wiki word embeddings with epoch and batch both set to 5 and 10, respectively. Conversely, the worst EM result was 23.2%, which occurred when using Skip-gram WWW word embeddings with both epoch and batch set to 5. Furthermore, the worst F1-measure result was 50.4%, which was also obtained using Skip-gram WWW word embeddings and with epoch and batch both set to 5 and 10, respectively.

When the POS word embedding dimension was set to 14, the best EM and F1-measure results obtained were 45.8% and 73.55%, respectively. These results were achieved using CBOW wiki word embeddings with epoch and batch set to 5 and 10, respectively. On the other hand, the worst EM and F1-measure results were 24.9% and 50.75%, respectively, obtained when using Skip-gram WWW word embeddings with both epoch and batch set to 5. Similarly, when the POS word embedding dimension was set to 32, the best EM and F1-measure results achieved were

**Table 4.** F1-measure after adding POS word embedding layer and replacing bi-GRU by bi-LSTM

F1 1000 test set	POS Dim = 20		POS Dim = 3		POS Dim = 14		POS Dim = 32	
Epoch	5	5	5	5	5	5	5	5
Batch	5	10	5	10	5	10	5	10
CBOW wiki	60.89	74.09	70.34	75.22	70.87	73.55	73.06	68.04
SG wiki	54.18	53.74	52.1	53.57	53.3	53.37	54.91	51.01
CBOW www	70.94	69.78	70.67	70.05	68.23	71.1	70.46	68.62
SG www	50.96	48.45	51.81	50.4	50.75	54.74	50.44	49.86

46% and 73.06%, respectively. These results were obtained using CBOW wiki word embeddings with both epoch and batch set to 5. Conversely, the worst EM and F1-measure results were 25.4% and 49.86%, respectively, obtained when using Skip-gram WWW word embeddings with epoch and batch set to 5 and 10, respectively. These results provide further insights into the performance of the improved-BIDAF model with different POS word embedding dimensions and word embedding types. It appears that using CBOW wiki word embeddings generally leads to better results compared to Skip-gram WWW word embeddings across different POS word embedding dimensions.

To summarize, the experimental results consistently showed that CBOW word embeddings outperformed Skip-gram word embeddings. Particularly, CBOW wiki word embeddings yielded the best results, while Skip-gram WWW word embeddings resulted in the worst performance. The highest EM and F1-measure results (46.6% and 75.22%, respectively) were achieved using a POS word embedding dimension of 3, CBOW wiki word embeddings, and with epoch and batch both set to 5 and 10, respectively. Interestingly, this configuration adheres to the rule of thumb discussed in Section 7, indicating that a dimension size equal to the fourth root of the number of categories might lead to optimal performance. Additionally, the replacement of bi-LSTM with bi-GRU improved the results, leading to a 2% increase in EM and a 1.13% increase in the F1-measure. On the other hand, the lowest EM result was 23.2%, obtained when using a POS word embedding dimension of 3 with Skip-gram WWW word embeddings and both epoch and batch set to 5. Meanwhile, the worst F1-measure result was 48.45%, obtained when using a POS word embedding dimension of 20 with Skip-gram WWW word embeddings and with epoch and batch set to 5 and 10, respectively. These observations shed light on the impact of different configurations on the performance of the improved-BIDAF model, emphasizing the importance of word embedding types and dimensions, as well as the choice of recurrent NN architecture.

8.3. Fine-tuned AraBERT BIDAF model

Table 5 presents the F1-measure results obtained from the experiments conducted using the fine-tuned AraBERT BIDAF model. There were four different configurations tested:

- Pretrained AraBERT with BIDAF using bi-LSTM, with an epoch set to 5 and two different batch sizes (5 and 10).
- Pretrained AraBERT with BIDAF using bi-GRU, with an epoch set to 5 and two different batch sizes (5 and 10).

For all four models, the EM metric resulted in 0%, indicating that none of the models achieved EMs with the golden answers. However, the highest F1-measure obtained was 34.12%. This result was

**Table 5.** F1-measure for the fine-tuned AraBERT BIDAf model

F1 of 1000 test set	AraBERT BIDAf bi-LSTM	AraBERT BIDAf bi-GRU
Epoch = 5 Batch = 5	33.01%	30.52%
Epoch = 5 Batch = 10	33.44%	34.12%

**Table 6.** The models selected for comparison with the highest results

	EM (1000 test set)	F1 (1000 test set)	EM (100 test set)	F1 (100 test set)
Human	N/A	N/A	42%	71.62%
Seq2Seq	3.3%	34.53%	4%	36.9%
AraBERT	0.1%	19.73%	0%	15.96%
Fine-tuned AraBERT	0.2%	33.1%	0%	32.93%
BIDAf1	23.3%	48.14%	24%	43.14%
BIDAf2	39.5%	66.25%	38%	67.54%
Improved-BIDAf1	44%	74.04%	45%	73.61%
Improved-BIDAf2	46.6%	75.22%	47%	78.52%

achieved when using pre-trained AraBERT embeddings and replacing bi-LSTM with bi-GRU within the BIDAf model, with an epoch set to 5 and a batch set to 10, showing an improvement of 14.39% from the baseline AraBERT BIDAf model. These results indicate that while the models did not perform well in terms of EM, the F1-measure improved slightly in the configuration with pre-trained AraBERT embeddings and bi-GRU.

9. Comparison and discussion

In this section, a comprehensive comparison is conducted between the improved BIDAf model, human performance, several baseline models on the ASER dataset, and other models.

9.1. Improved BIDAf and baseline models

Table 6 presents the results for the EM and F1-measure for each of these models on both the 1,000 and 100 testing sets. It is important to note that the human performance evaluation was only conducted on the 100-testing set. The baseline models include:

- Sequence-to-sequence models using bi-LSTM as both the encoder and decoder.
- BIDAf1 model, which replaces the character embedding layer with the Arabic fastText embedding layer. This model represents the original BIDAf.
- BIDAf2 model, which is implemented without the character embedding layer. It also represents the original BIDAf.
- AraBERT BIDAf model using bi-GRU.
- Fine-tuned AraBERT BIDAf model using bi-LSTM.

Both the improved-BIDAF1 and improved-BIDAF2 models use a POS word embedding layer. The main difference between these two models is that the latter replaces bi-LSTM with bi-GRU. Both of these models are designed for the Arabic language. The results in Table 6 provide insights into the performance of these models on the ASER dataset. The comparison with human performance on the 100-testing set allows for assessing how well the models perform relative to human-level understanding and comprehension.

The results demonstrate that Improved-BIDAF2 achieved excellent performance on both the 1,000-record testing set and the 100-record testing set. It surpassed all other models, including human performance on the smaller testing set. For the 100-record testing set, the human performance achieved an EM and F1-measure of 42% and 71.62%, respectively. However, Improved-BIDAF2 outperformed human performance, achieving an EM and F1-measure of 47% and 78.52%, respectively. This represents a gap of 5% in EM and 6.8% in F1-measure, showing the superiority of Improved-BIDAF2 over human performance on this particular dataset. The second-best results were obtained by Improved-BIDAF1, with an EM and F1-measure of 45% and 73.61%, respectively. Interestingly, human performance ranked third on the 100-record testing set. The authors attribute the improved performance of both Improved-BIDAF1 and Improved-BIDAF2 to the addition of a POS word embedding layer. This layer contributes semantic features to the models, leading to enhanced performance and better comprehension of the data. Overall, the results highlight the effectiveness of the POS word embedding layer in boosting the performance of the models and achieving results that even surpass human-level understanding in some cases.

Indeed, the addition of POS word embeddings to the BIDAF model has proven to be beneficial in enhancing the model's performance. In the context of the highly phonetic nature of the Arabic language, the writing reflects the pronunciation, which can lead to different meanings for homographic words based on their POS tags. This is exemplified by words like (ورد) which can mean "mentioned" if tagged as a verb and "flower" if tagged as a noun. In such cases, character embeddings might not effectively differentiate between the two meanings, while POS word embeddings can capture these semantic nuances, leading to improved comprehension and disambiguation. Moreover, the utilization of POS word embeddings can aid in resolving the out-of-vocabulary problem, where the model may encounter words not present in its training vocabulary. By considering the POS tags, the model can still gain insights into the context and meaning of such OOV words, enhancing its ability to provide meaningful answers.

Additionally, in the improved-BIDAF2 model, the replacement of bi-LSTM with bi-GRU has resulted in performance improvements, specifically the model's EM and F1-measure increased by 2.6% and 1.18%, respectively in the 1,000 record testing set, and by 2% and 4.91%, respectively, in the 100-testing set. Despite LSTM's reputation for performing well with long sequences, the results demonstrated that GRU performed better in this particular scenario. The model achieved higher EM and F1-measure scores on both the 1,000-record and 100-record testing sets, showcasing the effectiveness of using GRU in this context. Overall, the combination of POS word embeddings and bi-GRU has proven to be a successful enhancement in the improved-BIDAF2 model, contributing to its superior performance on the ASER dataset. These improvements allow the model to better understand the complex semantics of the Arabic language and provide more accurate answers.

In order to show the superiority of Improved-BIDAF1 and Improved-BIDAF2 to the baseline models, we have performed statistical evidence using the *t*-test as depicted in Table 7. The first null hypothesis  $H_0$  states that there is no significant performance difference between the Seq2Seq and the Improved-BIDAF1 model. However, the *t*-test resulted in ( $p - value = 0$ ), and since the  $p - value < \alpha(0.05)$ ,  $H_0$  is rejected, and the Seq2Seq population's average is not equal to the Improved-BIDAF1 population's average. As a result, the difference between the averages of Seq2Seq and Improved-BIDAF1 is big enough to be statistically significant. The second null hypothesis  $H_0$  states that there is no significant performance difference between the

**Table 7.** *t*-test statistics results

Models	Null hypothesis $H_0$	<i>t</i> -test result	$H_0$	Population's	
				average	Result
Seq2Seq and improved-BIDAF1 models.	The first states that there is no significant performance difference between the models.	( $p - value = 0$ ), $p - value < \alpha(0.05)$	Rejected	Not equal	The difference between the averages is big enough to be statistically significant.
BIDAF1 and improved-BIDAF1 models.	The second states that there is no significant performance difference between the models.	( $p - value = 0$ ), $p - value < \alpha(0.05)$	Rejected	Not equal	The difference between the averages is big enough to be statistically significant.
BIDAF2 and improved-BIDAF1 models.	The third states that there is no significant performance difference between the models.	( $p - value = 1.004e - 10$ ), $p - value < \alpha(0.05)$	Rejected	Not equal	The difference between the averages is big enough to be statistically significant.
AraBERT BIDAF and improved-BIDAF1 models.	The fourth states that there is no significant performance difference between the models.	( $p - value = 0$ ), $p - value < \alpha(0.05)$	Rejected	Not equal	The difference between the averages is big enough to be statistically significant.
AraBERT BIDAF and improved-BIDAF1 models.	The fifth states that there is no significant performance difference between the models.	( $p - value = 0.7035$ ), $p - value > \alpha(0.05)$	Cannot be rejected	Not equal	The difference between the averages is very small.
Seq2Seq and improved-BIDAF2 models.	The sixth states that there is no significant performance difference between the models.	( $p - value = 0$ ), $p - value < \alpha(0.05)$	Rejected	Not equal	The difference between the averages is big enough to be statistically significant.
BIDAF1 and improved-BIDAF2 models.	The seventh states that there is no significant performance difference between the models.	( $p - value = 0$ ), $p - value < \alpha(0.05)$	Rejected	Not equal	The difference between the averages is big enough to be statistically significant.
BIDAF2 and improved-BIDAF2 models.	The eighth states that there is no significant performance difference between the models.	( $p - value = 1.619e - 13$ ), $p - value < \alpha(0.05)$	Rejected	Not equal	The difference between the averages is big enough to be statistically significant.
AraBERT BIDAF and improved-BIDAF2 models.	The ninth states that there is no significant performance difference between the models.	( $p - value = 0$ ), $p - value < \alpha(0.05)$	Rejected	Not equal	The difference between the averages is big enough to be statistically significant.
Fine-tuned AraBERT BIDAF and improved-BIDAF2 models.	The tenth states that there is no significant performance difference between the models.	( $p - value = 0.7035$ ), $p - value > \alpha(0.05)$	Cannot be rejected	Not equal	The difference between the averages is very small.

BIDAF1 and the Improved-BIDAF1 model. However, the  $t$ -test resulted in ( $p - value = 0$ ), and since the  $p - value < \alpha(0.05)$ ,  $H_0$  is rejected, and the BIDAF1 population's average is not equal to the Improved-BIDAF1 population's average. As a result, the difference between the averages of BIDAF1 and Improved-BIDAF1 is big enough to be statistically significant.

The third null hypothesis  $H_0$  states that there is no significant performance difference between the BIDAF2 and the Improved-BIDAF1 model. However, the  $t$ -test resulted in ( $p - value = 1.004e - 10$ ), and since the  $p - value < \alpha(0.05)$ ,  $H_0$  is rejected, and the BIDAF2 population's average is not equal to the Improved-BIDAF1 population's average. As a result, the difference between the averages of BIDAF2 and Improved-BIDAF1 is big enough to be statistically significant. The fourth null hypothesis  $H_0$  states that there is no significant performance difference between the AraBERT BIDAF and the Improved-BIDAF1 model. However, the  $t$ -test resulted in ( $p - value = 0$ ), and since the  $p - value < \alpha(0.05)$ ,  $H_0$  is rejected, and the AraBERT BIDAF population's average is not equal to the Improved-BIDAF1 population's average. As a result, the difference between the averages of AraBERT BIDAF and Improved-BIDAF1 is big enough to be statistically significant. The fifth null hypothesis  $H_0$  states that there is no significant performance difference between the fine-tuned AraBERT BIDAF and the Improved-BIDAF1 model. However, the  $t$ -test resulted in ( $p - value = 0.7035$ ), and since the  $p - value > \alpha(0.05)$ ,  $H_0$  cannot be rejected, and the fine-tuned AraBERT BIDAF population's average is not equal to the Improved-BIDAF1 population's average. As a result, the difference between the averages of fine-tuned AraBERT BIDAF and Improved-BIDAF1 is very small.

The sixth null hypothesis  $H_0$  states that there is no significant performance difference between the Seq2Seq and the Improved-BIDAF2 model. However, the  $t$ -test resulted in ( $p - value = 0$ ), and since the  $p - value < \alpha(0.05)$ ,  $H_0$  is rejected, and the Seq2Seq population's average is not equal to the Improved-BIDAF2 population's average. As a result, the difference between the averages of Seq2Seq and Improved-BIDAF2 is big enough to be statistically significant. The seventh null hypothesis  $H_0$  states that there is no significant performance difference between the BIDAF1 and the Improved-BIDAF2 model. However, the  $t$ -test resulted in ( $p - value = 0$ ), and since the  $p - value < \alpha(0.05)$ ,  $H_0$  is rejected, and the BIDAF1 population's average is not equal to the Improved-BIDAF2 population's average. As a result, the difference between the averages of BIDAF1 and Improved-BIDAF2 is big enough to be statistically significant.

The eighth null hypothesis  $H_0$  states that there is no significant performance difference between the BIDAF2 and the Improved-BIDAF2 model. However, the  $t$ -test resulted in ( $p - value = 1.619e - 13$ ), and since the  $p - value < \alpha(0.05)$ ,  $H_0$  is rejected, and the BIDAF2 population's average is not equal to the Improved-BIDAF2 population's average. As a result, the difference between the averages of BIDAF2 and Improved-BIDAF2 is big enough to be statistically significant. The ninth null hypothesis  $H_0$  states that there is no significant performance difference between the AraBERT BIDAF and the Improved-BIDAF2 model. However, the  $t$ -test resulted in ( $p - value = 0$ ), and since the  $p - value < \alpha(0.05)$ ,  $H_0$  is rejected, and the AraBERT BIDAF population's average is not equal to the Improved-BIDAF2 population's average. As a result, the difference between the averages of AraBERT BIDAF and Improved-BIDAF2 is big enough to be statistically significant. The tenth null hypothesis  $H_0$  states that there is no significant performance difference between the fine-tuned AraBERT BIDAF and the Improved-BIDAF2 model. However, the  $t$ -test resulted in ( $p - value = 0.7035$ ), and since the  $p - value > \alpha(0.05)$ ,  $H_0$  cannot be rejected, and the fine-tuned AraBERT BIDAF population's average is not equal to the Improved-BIDAF2 population's average. As a result, the difference between the averages of fine-tuned AraBERT BIDAF and Improved-BIDAF2 is very small.

Tables 8 and 9 illustrate EM and F1-measure results, respectively, for human and improved-BIDAF models on the 100-testing set according to the evaluation label. Improved-BIDAF2 outperformed improved-BIDAF1 within the “exact-match” category, it also outperformed human performance within the same category in both EM and F1-measure with an increase of 8.41% and 6.89%, respectively. Within the “Sentence-level paraphrasing” category improved-BIDAF1

Table 8. EM according to evaluation label

Category	Human	Improved-BIDAF1	Improved-BIDAF2
Exact match (54)	45.29%	44.44%	53.7%
Sentence-level paraphrasing (34)	40.91%	47.06%	38.24%
Partial clues (10)	30%	40%	30%
Multiple sentences (2)	31.81%	50%	100%

Table 9. F1-measure according to evaluation label

Category	Human	Improved-BIDAF1	Improved-BIDAF2
Exact match (54)	74.25%	70.94%	81.14%
Sentence-level paraphrasing (34)	68.73%	77.35%	72.35%
Partial clues (10)	67.56%	80%	81.07%
Multiple sentences (2)	69.83%	50%	100%

outperformed improved-BIDAF2 and human performance, while improved-BIDAF2 ranked third according to EM within this category and ranked second according to F1-measure. Within the “partial-clues” category, improved-BIDAF1 outperformed improved-BIDAF2 and human performance within the EM measure, while Improved-BIDAF2 has the highest F1-measure within this category. Improved-BIDAF2 outperformed improved-BIDAF1 and human performance with both EM and F1-measure equal to 100% within the category “multiple-sentences”. The gap difference between improved-BIDAF2 and human performance is significant in this category which reached an EM and F1-measure of 68.19% and 30.17%, respectively.

Tables 10 and 11 provide insights into the EM and F1-measure results for both human and improved-BIDAF models on the 100-testing set, categorized based on domain coverage. In the “technology” domain, both the human performance and improved-BIDAF models achieved an EM of 0%, indicating that neither were able to provide EMs with the golden answers in this domain. On the other hand, human performance reached the highest EM of 81.81% in both the “culture and art” and “coverages” domains. This suggests that humans were able to perform quite well in these domains, with a relatively high level of understanding and comprehension. In the “culture and art” domain, none of the models outperformed human performance, as they all achieved an EM of 0%. However, in the “coverages” domain, both improved models achieved an EM of 100%, indicating that they were able to provide EMs with the golden answer in this domain, outperforming human performance. These results show that while the human performance was strong in certain domains, the improved-BIDAF models were able to excel in the “coverages” domain, where they achieved a perfect match with the golden answer. This highlights the effectiveness of the enhancements made to the BIDAF model, particularly with the addition of POS word embeddings and the replacement of bi-LSTM with bi-GRU, in capturing the intricacies of the language and domain-specific information, leading to improved performance in certain domains.

In the “medicine and health” domain, human performance achieved an EM of 45.45% and an F1-measure of 90.4%. This indicates that humans were able to provide correct answers for approximately 45.45% of the questions in this domain, and the average overlap between their answer and the golden answer was around 90.4%. In contrast, both the human and improved-BIDAF models achieved an EM of 0% in the “medicine and health” domain. This suggests that neither humans nor the models were able to provide an EM with the golden answer for the question

**Table 10.** EM according to domain

Domain category	Translation	Human	Improved-BIDAF1	Improved-BIDAF2
عربي (18)	Arabic	42.93%	55.56%	50%
(18) تقارير وحوارات	Reports and dialogues	37.37%	41.67%	44.44%
(12) ثقافة	Culture	42.42%	58.33%	50%
(12) صحة	Health	46.21%	37.5%	66.67%
(8) منوعات	Mix	29.54%	40%	50%
(5) جولة الصحافة	Press tour	40%	25%	40%
(4) دولي	International	40.9%	50%	25%
(4) القدس	Jerusalem	40.9%	33.33%	50%
(3) اقتصاد	Economy	57.58%	66.67%	33.33%
(3) اقتصاد	Culture and art	81.81%	0%	33.33%
(3) فن	Art	36.36%	0%	33.33%
(2) تكنولوجيا	Technology	0%	0%	0%
(2) سياسة	Politics	22.73%	0%	0%
(1) طب وصحة	Medicine and health	45.45%	0%	0%
(1) حقوق وحرريات	Law and freedom	72.73%	100%	100%
(1) علوم	Science	72.73%	0%	100%
(1) تغطيات	Coverages	81.81%	100%	100%
(1) شخصيات	Figures	9.09%	100%	100%
(1) امرأة	Woman	72.73%	0%	0%

in this domain. However, improved-BIDAF2 reached an F1-measure of 85.71% in the “medicine and health” domain, making it the second-best model in this domain based on the F1-measure. While it did not outperform human performance in terms of F1-measure, it demonstrated a relatively high level of overlap between its answers and the golden answers. Similarly, in the “woman” domain, the human performance achieved an EM of 72.73% and an F1-measure of 80.19%. Again, neither of the models, including improved-BIDAF2, outperformed human performance in this domain. All the models achieved an EM and F1-measure of 0% in this domain, indicating that they were unable to provide EMs or significant overlap with the golden answers.

These results highlight the challenges posed by specific domains, such as “medicine and health” and “woman,” where even the improved-BIDAF models struggled to achieve high performance compared to human understanding. The discrepancies between human performance and the model results in these domains suggest that there may be domain-specific complexities and nuances that are difficult for the models to capture effectively. Further research and fine-tuning of the models may be necessary to improve their performance in such challenging domains.

The comparison between improved-BIDAF1 and improved-BIDAF2 reveals interesting findings regarding their performance in different domains. While both models showed improvements over the baseline BIDAF, there are some variations in their domain-specific performance.

Table 11. F1-measure according to domain

Domain category	Translation	Human	Improved-BIDAF1	Improved-BIDAF2
عربي (18)	Arabic	72.64%	76.72%	77.34%
تقارير وحوارات (18)	Reports and dialogues	67.9%	83.08%	82.57%
ثقافة (12)	Culture	71.5%	67.86%	77.79%
صحة (12)	Health	70.9%	86.34%	88.59%
منوعات (8)	Mix	65.35%	63.99%	82.75%
جولة الصحافة (5)	Press tour	74.81%	77.14%	89.78%
دولي (4)	International	84.1%	72.14%	69.74%
القدس (4)	Jerusalem	53.79%	57.55%	50%
اقتصاد (3)	Economy	76.12%	86.9%	88.57%
ثقافة وفن (3)	Culture and art	92.15%	95.83%	83.08%
فن (3)	Art	64.89%	65.56%	88.24%
تكنولوجيا (2)	Technology	65.5%	11.35%	26.67%
سياسة (2)	Politics	69.21%	38.87%	34.19%
طب وصحة (1)	Medicine and health	90.4%	47.06%	85.71%
حقوق وحریات (1)	Law and freedom	86.83%	100%	100%
علوم (1)	Science	74.21%	25%	100%
تغطیات (1)	Coverages	95.15%	100%	100%
شخصیات (1)	Figures	79.09%	100%	100%
مرأة (1)	Woman	80.19%	0%	0%

Improved-BIDAF2 achieved an EM of 0% in only 4 domains, while improved-BIDAF1 had 7 domains with an EM of 0%. This suggests that improved-BIDAF2 performed better in a larger number of domains, as it had fewer domains with zero EM scores. Furthermore, improved-BIDAF2 achieved an EM and F1-measure of 100% in 4 domains, while improved-BIDAF1 achieved this level of performance in 3 domains. It is noteworthy that none of the domains resulted in 100% EM for human performance, indicating that the models were able to outperform humans in certain specific domains.

The results strongly suggest that replacing bi-LSTM with bi-GRU in the improved-BIDAF2 model led to significant enhancements in performance across different domains. The bi-GRU architecture appears to have better captured the linguistic patterns and context in various domains, enabling improved-BIDAF2 to achieve better results compared to improved-BIDAF1 in multiple scenarios.

These findings highlight the importance of selecting appropriate NN architectures for specific tasks and domains. The replacement of bi-LSTM with bi-GRU in the improved-BIDAF2 model demonstrated its superiority in handling the complexities and nuances of different domains, resulting in more accurate and robust performance across the dataset.

### 9.2. Qualitative analysis

For qualitative analysis, examples having different lengths of passages, questions, and answers were selected from the testing set. Table 12 presents the chosen lengths, the corresponding domain of each example, and the performance of the models. All the selected examples are available and shown in Appendix A, which includes Figures 5–12.

The first example belongs to the “Arabic” domain and consists of a long passage (52 words), a long question (17 words), and a long answer (20 words). Improved BIDAf1 and improved BIDAf2 are the only models that provided the EM of the answer. Noting that “Arabic” is the largest domain of the ASER dataset with a coverage of 18.24 per cent, one will suggest that all models should perform well on larger domains having more training examples. However, both BIDAf1 and BIDAf2 could not provide the correct answer scoring an EM and F1-measure of 0%. On the other hand, when an example was selected from the same domain with long passage (45 words), long question (16 words), and short answer (3 words), BIDAf2 was able to provide parts of the answer, having EM of 0% and F1-measure of 66.7%, and improved BIDAf1 and improved BIDAf2 were the only models to answer this category correctly scoring EM and F1-measure of 100

The third example consists of a long passage (61 words), a short question (4 words), and a long answer (34 words). None of the models were able to provide the correct answer, although the example is from the domain “Health” which is the fourth largest domain within ASER. The bad performance of all models is due to having few words within the question, which makes it difficult for the model to look for the correct answer. However, improved BIDAf2 provided the closest answer with an F1-measure of 98.5%. The fourth example consists of a long passage (forty-two words), a short question (four words), and a short answer (one word) from the domain “Art.” Domain “Art” covers a few training examples within the ASER dataset. Improved BIDAf2 was the only model to provide the correct answer with EM and F1 measures of 100%, while BIDAf1 scored 0% in both measures. Most of the models were able to answer questions of the remaining categories correctly with an EM and F1-measure of 100% except for AraBERT BIDAf before and after fine-tuning and seq2seq models. These categories have short passages and varying lengths of questions and answers.

These results demonstrate that improved BIDAf is capable of effectively handling long passages from diverse and complex sources, making it well-suited for the Arabic language, which is known for its lengthy sentences and intricate structure. The enhancements made to the BIDAf model have also contributed to its ability to generalize well across various questions and domains.

### 9.3. Improved BIDAf and other models

Table 13 provides a comprehensive comparison of improved-BIDAf with other models that were experimented on Arabic text. Despite improved-BIDAf being experimented on a smaller dataset (ASER) compared to the models in the table, it demonstrated superior performance in terms of F1-measure, outperforming QANet, BERT, and BIDAf by considerable margins of 30.82%, 13.92%, and 9.22%, respectively.

In terms of EM, improved-BIDAf also outperformed QANet and BERT, achieving higher scores with gaps of 17.2% and 12.4%, respectively. However, BIDAf managed to outperform improved-BIDAf by a gap of 9.4% in EM. The authors attribute the differences in performance to the variations in the datasets experimented. Notably, the size of the Arabic SQuAD dataset is much larger, containing 70,000 records, whereas ASER comprises only 10,000 records. Additionally, the answer lengths in SQuAD tend to be shorter, with a maximum length of forty-three tokens, while ASER has answers with a maximum length of seventy-five tokens.

The differences in dataset size and answer length likely influence the models’ performance. A larger dataset provides more diverse training examples and potentially allows the models to

**Table 12.** EM and F1 measures for one record having different lengths

Passage	Question	Answer	QID	Question category	Dataset coverage	EM = 100 per cent	F1-measure
Long (52 words)	Long (17 words)	Long (20 words)	9083	Arabic	18.24 per cent	Improved- BIDAF1, improved- BIDAF2	BIDAF1 = 0 per cent, BIDAF2 = 0 per cent, improved-BIDAF1 = 100 per cent, improved BIDAF2 = 100 per cent, fine-tuned AraBERT BIDAF = 33.3 per cent, AraBERT BIDAF = 36.4 per cent, Seq2seq = 48.9 per cent
Long (45 words)	Long (16 words)	Short (3 words)	9013	Arabic	18.24 per cent	Improved- BIDAF1, improved- BIDAF2	BIDAF1 = 23.1 per cent, BIDAF2 = 66.7 per cent, improved-BIDAF1 = 100 per cent, improved BIDAF2 = 100 per cent, improved BIDAF2 = 100 per cent, fine-tuned AraBERT BIDAF = 0 per cent, Seq2seq = 0 per cent
Long (61 words)	Short (4 words)	Long (34 words)	9592	Health	11.97 per cent	None	BIDAF1 = 24.6 per cent, BIDAF2 = 73.9 per cent, improved-BIDAF1 = 24.6 per cent, improved-BIDAF2 = 98.5 per cent, fine-tuned AraBERT BIDAF = 76.5 per cent, AraBERT BIDAF = 97.5 per cent, Seq2seq = 20.8 per cent
Long (42 words)	Short (4 words)	Short (1 word)	9891	Art	2.9 per cent	Improved- BIDAF2	BIDAF1 = 0 per cent, BIDAF2 = 66.7 per cent, improved-BIDAF1 = 66.7 per cent, improved-BIDAF2 = 100 per cent, fine-tuned AraBERT BIDAF = 0 per cent, AraBERT BIDAF = 0 per cent, Seq2seq = 50 per cent
Short (17 words)	Long (10 words)	Long (9 words)	9035	Arabic	18.24 per cent	BIDAF1, BIDAF2, improved- BIDAF1, improved- BIDAF2	BIDAF1 = 100 per cent, BIDAF2 = 100 per cent, improved-BIDAF1 = 100 per cent, improved-BIDAF2 = 100 per cent, fine-tuned AraBERT BIDAF = 0 per cent, AraBERT BIDAF = 0 per cent, Seq2seq = 94.1 per cent
Short (21 words)	Long (15 words)	Short (3 words)	9621	Mix	7.93 per cent	BIDAF1, BIDAF2, improved- BIDAF1, improved- BIDAF2	BIDAF1 = 100 per cent, BIDAF2 = 100 per cent, improved-BIDAF1 = 100 per cent, improved-BIDAF2 = 100 per cent, fine-tuned AraBERT BIDAF = 50 per cent, AraBERT BIDAF = 50 per cent, Seq2seq = 57.1 per cent
Short (17 words)	Short (7 words)	Long (3 words)	9620	Mix	7.93 per cent	BIDAF1, BIDAF2, improved- BIDAF1, improved- BIDAF2	BIDAF1 = 100 per cent, BIDAF2 = 100 per cent, improved-BIDAF1 = 100 per cent, improved-BIDAF2 = 100 per cent, fine-tuned AraBERT BIDAF = 0 per cent, AraBERT BIDAF = 0 per cent, Seq2seq = 78 per cent
Short (13 words)	Short (4 words)	Short (7 words)	9629	Mix	7.93 per cent	BIDAF2, improved- BIDAF1, improved- BIDAF2	BIDAF1 = 33.3 per cent, BIDAF2 = 100 per cent, improved-BIDAF1 = 100 per cent, improved-BIDAF2 = 100 per cent, fine-tuned AraBERT BIDAF = 0 per cent, AraBERT BIDAF = 0 per cent, Seq2seq = 0 per cent

**Table 13.** Comparison of improved-BIDAF with other models

	QANet	BERT	BIDAF	Improved-BIDAF
Dataset	Arabic SQuAD	Arabic SQuAD	Arabic SQuAD	ASER
Dataset size	48,344	48,344	70,000	10,000
EM	29.4%	34.2%	56%	46.6%
F1	44.4%	61.3%	66%	75.22%

generalize better to unseen data. Moreover, shorter answer lengths in SQuAD may facilitate easier comprehension and extraction, whereas longer answer lengths in ASER present additional challenges for the models.

Despite these differences, improved-BIDAF demonstrated its strength in handling the ASER dataset, outperforming other models in terms of F1-measure, and achieving competitive results in EM. The performance differences underscore the significance of dataset characteristics and demonstrate the capability of improved-BIDAF in tackling the complexities of the ASER dataset with longer answer lengths.

#### 9.4. Limitations and future directions

This study has two primary limitations. First, while the ASER dataset comprises 10,000 records, it remains relatively small compared to datasets available in other languages. This limitation may have constrained the generalizability of current findings to a larger dataset. Unfortunately, the scarcity of dependable large datasets for MRC tasks in the Arabic language means this limitation cannot be easily addressed.

Second, we encountered a lack of preexisting embeddings for POS tags in the Arabic language trained on substantial datasets, necessitating the creation of a dataset for POS tags and subsequent embeddings.

Moving forward, future research could aim to address these limitations by exploring the study's concepts on a larger, more diverse dataset. Additionally, experimentation with an improved BIDAF model could be conducted for classification and plagiarism tasks.

### 10. Conclusion

In this research paper, the authors proposed an enhanced version of the BIDAF model specifically designed for the Arabic MRC task. The model was evaluated on the ASER dataset, an Arabic Span-Extraction Reading Comprehension Benchmark. The improvements to the BIDAF model involved two key modifications: replacing the character embedding layer with a POS word embedding layer and substituting the bi-LSTM with bi-GRU.

Experimental results demonstrated a significant performance gap between the baseline models experimented on ASER and the improved-BIDAF model. The improved-BIDAF model showcased superior performance, achieving higher EM and F1-measure scores compared to the baseline models. Notably, the model even surpassed human performance in this specific task, achieving an EM and F1-measure increase of 5% and 6.9%, respectively, over human performance.

The use of POS word embeddings in the BIDAF model was instrumental in enhancing its performance. POS tags provide semantic meaning to words, and incorporating this information through the POS word embeddings allowed the model to better capture the nuances and context of the Arabic language, leading to improved comprehension and more accurate answers.

The study's findings highlight the potential of deep learning models in outperforming human performance in certain tasks, such as MRC. By effectively leveraging advanced NN architectures and linguistic features like POS word embeddings, these models can exhibit remarkable capabilities in understanding and processing complex natural language data. The results encourage further exploration and development of deep learning approaches for natural language understanding tasks, specifically for the Arabic language.

**Competing interests.** The author(s) declare none.

## References

- Aftab E. and Malik M.K. (2022). erock at qur'an qa 2022: Contemporary deep neural networks for qur'an based reading comprehension question answers. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pp. 96–103.
- Alian M. and Al-Naymat G. (2022). Questions clustering using canopy-k-means and hierarchical-k-means clustering. *International Journal of Information Technology* **14**(7), 3793–3802.
- Alkhatnai M., Amjad H.I., Amjad M., Gelbukh A. (2020). Methods and trends of machine reading comprehension in the arabic language. *Computación y Sistemas* **24**(4), 1607–1615.
- Alnefaie S., Atwell E. and Alsalka M.A (2023). Haqa and quqa: Constructing two arabic question-answering corpora for the quran and hadith. In *Proceedings of the Conference Recent Advances in Natural Language Processing-Large Language Models for Natural Language Processings*, Shoumen, BULGARIA: INCOMA Ltd, pp. 90–97.
- Atef A., Mattar B., Sherif S., Elrefai E. and Torki M. (2020). Aqad: 17,000+ arabic questions for machine comprehension of text. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, pp. 1–6.
- Baradaran R., Ghiasi R., Amirkhani H. (2022). A survey on machine reading comprehension systems. *Natural Language Engineering* **28**(6), 683–732.
- Biltawi M., Awajan A. and Tedmori S. (2020a). Towards building an open-domain corpus for Arabic reading comprehension. In *Proc. 35th International Business Information Management Association (IBIMA)*, pp. 1–27.
- Biltawi M., Awajan A. and Tedmori S. (2020b). Arabic reading comprehension benchmarks created semiautomatically. In *2020 21st International Arab Conference on Information Technology (ACIT)*, IEEE, pp. 1–6.
- Biltawi M.M., Tedmori S. and Awajan A. (2021). Arabic question answering systems: gap analysis. *IEEE Access* **9**, 63876–63904.
- Biltawi M.M., Awajan A., Tedmori S. (2023). Arabic span extraction-based reading comprehension benchmark (aser) and neural baseline models. *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**(5), 1–29.
- Chen D. (2018). *Neural Reading Comprehension and Beyond*. Stanford University, PhD thesis.
- Dong R., Wang X., Dong L. and Zhang Z. (2023). Multi-passage extraction-based machine reading comprehension based on verification sorting. *Computers and Electrical Engineering* **106**, 108576.
- Gupta S., Rawat B.P.S. and Yu H. (2020). Conversational machine comprehension: a literature review. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2739–2753.
- Hermann K., Kočiský T., Grefenstette E., Espeholt L., Kay W., Suleyman M. and Blunsom P. (2015). Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems* **28**, 1693–1701.
- Hu M., Wei F., Peng Y., Huang Z., Yang N. and Li D. (2019). Read+ verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, pp. 6529–6537.
- Kenton J.D.M.-W.C. and Toutanova L.K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186.
- Li X., Cheng G., Chen Z., Sun Y. and Qu Y. (2022). Adalogn: Adaptive logic graph network for reasoning-based machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7147–7161.
- Liu S., Zhang X., Zhang S., Wang H. and Zhang W. (2019). Neural machine reading comprehension: methods and trends. *Applied Sciences* **9**(18), 3698.
- Malhas R. and Elsayed T. (2022). Arabic machine reading comprehension on the Holy Qur'an using CL-AraBERT. *Information Processing & Management* **59**(6), 103068.
- Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv: <https://arxiv.org/abs/1301.3781>.
- Mozannar H., Maamary E., Hajal K.E. and Hajj H. (2019). Neural arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pp. 108–118.
- Nguyen T., Rosenberg M., Song X., Gao J., Tiwary S., Majumder R. and Deng L. (2016). Ms marco: a human generated machine reading comprehension dataset. *Choice* **2640**, 660.

- Patel K. and Bhattacharyya P.** (2017). Towards lower bounds on number of dimensions for word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 31–36.
- Rajpurkar P., Zhang J., Lopyrev K. and Liang P.** (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392.
- Richardson M., Burges C.J.C. and Renshaw E.** (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 193–203.
- Seo M., Kembhavi A., Farhadi A. and Hajishirzi H.** (2016). Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*. 2022. arXiv preprint [arXiv:1611.01603](https://arxiv.org/abs/1611.01603).
- Soliman A.B., Eissa K. and El-Beltagy S.R.** (2017). Aravec: a set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science* 117, 256–265.
- Toutanova K., Klein D., Manning C.D. and Singer Y.** (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 252–259.
- Trischler A., Wang T., Yuan X., Harris Justin, Sordoni A., Bachman P. and Suleman K.** (2017). Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. pp. 191–200. Association for Computational Linguistics.
- Weissenborn D., Wiese G. and Seiffe L.** (2017). Making neural qa as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 271–280.
- Welbl J., Liu N.F. and Gardner M.** (2017). Crowdsourcing multiple choice science questions. *W-NUT 2017*, 94.
- Xin Z., An Y., Sujian L. and Yizhong W.** (2019). Machine reading comprehension: a literature review, arXiv preprint [arXiv:1907.01686](https://arxiv.org/abs/1907.01686).
- Zeng C., Li S., Li Q., Hu J. and Hu J.** (2020). A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences* 10(21), 7640.

Appendix A. Examples from ASER dataset with the answers provided by the experimented models

<p><b>Question:</b> ما الذي ينبغي على المجتمع الدولي عمله بناء على ما قاله الناشئ في لقائه مع الجزيرة نت؟</p> <p><b>Passage:</b> وطالب الناشئ في لقاء خاص مع الجزيرة نت المجتمع الدولي بتحمل مسؤولياته وضمان إجراء الانتخابات في موعدها المحدد، وأضاف "من الآن وحتى حلول موعد الانتخابات نحن على جاهزية تامة، وعلى المجتمع الدولي أن يجهز لنا أرضنا ويتدخل ليسهل الاتصالات بين أبناء شعبنا حتى نتتمكن من إجراء الانتخابات البلدية والتشريعية والرئاسية حسب القوانين."</p> <p><b>Gold answer:</b> أن يجهز لنا أرضنا ويتدخل ليسهل الاتصالات بين أبناء شعبنا حتى نتمكن من إجراء الانتخابات البلدية والتشريعية والرئاسية حسب القوانين</p>	<p><b>BIDAF1:</b> المجتمع الدولي يتحمل مسؤولياته</p>
	<p><b>BIDAF2:</b> يتحمل مسؤولياته</p>
	<p><b>Improved BIDAF1:</b> ان يجهز لنا ارضنا ويتدخل ليسهل الاتصالات بين ابناء شعبنا حتي نتمكن من اجراء الانتخابات البلدية والتشريعية والرئاسيه حسب القوانين</p>
	<p><b>Improved BIDAF2:</b> ان يجهز لنا ارضنا ويتدخل ليسهل الاتصالات بين ابناء شعبنا حتي نتمكن من اجراء الانتخابات البلدية والتشريعية والرئاسيه حسب القوانين</p>
	<p><b>Finetuned AraBERT-BIDAF:</b> جاهزيه تامه وعلي المجتمع الدولي ان يجهز لنا ارضنا ويتدخل</p>
	<p><b>AraBERT-BIDAF:</b> من الان وحتى حلول موعد الانتخابات نحن علي جاهزيه تامه وعلي المجتمع الدولي ان يجهز لنا ارضنا ويتدخل</p>
	<p><b>Seq2seq:</b> موعدھا المحدد واضاف من الان وحتى حلول موعد الانتخابات نحن علي جاهزيه تامه وعلي المجتمع الدولي ان يجهز لنا ارضنا ويتدخل ليسهل الاتصالات بين ابناء</p>

Figure 5. Example from ASER: Long passage, long question with long answer.

<p><b>Question:</b> من هو المتحدث الذي اعلن عبر صفحته على فيسبوك السيطرة على مقر الفرقة المدرعة الاولى المنحلة؟</p> <p><b>Passage:</b> وأشارت مصادر متطابقة إلى أن الحوثيين سيطروا كذلك على وزارة الإعلام ووزارة الصحة.وفي وقت لاحق، أعلن المتحدث باسم الحوثيين محمد عبد السلام عبر صفحته على فيسبوك السيطرة على مقر الفرقة المدرعة الاولى المنحلة، أي مقر اللواء علي محسن الأحمر الذي يبدو أنه تمكن من الفرار.</p> <p><b>Gold answer:</b> محمد عبد السلام</p>	<p><b>BIDAF1:</b> المتحدث باسم الحوثيين محمد عبد السلام عبر صفحته علي فيسبوك السيطرة علي مقر الفرقة المدرعه الاولى المنحلة اي مقر اللواء علي محسن الاحمر</p>
	<p><b>BIDAF2:</b> المتحدث باسم الحوثيين محمد عبد السلام</p>
	<p><b>Improved BIDAF1:</b> محمد عبد السلام</p>
	<p><b>Improved BIDAF2:</b> محمد عبد السلام</p>
	<p><b>Finetuned AraBERT-BIDAF:</b> No answer provided.</p>
	<p><b>AraBERT-BIDAF:</b> يبدو انه تمكن من</p>
	<p><b>Seq2seq:</b> الصحة وفي وقت</p>

Figure 6. Example from ASER: Long passage, long question with short answers.

<p><b>Question:</b> ماذا قال الخبير الألماني؟</p> <p><b>Passage:</b> وأكد الخبير الألماني ضرورة أن يتخلى هواة ممارسة الرياضة عن الاعتقاد بأن ممارسة الأنشطة الحركية في الصباح دون تناول الطعام تمكنهم من إنقاص وزنهم بشكل جيد، وقال "يعتقد الكثيرون أن ممارسة الرياضة في الصباح الباكر تعمل على حرق مخزون الدهون في الجسم، لأن مخزون السكر يكون قد تم حرقه بالفعل خلال الليل، ويكون جسم الإنسان خالياً من الطعام في هذا الوقت".</p> <p><b>Gold answer:</b> "يعتقد الكثيرون أن ممارسة الرياضة في الصباح الباكر تعمل على حرق مخزون الدهون في الجسم، لأن مخزون السكر يكون قد تم حرقه بالفعل خلال الليل، ويكون جسم الإنسان خالياً من الطعام في هذا الوقت".</p>	<p><b>BIDAF1:</b> ضروره ان يتخلي هواه ممارسه الرياضه عن الاعتقاد بان ممارسه الانشطه الحركيه في الصباح دون تناول الطعام تمكنهم من انقاص وزنهم بشكل جيد</p>
	<p><b>BIDAF2:</b> ضروره ان يتخلي هواه ممارسه الرياضه عن الاعتقاد بان ممارسه الانشطه الحركيه في الصباح دون تناول الطعام تمكنهم من انقاص وزنهم بشكل جيد وقال يعتقد الكثيرون ان ممارسه الرياضه في الصباح الباكر تعمل علي حرق مخزون الدهون في الجسم لان مخزون السكر يكون قد تم حرقه بالفعل خلال الليل ويكون جسم الانسان خاليا من الطعام في هذا الوقت</p>
	<p><b>Improved BIDAF1:</b> ضروره ان يتخلي هواه ممارسه الرياضه عن الاعتقاد بان ممارسه الانشطه الحركيه في الصباح دون تناول الطعام تمكنهم من انقاص وزنهم بشكل جيد</p>
	<p><b>Improved BIDAF2:</b> الكثيرون ان ممارسه الرياضه في الصباح الباكر تعمل علي حرق مخزون الدهون في الجسم لان مخزون السكر يكون قد تم حرقه بالفعل خلال الليل ويكون جسم الانسان خاليا من الطعام في هذا الوقت</p>
	<p><b>Finetuned AraBERT-BIDAF:</b> الرياضه عن الاعتقاد بان ممارسه الانشطه الحركيه في الصباح دون تناول الطعام تمكنهم من انقاص وزنهم بشكل جيد وقال يعتقد الكثيرون ان ممارسه الرياضه في الصباح الباكر تعمل علي حرق مخزون الدهون في الجسم لان مخزون السكر يكون قد تم حرقه بالفعل خلال الليل ويكون جسم الانسان</p>
	<p><b>AraBERT-BIDAF:</b> الرياضه عن الاعتقاد بان ممارسه الانشطه الحركيه في الصباح دون تناول الطعام تمكنهم من انقاص وزنهم بشكل جيد وقال يعتقد الكثيرون ان ممارسه الرياضه في الصباح الباكر تعمل علي حرق مخزون الدهون في الجسم لان مخزون السكر يكون قد تم حرقه بالفعل خلال الليل ويكون جسم الانسان</p>
	<p><b>Seq2seq:</b> ضروره ان يتخلي هواه ممارسه الرياضه عن الاعتقاد بان ممارسه الانشطه الحركيه في الصباح</p>

Figure 7. Example from ASER: Long passage, short question with long answer.

<p><b>Question:</b> ما اسم نجل الحلاني؟</p> <p><b>Passage:</b> وعودة إلى الحلاني، فإن نجله الوليد شارك في برنامج على إحدى المحطات اللبنانية "صدفة"، وفاز بالمركز الأول "صدفة"، وأصدر أغنيته الخاصة ونزلت للأسواق، وبات الحديث عن صوته وإحساسه و"نبوغه الفني"، وتوافق فترة صعوده حملة ترويج كبيرة في الإعلام اللبناني ومواقع التواصل الاجتماعي.</p> <p><b>Gold answer:</b> الوليد</p>	<p><b>BIDAF1:</b> واصدر اغنيته الخاصه</p>
	<p><b>BIDAF2:</b> نجله الوليد</p>
	<p><b>Improved BIDAF1:</b> نجله الوليد</p>
	<p><b>Improved BIDAF2:</b> الوليد</p>
	<p><b>Finetuned AraBERT-BIDAF:</b> علي احدي المحطات اللبنانيه صدفة وفاز بالمركز الاول صدفة واصدر اغنيته الخاصه ونزلت للأسواق وبات الحديث عن صوته واحساسه ونبوغه الفني وتوافق فترة صعوده حملة ترويج</p>
	<p><b>AraBERT-BIDAF:</b> علي احدي المحطات اللبنانيه صدفة وفاز بالمركز الاول صدفة واصدر اغنيته الخاصه ونزلت للأسواق وبات الحديث عن صوته واحساسه ونبوغه الفني وتوافق فترة صعوده حملة ترويج</p>
	<p><b>Seq2seq:</b> فان نجله الوليد</p>

Figure 8. Example from ASER: Long passage, short question with short answer.

<p><b>Question:</b> ماذا وصف مصدر مسؤول في الخارجية الكويتية عن الاجراء الاسرائيلي؟</p> <p><b>Passage:</b> ووصف مصدر مسؤول في الخارجية الكويتية الاجراء الاسرائيلي بأنه استفزاز لمشاعر المسلمين وانتهاك لحرية ممارسة الشعائر الدينية.</p> <p><b>Gold answer:</b> بأنه استفزاز لمشاعر المسلمين وانتهاك لحرية ممارسة الشعائر الدينية.</p>	<b>BIDAF1:</b> بأنه استفزاز لمشاعر المسلمين وانتهاك لحرية ممارسه الشعائر الدينية
	<b>BIDAF2:</b> بأنه استفزاز لمشاعر المسلمين وانتهاك لحرية ممارسه الشعائر الدينية
	<b>Improved BIDAF1:</b> بأنه استفزاز لمشاعر المسلمين وانتهاك لحرية ممارسه الشعائر الدينية
	<b>Improved BIDAF2:</b> بأنه استفزاز لمشاعر المسلمين وانتهاك لحرية ممارسه الشعائر الدينية
	<b>Finetuned AraBERT-BIDAF:</b> No answer provided.
	<b>AraBERT-BIDAF:</b> No answer provided.
	<b>Seq2seq:</b> استفزاز لمشاعر المسلمين وانتهاك لحرية ممارسه الشعائر الدينية

Figure 9. Example from ASER: Short passage, long question with long answer.

<p><b>Question:</b> كم كان يبلغ متوسط عمر المرأة الأمريكية - التي تحمل لأول مرة- في عام 1970؟</p> <p><b>Passage:</b> ويبلغ متوسط عمر المرأة الأميركية -التي تحمل لأول مرة حاليا- نحو 26 عاما، بينما عام 1970 كان يصل إلى 21.4 عاما.</p> <p><b>Gold answer:</b> نحو 26 عاما</p>	<b>BIDAF1:</b> نحو 26 عاما
	<b>BIDAF2:</b> نحو 26 عاما
	<b>Improved BIDAF1:</b> نحو 26 عاما
	<b>Improved BIDAF2:</b> نحو 26 عاما
	<b>Finetuned AraBERT-BIDAF:</b> عاما
	<b>AraBERT-BIDAF:</b> عاما
	<b>Seq2seq:</b> مره حاليا نحو 26

Figure 10. Example from ASER: Short passage, long question with short answer.

<p><b>Question:</b> على كم تمساح تم القضاء خلال الليل؟</p> <p><b>Passage:</b> وتم القضاء على تمساحين اثنين خلال الليل ولكن بعد شقهما اتضح أنهما لم يبتلعا أي كائن بشري.</p> <p><b>Gold answer:</b> على تمساحين اثنين</p>	<b>BIDAF1:</b> على تمساحين اثنين خلال الليل ولكن بعد شقهما اتضح انهما لم يبتلعا اي كائن بشري
	<b>BIDAF2:</b> على تمساحين اثنين
	<b>Improved BIDAF1:</b> على تمساحين اثنين
	<b>Improved BIDAF2:</b> على تمساحين اثنين
	<b>Finetuned AraBERT-BIDAF:</b> No answer provided.
	<b>AraBERT-BIDAF:</b> كائن بشري
	<b>Seq2seq:</b> بعد شقهما اتضح

Figure 11. Example from ASER: Short passage, short question with short answer.

<p>Question: بماذا سيقوم القمر الصناعي؟</p> <p>Passage: وأوضحت الصحيفة أن القمر الاصطناعي سيقوم باستكشاف الموارد الطبيعية وبنية طبقات سطح القمر .</p> <p>Gold answer: باستكشاف الموارد الطبيعية وبنية طبقات سطح القمر.</p>	BIDAF1:
	باستكشاف الموارد الطبيعية وبنية طبقات سطح القمر
	BIDAF2:
	باستكشاف الموارد الطبيعية وبنية طبقات سطح القمر
	Improved BIDAF1:
	باستكشاف الموارد الطبيعية وبنية طبقات سطح القمر
	Improved BIDAF2:
	باستكشاف الموارد الطبيعية وبنية طبقات سطح القمر
	Finetuned AraBERT-BIDAF:
	No answer provided.
	AraBERT-BIDAF:
	No answer provided.
	Seq2seq:
	الاصطناعي سيقوم باستكشاف الموارد الطبيعية وبنية طبقات سطح

Figure 12. Example from ASER: Short passage, short question with long answer.