Applied Intelligence (2021) 51:8484–8497
https://doi.org/10.1007/s10489-021-02348-9

ORIGINAL SUBMISSION

Automatic question-answer pairs generation and question similarity mechanism in question answering system

Shivani G. Aithal[1] · **Abishek B. Rao**[1] · **Sanjay Singh**[1]

Accepted: 11 March 2021 / Published online: 7 April 2021

Abstract

With the swift growth of the information over the past few years, taking full benefit is increasingly essential. Question Answering System is one of the promising methods to access this much information. The Question Answering System lacks humans' common sense and reasoning power and cannot identify unanswerable questions and irrelevant questions. These questions are answered by making unreliable and incorrect guesses. In this paper, we address this limitation by proposing a Question Similarity mechanism. Before a question is posed to a Question-Answering system, it is compared with possible generated questions of the given paragraph, and then a Question Similarity Score is generated. The Question Similarity mechanism effectively identifies the unanswerable and irrelevant questions. The proposed Question Similarity mechanism incorporates a human way of reasoning to identify unanswerable and irrelevant questions. This mechanism can avoid the unanswerable and irrelevant questions altogether from being posed to the Question Answering system. It helps the Question Answering Systems to focus only on the answerable questions to improve their performance. Along with this, we introduce an application of the Question Answering System that generates the question-answer pairs given a passage and is useful in several fields.

Keywords Question answering · Question generation · Question similarity · Universal sentence encoder · Question comprehension

1 Introduction

The Question Answering System (QAS) plays an important role in getting questions and automatically answering them using a knowledge information system. This paper blends the essence of Question Generation, Question Comprehension, and Question Answering to overcome the Question Answering System's limitations.

Question Answering System had existed way back in the 1960s. The first-ever question answering system introduced was BASEBALL [16]. It was built with a sequence of handwritten rules, and all baseball figures were stored in a database accumulated over the year. Later, LUNAR [42] was introduced during the Apollo mission to answer questions. This system was built to answer the moon's geological patterns and other related information about the APOLLO mission. The customized nature of this system leads to the generation of highly accurate answers.

As the research evolved, the Question Answering System started gaining higher credibility due to data outbursts. The Natural Language Processing (NLP) systems were introduced to reach realistic language understanding [18]. Using the NLP concept, there has been significant research in Question Answering Systems during the past four decades. Early examples of primordial NLP systems are ELIZA [40], SHRDLU [41], which were developed to understand language between humans and machines. Although ELIZA was closer to a human conversation but was much less intelligent and knew almost nothing. SHRDLU on the other hand, was able

✉ Sanjay Singh
sanjay.singh@manipal.edu

Shivani G. Aithal
kshivaniaithal19@gmail.com

Abishek B Rao
abishekbrao1996@gmail.com

[1] Department of Information and Communication Technology, Manipal Institute of Technology, MAHE, Manipal 576104, India

to reason about the block world. Although the conversation was limited to the block's world, so not convincingly human-like, it does know what it is talking about.

Later in the year, 2011 IBM Watson [15] gained worldwide attention, which uses NLP to analyze human speech for meaning and syntax. Way back, it was commonly referred to as a brain. In recent years, search engines (Google), chatbots (SIRI, ALEXA, and CORTANA) are becoming better at going beyond by answering the exact answer to our question. The Question Answering System has also seen significant changes in the architecture from basic Recurrent Neural Network (RNN) to transformers [8, 12] over the years.

The Question Answering System is classified into an Open-domain Question Answering System, and Closed-domain Question Answering System [24]. The open-domain question answering systems like [10, 17] can handle nearly any questions based on world knowledge. This type of Question Answering System has access to more data to extract the answer. The closed-domain question answering systems are domain-specific [2, 9, 45]. Closed-domain question answering systems answers from either a pre-structured database or the collection of domain-specific natural language documents.

According to the studies [32, 33], the human accuracy of answering the question is 89.45%, and the state-of-the-art Question Answering System's accuracy is 93.01%. Although system accuracy exceeds human accuracy, such a Question-Answering system lacks reasoning power as humans do [30, 34, 44], to identify the questions and understand them. The SQUAD 2.0 dataset [31] provides unanswerable questions with plausible answers; however, identifying the unanswerable question remains a challenge.

The limitations of the question answering system are:

- **Unanswerable Questions:** A question that is incorrect and related to the context is posed to the Question Answering System. The Question Answering System, which has outstripped human accuracy, should know that the question is unanswerable and should not generate the answer. However, the SQUAD 1.1 dataset models answer such unanswerable questions by unreliable guesses on questions for which the correct answer is not stated. It indicates that these models lack a rational way of reasoning. Even though the SQUAD 2.0 dataset introduced unanswerable questions in the dataset, identifying the unanswerable questions remains unsolved.
- **Irrelevant Questions:** When the Question Answering System is posed with irrelevant questions that are out of context, the system still generates an understandable but nonsensical answer. On the other hand, humans do not provide such nonsensical answers; instead, they will identify that the question is irrelevant and out of context.

The contributions of this paper are as follows:

1. To automatically generate the possible question-answer pairs, given a passage.
2. We introduce a Question Similarity mechanism, where it will identify the unanswerable and irrelevant questions.
3. We combine the Question Generation System with Question Answering System to create an application called Automatic Question-Answer Pairs Generation System.

Rest of the paper is organized as follows. Section 2 provides the related work on Question Answering Systems. Section 3 explains about the automatic question-answer pair generation, and question similarity mechanism. Section 4 provides details about the datasets used and the experiments. The experimental results are presented in Section 5, and Section 6 discusses about the results. Finally, in Section 7 we conclude this paper.

2 Related works

In recent years, several works are proposed to tackle world knowledge by combining search factors based on bi-gram hashing, TF-IDF matching [7] and machine reading comprehension [22, 29]. It brought the Question Answering System a good beginning. The most recent QAS is the Bidirectional Encoder Representations from Transformers (BERT) [11]. It uses neural models such as transformers to pre-train the large corpora of data. Such a latest refinement has led to remarkable gains in NLP tasks such as Question Answering, Text Summarization, and many classification problems. Besides BERT, for a broad range of applications, researchers have lately exhibited the efficiency of neural models using pretraining language modeling by taking BERT as a base model. By combining different neural architectures with the BERT language model and exploiting its embeddings, cutting-edge results in English has been achieved [5]. BERT model with the advancement of the research, a few systems such as the end-to-end interactive chatbot system like BERTserini [43], a lighter version of BERT called ALBERT [21], and an all-purpose language model called DistilBERT [36] were introduced.

The model is trained on a specific dataset after pre-training with a large corpus of data to answer the questions either in an open-domain or closed-domain question answering system. There are few datasets for question answering systems such as the CuratedTREC dataset [1], WebQuestions dataset [3] that answer questions from Freebase [4], and the Stanford Question Answering Dataset (SQuAD) [33], which is based on Wikipedia knowledge source.

The SQuAD is one of the most significant general-purpose Open-domain Question Answering datasets currently available among all these datasets. There are two

Table 1 Examples showing unanswerable and irrelevant questions resulting in incorrect answers

SQUAD 2.0	
Passage	From Italy, the disease spread northwest across Europe, striking France, Spain, Portugal, and England by June 1348, then turned and spread east through Germany and Scandinavia from 1348 to 1350. It was introduced in Norway in 1349 when a ship landed at Askoy, then spread to Bjorgvin (modern Bergen) and Iceland. Finally, it spread to northwestern Russia in 1351. The plague was somewhat less common in parts of Europe that had smaller trade relations with their neighbors, including the Kingdom of Poland, the majority of the Basque Country, isolated parts of Belgium and the Netherlands, and isolated alpine villages throughout the continent.
Q_1: Unanswerable Question	In what month and year did the disease spread into the Kingdom of Poland?
BERT model output (incorrect)	june 1348
Q_2: Irrelevant Question	Which is the UK's largest digital subscription television company?
BERT model output (incorrect)	which

versions of SQuAD dataset: SQuAD 1.1 [33] and SQuAD 2.0 [32]. The dataset SQuAD 2.0 contains unanswerable questions with plausible answers in addition to the SQuAD 1.1 dataset.

However, as seen in Table 1, when unanswerable and irrelevant questions are asked to the system, the model would make unreliable and incorrect guesses and answers to such questions.

Along with the Question Answering System (QAS) side, the Question Generation System (QGS) plays a vital role in making the model understand the question and answer it. According to Sun et al. [39], there is a close relation between Question Answering and Question Generation. The question generation task has seen many training objectives. Works such as [13, 25, 37] does not capture long-term dependencies but concentrate on the most recent tokens. Even though these papers provide a good result, these works lack capturing long-term dependencies [19, 22]. The work proposed by Qi et al. [29] has a future n-gram as a training objective, thus providing excellent results in question generation tasks.

When we extensively tested the Question Answering System keeping in mind how the answer is generated, it is found that Question Comprehension plays a significant role in the question answering system [38]. Also, systems like [46] introduce a pair-to-sequence model that captures the interaction between the question asked and the given paragraph. Specific systems like ParaQG [20] try to generate the questions from the paragraph. Systems like [35] pick up the keywords from the question and paragraph and match them using RNN. Pota et al., [27] used Convolution Neural Networks (CNNs) to classify the questions. The question classification plays a vital role in extracting the correct answer in the Question Answering System. The method proposed by Esposito et al., [14] extracts the most relevant terms from the questions, and then these words are placed in the context. This document collection is later used in the QA system. Some other work like [28] uses Part of Speech (POS) tagging based on a deep neural network. Here the POS is tagged at the character level, and then it is eventually fed to Bi-LSTM. This method handles rare and Out-of-Vocabulary words as well as common and known words.

3 Methodology

This section introduces an automatic Question-Answer pairs generation system, a combination of Question Answering System and Question Generation System. To address the limitations of the Question Answering System, we propose a Question Similarity mechanism. The possible generated questions are from the state-of-the-art question generation

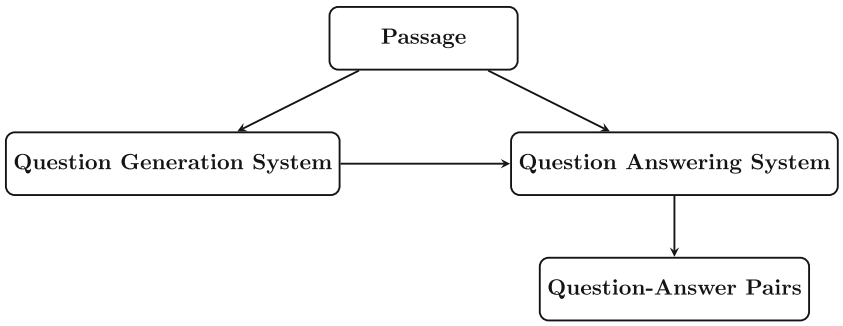

Fig. 1 Block diagram depicting Question-Answer pairs generation system

system called ProphetNet [29] and the Question posed is from the SQuAD 2.0 dataset. The Question Similarity mechanism calculates the cosine similarity between the possible generated questions from the given paragraph and the question posed.

3.1 Automatic question-answer pairs generation system

The automatic question-answer pairs generation system uses pre-trained weights of a state-of-the-art question generation system called ProphetNet [29] to generate the questions, and BERT [11] model to generate the answers for the generated questions.

As shown in Fig. 1, first, we provide the passage as input to both the question generation system and answering system. Once the question generation system generates the possible set of questions based on the answer spans, which are found by a noun and verb phrases in the passage, the generated questions are given to the question answering system. The question answering system based on the passage and the set of generated questions generates the answers. Finally, we get the Question-Answer pairs from this system.

3.2 Question similarity mechanism

In addition to automatically generating Question-Answer pairs, if additional questions are posed to the system, such questions are identified either as answerable or unanswerable and irrelevant before passing it to the Question Answering System. To identify the questions, we introduce a mechanism called a Question Similarity mechanism. This mechanism calculates the cosine similarity between the generated questions and the question posed.

As shown in Fig. 2, the passage is initially passed to the Question Generation System to generate the possible set of questions on the given paragraph based on the answer spans derived on the noun and verb phrases.

Let GQ and QP be the set of generated questions and the question posed with $|GQ| = m$ and $|QP| = 1$. The sentence embeddings for the generated questions is obtained using Universal Sentence Encoder [6], which gives better results than the pre-trained word embeddings such as those produced by GloVe [26] and word2vec [23] and it is given by,

$$X_{SE}^{GQ} = \{E_{GQ}^{(i)} \in \mathbb{R}^{512}; i = 1, \ldots, m\}. \quad (1)$$

where,

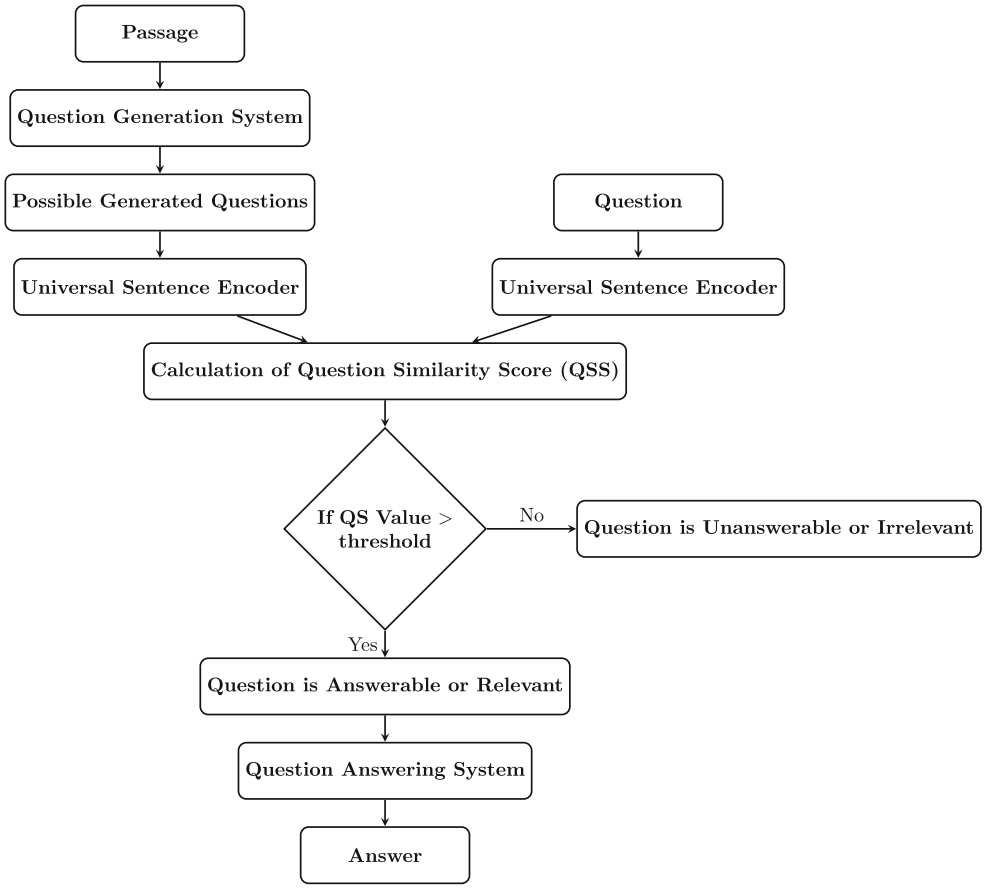

Fig. 2 Block diagram to identify the unanswerable or irrelevant questions

- X_{SE}^{GQ} is the set of Sentence Embeddings (SE) for the Generated Questions (GQ), and
- E_{GQ} is the sentence embeddings for each Generated Question (GQ).

Similarly, we obtain the sentence embeddings for the question posed as

$$X_{SE}^{QP} = E_{QP}^{(i)} \in \mathbb{R}^{512}; i = 1. \tag{2}$$

where,

- X_{SE}^{QP} is the set of Sentence Embeddings (SE) for the Questions Posed (QP), and
- E_{QP} is the sentence embeddings for each Question Posed (QP).

The cosine similarity between the generated questions and the question posed is computed as per the (3).

$$\begin{aligned} \text{Cosine Similarity}(E_{GQ}^{(i)}, X_{SE}^{QP}) &= \cos(E_{GQ}^{(i)}, X_{SE}^{QP}) \\ &= \frac{\langle E_{GQ}^{(i)}, X_{SE}^{QP} \rangle}{||E_{GQ}^{(i)}||||X_{SE}^{QP}||}, \quad i = 1, \ldots, m \end{aligned} \tag{3}$$

where $\langle E_{GQ}^{(i)}, X_{SE}^{QP} \rangle$ denotes the inner product of $E_{GQ}^{(i)}$, and X_{SE}^{QP}.

To calculate Question Similarity Score (QSS), we need to identify the question among the generated questions, whose cosine similarity is highest with respect to the posed question. We call it as *Highest Similarity Score Question*, and it is obtained by (4).

$$\text{Highest Similarity Score Question} = \underset{i \in \{1,\ldots,m\}}{argmax} \cos(E_{GQ}^{(i)}, X_{SE}^{QP}). \tag{4}$$

Now, the Question Similarity Score between the generated question (identified as per the (4)) and the question posed is given by,

$$\text{Question Similarity Score}(E_{GQ}^{(j)}, X_{SE}^{QP}) = \cos(E_{GQ}^{(j)}, X_{SE}^{QP}) \tag{5}$$

where $E_{GQ}^{(j)}$, and X_{SE}^{QP} are the sentence embeddings for the jth generated questions (as obtained by (4)) and the question posed respectively.

3.3 Question Posed

Question Answering System is posed with several question types. The questions are classified into unanswerable, irrelevant, or answerable

- Unanswerable: When the context is available in the passage, but the user poses the question in a very complex way, which is unanswerable by the question answering system, this question is labeled an unanswerable question.
- Irrelevant: When the user poses a question that is out of context with the given passage, this question is labeled as irrelevant.
- Answerable: It is defined as the question whose context is available in the given passage, and this question is answerable by the question answering system.

3.4 Question similarity score

The question similarity mechanism is used as a question filter to the Question Answering System. This mechanism identifies and filters unanswerable, irrelevant, and answerable questions based on the threshold value. The range of the QSS threshold and the corresponding label of the posed question is given in Table 2.

In our experiment, 1000 questions are chosen for unanswerable questions, irrelevant questions, and answerable questions from the SQuAD 2.0 dataset. We have found that the Irrelevant questions have question similarity scores in the range of 0.00 to 0.50 and unanswerable questions have their question similarity scores in the range 0.50 to 0.80. Further, we experimented to check the question similarity scores for the answerable questions and found that the question similarity scores are in the range of 0.85 to 1.00. So, we set the threshold values to be in the range of $0.00 - 0.50$ if the posed question is *Irrelevant*, $0.50 - 0.85$ if the posed question is *Unanswerable*, and $0.85 - 1.00$ if the posed question is *Answerable* question. If the question posed crosses the threshold value, it is identified as an answerable or relevant question, and it is passed to the question answering system to get the answer to that question. If the question posed does not cross the threshold, then as per the Table 2 it is identified either as irrelevant or unanswerable.

4 Data and Experiments

The following data are used for the experiments:

1. We have used SQuAD 2.0 [32] dataset for our experiments. It consists of 50,000 additional questions to that of SQuAD 1.1 [33] dataset, which has 100,000 answerable questions.

Table 2 Labeling of posed question

Question similarity score	Label of posed question
$0.00 - 0.50$	Irrelevant question
$0.50 - 0.85$	Unanswerable question
$0.85 - 1.00$	Answerable question

When the Question Similarity Score (QSS) is within a particular range then the corresponding label is assigned to that posed question

Table 3 This table illustrates possible generated questions from the passage, which is randomly taken from SQUAD 2.0 dataset

SQUAD 2.0	
Passage 1	In the late 17th century, Robert Boyle proved that air is necessary for combustion. English chemist John Mayow (1641-1679) refined this work by showing that fire requires only a part of air that he called spiritus nitroaereus or just nitroaereus. In one experiment, it is found that placing either a mouse or a lit candle in a closed container over water caused the water to rise and replace one-fourteenth of the air's volume before extinguishing the subjects. From this, he surmised that nitroaereus is consumed in both respiration and combustion.
Possible Generated Questions	G_1. when did robert boyle prove that air is necessary for combustion?
	G_2. who proved that air is necessary for combustion?
	G_3. how did robert boyle prove that air is necessary for combustion?
	G_4. what did robert boyle prove that air is necessary for?
	G_5. who refined boyle' s work?
	G_6. how did john mayow improve on boyle's work?
	G_7. what did john mayow improve on boyle's work?
	G_8. how did john mayow refined boyle's work?
	G_9. john mayow refined boyle' s work to show that only a part of air is required for what?
	G_{10}. what part of air does fire require?
	G_{11}. what did mayow show that fire requires?
	G_{12}. what did john mayow call the part of air that fire requires?
	G_{13}. what type of experiment did mayow use to prove that air is necessary for combustion?
	G_{14}. what did mayow do with a mouse or a candle in a closed container?
	G_{15}. along with a mouse, what did mayow place in a closed container over water?
	G_{16}. what kind of container did mayow place a mouse or a candle in?
	G_{17}. what did mayow place a mouse or candle in to extinguish them?
	G_{18}. what was the effect of placing a mouse or a candle in a closed container over water?
	G_{19}. what did mayow find when placing a mouse or a candle in a closed container over water?
	G_{20}. what did mayow think the water did with one-fourteenth of the air's volume before extinguishing the mice?
	G_{21}. how much of the air's volume did water replace before extinguishing the mice or candle?
	G_{22}. john mayow found that water replaced one-fourteenth of what in the air?
	G_{23}. what did mayow do to the mice or the candle?
	G_{24}. who did mayow extinguish in his experiment?
	G_{25}. how did mayow know that nitroaereus is consumed in both respiration and combustion?
	G_{26}. what did mayow call the part of air that is consumed in both respiration and combustion?
	G_{27}. what is nitroaereus thought to be in both respiration and combustion?
	G_{28}. john mayow theorized that nitroaereus is consumed in what type of combustion?

The question generation system, ProphetNet [1] is used to generate all possible questions G_1 to G_{28} from the given passage

2. The pre-trained weights of state-of-the-art Question Generation System called ProphetNet [29] to generate the questions for a given paragraph.
3. The pre-trained weights of BERT [11] Question Answering System, which is fine-tuned on the SQuAD 1.1 dataset [33].
4. Pre-trained Universal Sentence Encoder (USE) [6] to generate the sentence embeddings for the questions (Tables 3, 4, 5, 6, 7, 8, 9 and 10).

5 Results

5.1 Automatic question-answer pairs generation system

This subsection shows the results produced by the automatic question-answer pairs generation system. We have generated automatic question-answer pairs for 100 passages from the SQuAD 2.0 dataset [7]. Tables 3, 6

Table 4 This table illustrates the generated question-answer pairs from the given passage

Passage 1: In the late 17th century, Robert Boyle proved that air is necessary for combustion. English chemist John Mayow (1641-1679) refined this work by showing that fire requires only a part of air that he called spiritus nitroaereus or just nitroaereus. In one experiment he found that placing either a mouse or a lit candle in a closed container over water caused the water to rise and replace one-fourteenth of the air's volume before extinguishing the subjects. From this, he surmised that nitroaereus is consumed in both respiration and combustion.

G_1. when did robert boyle prove that air is necessary for combustion?	A_1. late 17th-century
G_2. who proved that air is necessary for combustion?	A_2. robert boyle
G_3. how did robert boyle prove that air is necessary for combustion?	A_3. placing either a mouse or a lit candle in a closed container over water caused the water to rise and replace one-fourteenth of the air's volume before extinguishing the subjects
G_4. what did robert boyle prove that air is necessary for?	A_4. combustion
G_5. who refined boyle' s work?	A_5. john mayow
G_6. how did john mayow improve on boyle's work?	A_6. by showing that fire requires only a part of air
G_7. what did john mayow improve on boyle's work?	A_7. showing that fire requires only a part of air
G_8. how did john mayow refined boyle's work?	A_8. by showing that fire requires only a part of air
G_9. john mayow refined boyle' s work to show that only a part of air is required for what?	A_9. fire
G_{10}. what part of air does fire require?	A_{10}. spiritus nitroaereus
G_{11}. what did mayow show that fire requires?	A_{11}. a part of air
G_{12}. what did john mayow call the part of air that fire requires?	A_{12}. spiritus nitroaereus
G_{13}. what type of experiment did mayow use to prove that air is necessary for combustion?	A_{13}. placing either a mouse or a lit candle in a closed container over water
G_{14}. what did mayow do with a mouse or a candle in a closed container?	A_{14}. caused the water to rise and replace one-fourteenth of the air's volume before extinguishing the subjects
G_{15}. along with a mouse, what did mayow place in a closed container over water?	A_{15}. a lit candle
G_{16}. what kind of container did mayow place a mouse or a candle in ?	A_{16}. closed
G_{17}. what did mayow place a mouse or candle in to extinguish them?	A_{17}. a closed container over water
G_{18}. what was the effect of placing a mouse or a candle in a closed container over water?	A_{18}. caused the water to rise and replace one-fourteenth of the air's volume before extinguishing the subjects
G_{19}. what did mayow find when placing a mouse or a candle in a closed container over water?	A_{19}. caused the water to rise and replace one-fourteenth of the air's volume before extinguishing the subjects
G_{20}. what did mayow think the water did with one-fourteenth of the air's volume before extinguishing the mice?	A_{20}. rise and replace
G_{21}. how much of the air's volume did water replace before extinguishing the mice or candle?	A_{21}. one - fourteenth
G_{22}. john mayow found that water replaced one-fourteenth of what in the air?	A_{22}. volume
G_{23}. what did mayow do to the mice or the candle?	A_{23}. extinguishing the subjects
G_{24}. who did mayow extinguish in his experiment?	A_{24}. subjects
G_{25}. how did mayow know that nitroaereus is consumed in both respiration and combustion?	A_{25}. placing either a mouse or a lit candle in a closed container over water caused the water to rise and replace one-fourteenth of the air's volume before extinguishing the subjects
G_{26}. what did mayow call the part of air that is consumed in both respiration and combustion?	A_{26}. spiritus nitroaereus or just nitroaereus.
G_{27}. what is nitroaereus thought to be in both respiration and combustion?	A_{27}. consumed
G_{28}. john mayow theorized that nitroaereus is consumed in what type of combustion?	A_{28}. respiration

The questions generated by the question generation system is further given to the question answering system using BERT[24], which finds the answers A_1 to A_{28} for all the generated questions.

Table 5 This table illustrates how the question similarity method identifies unanswerable or irrelevant questions and addresses the limitations of QAS

Passage 1: In the late 17th century, Robert Boyle proved that air is necessary for combustion. English chemist John Mayow (1641-1679) refined this work by showing that fire requires only a part of air that he called spiritus nitroaereus or just nitroaereus. In one experiment he found that placing either a mouse or a lit candle in a closed container over water caused the water to rise and replace one-fourteenth of the air's volume before extinguishing the subjects. From this, he surmised that nitroaereus is consumed in both respiration and combustion.

Question Posed (QP)	Question Similarity Score$_{(Highest)}$	Question Label	Answer from BERT
QP_1: What English chemist showed that fire only needed nitroaereus ?	(G_2, QP_1) = 0.68	Unanswerable	john mayow
QP_2: How many square miles large was the region impacted by the 2010 drought?	(G_{22}, QP_2) = 0.44	Irrelevant	how many square miles
QP_3: who proved that air is necessary for combustion ?	(G_2, QP_3) = 1.0	Answerable	robert boyle

The first column specifies the question posed. The second column specifies the highest question similarity score between the question posed and generated questions. The third column indicates whether the question is answerable/unanswerable/irrelevant based on the highest question similarity score by comparing it with the threshold (0.85). The fourth column indicates the answers generated for the questions posed, bypassing the question to the BERT question answering system

and 9 show all possible questions generated from the passages by question generation system. These questions are further given to the question answering system and the passage to generate the answers to the possible generated questions. Table 4, Table 7, and Table 10 show the question-answer pairs generated by automatic question-answer pairs generation system. On manual reading, it is found that the question-answer pairs generated are of good quality (Table 11).

5.2 Question similarity mechanism

This subsection provides the results of the proposed question similarity mechanism. When a question is posed

Table 6 This table illustrates possible generated questions from the passage, which is randomly taken from SQUAD 2.0 dataset

SQUAD 2.0	
Passage, 2	The availability of the Bible in vernacular languages was important to the spread of the Protestant movement and development of the Reformed church in France. The country had a long history of struggles with the papacy by the time the Protestant Reformation finally arrived. Around 1294, a French version of the Scriptures was prepared by the Roman Catholic priest, Guyard de Moulin. A two-volume illustrated folio paraphrase version based on his manuscript, by Jean de Rely, was printed in Paris in 1487.
Possible Generated Questions	G_1. what was important to the protestant movement in france? G_2. why did france struggle with the papacy before the protestant reformation ? G_3. what did france have a long history of with the papacy ? G_4. france had a long history of struggles with what group by the time what finally arrived ? G_5. when did the protestant reformation arrive in france ? G_6. what did guyard de moulin prepare around 1294 ? G_7. what did guyard de moulin do with the french version of the Bible? G_8. what religion was guyard de moulin ? G_9. who prepared a french version of the scriptures around 1294 ? G_{10}. what was printed in Paris in 1487 ? G_{11}. on what manuscript was jean de rely's version of the Bible-based ? G_{12}. what was jean de rely's version based on ? G_{13}. who created the illustrated folio paraphrase version of the Bible in 1487 ? G_{14}. how was jean de rely's version of the Bible published ? G_{15}. where was jean de rely's version of the Bible printed ?

The question generation system called ProphetNet[1] is used to generate all possible questions G_1 to G_{15} from the given passage

Table 7 This table illustrates the generated question-answer pairs from the given passage

Passage 2: The availability of the Bible in vernacular languages was important to the spread of the Protestant movement and the development of the Reformed church in France. The country had a long history of struggles with the papacy by the time the Protestant Reformation finally arrived. Around 1294, a French version of the Scriptures was prepared by the Roman Catholic priest, Guyard de Moulin. A two-volume illustrated folio paraphrase version based on his manuscript, by Jean de Rely, was printed in Paris in 1487.

G_1. what was important to the protestant movement in france ?	A_1. the availability of the Bible in vernacular languages
G_2. why did france struggle with the papacy before the protestant reformation ?	A_2. the country had a long history
G_3. what did france have a long history of with the papacy ?	A_3. struggles
G_4. france had a long history of struggles with what group by the time what finally arrived ?	A_4. the papacy
G_5. when did the protestant reformation arrive in france ?	A_5. around 1294
G_6. what did guyard de moulin prepare around 1294 ?	A_6. a french version of the scriptures
G_7. what did guyard de moulin do with the french version of the Bible?	A_7. prepared
G_8. what religion was guyard de moulin ?	A_8. roman catholic
G_9. who prepared a french version of the scriptures around 1294 ?	A_9. guyard de moulin
G_{10}. what was printed in Paris in 1487 ?	A_{10}. a two-volume illustrated folio paraphrase version based on his manuscript
G_{11}. on what manuscript was jean de rely 's version of the Bible-based ?	A_{11}. guyard de moulin
G_{12}. what was jean de rely 's version based on ?	A_{12}. manuscript
G_{13}. who created the illustrated folio paraphrase version of the bible in 1487 ?	A_{13}. jean de rely
G_{14}. how was jean de rely 's version of the bible published ?	A_{14}. two-volume illustrated folio paraphrase
G_{15}. where was jean de rely 's version of the Bible printed ?	A_{15}. paris

The questions generated by the question generation system is further given to the question answering system using BERT[24], which finds the answers A_1 to A_{15} for all the generated questions

Table 8 This table illustrates how the question similarity method identifies unanswerable or irrelevant questions and addresses the limitations of QAS

Passage 2: The availability of the Bible in vernacular languages was important to the spread of the Protestant movement and development of the Reformed church in France. The country had a long history of struggles with the papacy by the time the Protestant Reformation finally arrived. Around 1294, a French version of the Scriptures was prepared by the Roman Catholic priest, Guyard de Moulin. A two-volume illustrated folio paraphrase version based on his manuscript, by Jean de Rely, was printed in Paris in 1487.

Question Posed (QP)	Question Similarity Score$_{(Highest)}$	Question Label	Answer from BERT
QP_1: Where was Jean de Rely from?	$(G_{12}, QP_1) = 0.67$	Unanswerable	paris
QP_2: What is the name of the cyclone?	$(G_{12}, QP_2) = 0.39$	Irrelevant	what
QP_3: What helped spread Protestantism in France?	$(G_1, QP_3) = 0.93$	Answerable	the availability of the bible in vernacular languages

The first column specifies the question posed. The second column specifies the highest question similarity score between the question posed and generated questions. The third column indicates whether the question is answerable/unanswerable/irrelevant based on the highest question similarity score by comparing it with the threshold (0.85). The fourth column indicates the answers generated for the questions posed, bypassing the question to the BERT question answering system

to the question answering system, the question similarity mechanism identifies whether the question posed is answerable or unanswerable and relevant or irrelevant questions. Both unanswerable and irrelevant questions are taken from the SQuAD 2.0 dataset [7] for the experiments.

We have carried out the experiments for random 100 passages from the SQuAD 2.0 dataset [7] with unanswerable and irrelevant questions. As shown in Tables 5, 8 and 12, when the cosine similarity score of generated question and posed question does not exceed the threshold of 0.85, it is marked or labeled either as an unanswerable or irrelevant question. Such a question will not be passed to the Question Answering System. So, the question posed with less than the threshold will not be passed to the Question Answering system. Our proposed question similarity mechanism does not allow the question-answering model to answer the unanswerable or irrelevant questions by incorrect guessing. We also present the Question Similarity Scores for answerable questions from SQuAD 2.0 dataset [7]. We found that the answerable questions get Question Similarity scores above 0.90. We can infer that the question similarity mechanism identifies the questions on par with human judgment.

We have experimented with 1000 questions for both Unanswerable and Irrelevant questions. In our experiments, we have used the BERT model trained on SQuAD 1.1 dataset. BERT model trained on SQuAD 2.0 should not predict answers for the Unanswerable questions. However, this model answers few Unanswerable questions. We introduced the Question Similarity mechanism with the BERT model trained on SQuAD 1.1; this mechanism helps to identify unanswerable and irrelevant questions. Irrelevant questions are not introduced in the SQuAD 2.0 dataset. For a particular passage in SQuAD 2.0 dataset, irrelevant questions are chosen randomly from the different passages. So that the randomly chosen questions will not be related to the context. The efficiency of the model is calculated by,

$$\text{Efficiency} = \frac{\text{No. of Unanswerable/Irrelevant questions not answered by the model}}{\text{Total No. of Unanswerable/Irrelevant questions}} \times 100 \tag{6}$$

Table 9 This table illustrates possible generated questions from the passage, which is randomly taken from SQUAD 2.0 dataset

SQUAD 2.0	
Passage 3	Formed in November 1990 by the equal merger of Sky Television and British Satellite Broadcasting, BSkyB became the UK's largest digital subscription television company. Following BSkyB's 2014 acquisition of Sky Italia and a majority 90.04% interest in Sky Deutschland in November 2014, its holding company British Sky Broadcasting Group plc changed its name to Sky plc. The United Kingdom operations also changed the company name from British Sky Broadcasting Limited to Sky UK Limited, still trading as Sky.
Possible Generated Questions	G_1. how was bskyb formed? G_2. when was bskyb formed?
	G_3. what caused the formation of bskyb?
	G_4. what was the name of the UK's largest digital subscription television company?
	G_5. what did bskyb do when it was formed?
	G_6. in what country did bskyb become the largest digital subscription television company?
	G_7. what type of company is bskyb?
	G_8. how many acquisitions did bskyb make in 2014?
	G_9. what was the name of the UK's largest digital subscription television company?
	G_{10}. what was the result of bskyb' s 2014 acquisition of sky deutschland?
	G_{11}. what percentage of sky deutschland did bskyb acquire in 2014?
	G_{12}. what type of company is british sky broadcasting group plc?
	G_{13}. what was the name of bskyb' s holding company?
	G_{14}. what was the name of british sky broadcasting group's holding company in november 2014?
	G_{15}. what did british sky broadcasting group plc change its name to in november 2014?
	G_{16}. who changed the company name from british Sky broadcasting limited to sky UK limited?
	G_{17}. what happened to british sky broadcasting group plc's name?
	G_{18}. what did the united kingdom operations of bskyb change their name to?
	G_{19}. what is the current trading name of sky UK limited?
	G_{20}. what is the trading name of sky UK limited?

The question generation system called ProphetNet[1] is used to generate all possible questions G_1 to G_{20} from the given passage

Table 10 This table illustrates the generated question-answer pairs from the given passage

Passage 3: Formed in November 1990 by the equal merger of Sky Television and British Satellite Broadcasting, BSkyB became the UK's largest digital subscription television company. Following BSkyB's 2014 acquisition of Sky Italia and a majority 90.04% interest in Sky Deutschland in November 2014, its holding company British Sky Broadcasting Group plc changed its name to Sky plc. The United Kingdom operations also changed the company name from British Sky Broadcasting Limited to Sky UK Limited, still trading as Sky.	
G_1. how was bskyb formed?	A_1. by the equal merger of sky television and british satellite broadcasting
G_2. when was bskyb formed?	A_2. november 1990
G_3. what caused the formation of bskyb?	A_3. the equal merger of sky television and british satellite broadcasting
G_4. what was the name of the UK's largest digital subscription television company?	A_4. bskyb
G_5. what did bskyb do when it was formed?	A_5. merger of sky television and british satellite broadcasting
G_6. in what country did bskyb become the largest digital subscription television company?	A_6. uk
G_7. what type of company is bskyb?	A_7. digital subscription television company
G_8. how many acquisitions did bskyb make in 2014?	A_8. 2014 acquisition of sky italia and a majority 90.04% interest in sky deutschland
G_9. what was the name of the UK's largest digital subscription television company?	A_9. bskyb
G_{10}. what was the result of bskyb' s 2014 acquisition of sky deutschland?	A_{10}. british sky broadcasting group plc changed its name to sky plc
G_{11}. what percentage of sky deutschland did bskyb acquire in 2014?	A_{11}. 90.04 %
G_{12}. what type of company is british Sky broadcasting group plc ?	A_{12}. holding company
G_{13}. what was the name of bskyb' s holding company?	A_{13}. british sky broadcasting group plc changed its name to sky plc
G_{14}. what was the name of british sky broadcasting group's holding company in november 2014?	A_{14}. sky plc
G_{15}. what did british sky broadcasting group plc change its name to in november 2014?	A_{15}. sky plc
G_{16}. who changed the company name from british Sky broadcasting limited to sky uk limited?	A_{16}. united kingdom operations
G_{17}. what happened to british sky broadcasting group UK's name ?	A_{17}. sky plc
G_{18}. what did the united kingdom operations of bskyb change their name to ?	A_{18}. sky UK limited
G_{19}. what is the current trading name of sky UK limited?	A_{19}. sky
G_{20}. what is the trading name of sky UK limited?	A_{20}. sky

The questions generated by the question generation system is further given to the question answering system using BERT[24], which finds the answers A_1 to A_{20} for all the generated questions

Table 11 This table illustrates how the question similarity method identifies unanswerable or irrelevant questions and addresses the limitations of QAS

Passage 3: Formed in November 1990 by the equal merger of Sky Television and British Satellite Broadcasting, BSkyB became the UK's largest digital subscription television company. Following BSkyB's 2014 acquisition of Sky Italia and a majority 90.04% interest in Sky Deutschland in November 2014, its holding company British Sky Broadcasting Group plc changed its name to Sky plc. The United Kingdom operations also changed the company name from British Sky Broadcasting Limited to Sky UK Limited, still trading as Sky.			
Question Posed (QP)	Question Similarity Score$_{(Highest)}$	Question Label	Answer from BERT
QP_1: What company no longer trades as Sky?	(G_{19}, QP_1) = 0.77	Unanswerable	sky uk limited
QP_2: In which part of italy plague was less common?	(G_3, QP_2) = 0.32	Irrelevant	deutschland
QP_3: What was the name of the uk's largest digital subscription television company?	(G_4, QP_3) = 0.93	Answerable	bskyb

The first column specifies the question posed. The second column specifies the highest question similarity score between the question posed and generated questions. The third column indicates whether the question is answerable/unanswerable/irrelevant based on the highest question similarity score by comparing it with the threshold (0.85). The fourth column indicates the answers generated for the questions posed, bypassing the question to the BERT question answering system

Table 12 Quantitative analysis of the BERT model trained on SQuAD 2.0 and BERT model trained on SQuAD 1.1 with Question Similarity Mechanism

Models	Unanswerable Questions	Efficiency of Unanswerable Questions	Irrelevant Questions	Efficiency of Irrelevant Questions
BERT trained on SQuAD 2.0	Model does not answers for **480** Unanswerable questions out of **1000** Unanswerable questions	**48%**	Model answers all Irrelevant questions, because this model cannot identify irrelevant questions	-
BERT trained on SQuAD 1.1 with Question Similarity Mechanism	Model does not answers for **910** Unanswerable questions out of **1000** Unanswerable questions	**91%**	Model does not answers for **1000** Irrelevant questions	**100%**

6 Discussion

The automatic question-answer pairs generation gives an overview of how the question answering system and question generation work as a twin task system to obtain satisfactory results. Also, on manual reading, we can infer that this system generates good question and answer pairs. The question-answer pairs generated to date are confined to generate only 'wh' questions and their answers. The majority of the question-answer pairs generation systems are rule-based systems. Whereas our proposed application generates all possible question-answer pairs using a machine learning approach.

In the question similarity mechanism, we show the work's significance by addressing the Question Answering System's challenge. Even though the works like [1, 3, 32, 33] introduced different techniques to overcome the limitations of the Question Answering System, the identification of the unanswerable questions remains an open challenge. The proposed Question Similarity mechanism does not require training. It improves the question answering systems' performance by focusing only on the answerable or relevant questions. By this, we can infer that the Question Similarity mechanism incorporates a human way of reasoning to identify unanswerable and irrelevant questions and hence addresses the limitation of QAS.

7 Conclusion

In this paper, we introduce an application by combining the Question Generation and Question Answering system called automatic question-pairs generation system, where all possible question and answer pairs will be generated. It has got various applications in different fields. Later, we introduce a Question Similarity mechanism that imitates human reasoning to identify whether the question posed is answerable questions or unanswerable and irrelevant questions. The existing question answering systems cannot identify whether the question posed is answerable or unanswerable and irrelevant. If the question posed is unanswerable or irrelevant, then such questions are not passed to the QAS. As there is no training process involved in this model, it requires less computational resources. This mechanism can be included with state-of-the-art Question Answering Systems so that the models can concentrate on answerable questions to improve their performance. The automatically generated question-answer pairs can be used as a dataset to train the Question Answering models.

Acknowledgements The authors would like to thank Dr.Harishchandra Hebbar from the School of Information Science (SOIS), Manipal, for providing access to the GPU-based computing facility.

We thank the anonymous reviewers whose insightful comments and suggestions have significantly improved this paper.

Author Contributions All authors have contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Shivani G Aithal and Abishek B Rao. Shivani G Aithal wrote the first draft of the manuscript, Abishek B Rao, and all authors commented on previous versions of the manuscript. Sanjay Singh did the supervision, reviewing, and editing. All authors read and approved the final manuscript.

Funding Open access funding provided by Manipal Academy of Higher Education, Manipal. For this study, we have not sought any funding from any agency.

Availability of data and material For this study, we have used the dataset available in the public domain. The source of the dataset is cited in the paper.

Code Availability The code is available on request.

Declarations

Competing interests The authors declare that they have no conflict of interest.

References

1. Baudiš P, Šedivý J (2015) Modeling of the Question Answering Task in the yodaQA System. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones G, San Juan E, Capellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction. Springer International Publishing, Cham, pp 222–228
2. Benamara F (2004) Cooperative question answering in restricted domains: the WEBCOOP experiment. In: Proceedings of the Conference on Question Answering in Restricted Domains, pp. 31–38. Association for Computational Linguistics, Barcelona. https://www.aclweb.org/anthology/W04-0506
3. Berant J, Chou A, Frostig R, Liang P (2013) Semantic Parsing on Freebase from Question-Answer Pairs. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1533–1544. Association for Computational Linguistics, Seattle. https://www.aclweb.org/anthology/D13-1160
4. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08. Association for Computing Machinery, New York, pp 1247–1250. https://doi.org/10.1145/1376616.1376746
5. Catelli R, Casola V, De Pietro G, Fujita H, Esposito M (2021) Combining contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical de-identification. Knowl-Based Syst 213:106649. https://doi.org/10.1016/j.knosys.2020.106649, https://www.sciencedirect.com/science/article/pii/S0950705120307784
6. Cer D, Yang Y, Kong S. y, Hua N, Limtiaco N, St. John R, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Strope B, Kurzweil R (2018) Universal Sentence Encoder for English. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 169–174. Association for Computational Linguistics, Brussels. https://doi.org/10.18653/v1/D18-2029, https://www.aclweb.org/anthology/D18-2029
7. Chen D, Fisch A, Weston J, Bordes A (2017) Reading wikipedia to answer Open-Domain questions. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vancouver, pp 1870–1879. https://doi.org/10.18653/v1/P17-1171
8. Chen Y, Li H (2020) DAM: Transformer-Based relation detection for Question Answering over Knowledge Base. Knowl-Based Syst 201-202:1–8. https://doi.org/10.1016/j.knosys.2020.106077
9. Cuteri B, Reale K, Ricca F (2019) A Logic-Based question answering system for cultural heritage. In: Calimeri F, Leone N, Manna M (eds) Logics in artificial intelligence. Springer International Publishing, Cham, pp 526–541
10. Dehghani M, Azarbonyad H, Kamps J, de Rijke M (2019) Learning to transform, combine, and reason in Open-Domain question answering. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19. Association for Computing Machinery, New York, pp 681–689. https://doi.org/10.1145/3289600.3291012
11. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training Of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
12. Di Gennaro G, Buonanno A, Di Girolamo A, Ospedale A, Palmieri FAN (2020) Intent Classification in Question-Answering Using LSTM Architectures. In: Esposito A, Faundez-Zanuy M, Morabito FC (eds) Progresses in Artificial Intelligence and Neural Systems. Springer Singapore, Singapore, pp 115–124. https://doi.org/10.1007/978-981-15-5093-5_11
13. Dutil F, Gulcehre C, Trischler A, Bengio Y (2017) Plan, Attend, Generate: Planning for Sequence-to-Sequence Models. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, pp 5480–5489
14. Esposito M, Damiano E, Minutolo A, De Pietro G, Fujita H (2020) Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Information Sciences 514:88–105. https://doi.org/10.1016/j.ins.2019.12.002
15. Ferrucci D, Nyberg E, Allan J, Barker K, Brown EW, Chu-Carroll J, Ciccolo AC, Duboué PA, Fan J, Gondek DC, Hovy E, Katz B, Lally A, McCord M, Morarescu P, Murdock B, Porter B, Prager JM, Strzalkowski T, Welty C, Zadrozny W (2009) IBM Research report towards the open advancement of question answering systems. Tech. Rep. RC24789 (w0904-093) IBM
16. Green BF, Wolf AK, Chomsky C, Laughery K (1961) Baseball: An Automatic Question-Answerer. In: Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference, IRE-AIEE-ACM '61 (Western), pp 219–224. Association for Computing Machinery, New York. https://doi.org/10.1145/1460690.1460714
17. Hermjakob E, Hovy U, Gerber L, Junk M, Lin CY (2000) Question answering in webclopedia. In: Proceedings of the TREC-9 conference, NIST, Gaithersburg, pp 1–10
18. Khurana D, Koli A, Khatter K, Singh S (2017) Natural language processing: State of the art. Current Trends and Challenges. arXiv:1708.05148
19. Krueger D, Maharaj T, Kramár J, Pezeshki M, Ballas N, Ke NR, Goyal A, Bengio Y, Courville AC, Pal C (2017) Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations. In: Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, pp 1–11
20. Kumar V, Muneeswaran S, Ramakrishnan G, Li YF (2019) ParaQG: A System for Generating Questions and Answers from Paragraphs. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pp. 175–180. Association for Computational Linguistics, Hong Kong. https://doi.org/10.18653/v1/D19-3030, https://www.aclweb.org/anthology/D19-3030
21. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In: Proceedings of Eighth International Conference on Learning Representation (ICLR), Addis Ababa, pp 1–17. https://iclr.cc/virtual_2020/poster_H1eA7AEtvS.html
22. Merity S, Keskar NS, Socher R (2018) Regularizing and optimizing LSTM language models. In: Proceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, pp 1–10
23. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed Representations of Words and Phrases and

their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems, vol 26, Curran Associates, Inc, pp 3111–3119
24. Mishra A, Jain SK (2016) A survey on question answering systems with classification. J King Saud Univ-Comput Inf Sci 28(3):345–361
25. Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. In: Dasgupta S, McAllester D (eds) Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 28. PMLR, Atlanta, pp 1310–1318. http://proceedings.mlr.press/v28/pascanu13.html
26. Pennington J, Socher R, Manning C (2014) GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha. https://doi.org/10.3115/v1/D14-1162. https://www.aclweb.org/anthology/D14-1162
27. Pota M, Esposito M, De Pietro G, Fujita H (2020) Best Practices of Convolutional Neural Networks for Question Classification. Appl Sci 10(14). https://doi.org/10.3390/app10144710, https://www.mdpi.com/2076-3417/10/14/4710
28. Pota M, Marulli F, Esposito M, De pietro G, Fujita H (2019) Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched Word Embeddings. Knowledge-Based Sys. 164:309–323. https://doi.org/10.1016/j.knosys.2018.11.003
29. Qi W, Yan Y, Gong Y, Liu D, Duan N, Chen J, Zhang R, Zhou M (2020) Prophetnet: Predicting Future N-gram for Sequence-to-Sequence Pre-training In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 2401–2410. https://doi.org/10.18653/v1/2020.findings-emnlp.217
30. Qiao C, Hu X (2020) A neural knowledge graph evaluator: Combining structural and semantic evidence of knowledge graphs for predicting supportive knowledge in scientific QA. Inf Process Manag 57(6):102309. https://doi.org/10.1016/j.ipm.2020.102309
31. Rajpurkar P (2020) Performance of Unanswerble questions in SQUAD 2.0. https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/ (2020) [Online; accessed 10
32. Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: Unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Melbourne, pp 784–789. https://doi.org/10.18653/v1/P18-2124, https://www.aclweb.org/anthology/P18-2124
33. Rajpurkar P, Zhang J, Lopyrev K, 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. Association for Computational Linguistics, Austin, pp 2383–2392. https://doi.org/10.18653/v1/D16-1264
34. Ray A, Christie G, Bansal M, Batra D, Parikh D (2016) Question relevance in VQA: identifying Non-Visual and False-Premise questions. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, pp 919–924. https://doi.org/10.18653/v1/D16-1090
35. Reddy S, Raghu D, Khapra MM, Joshi S (2017) Generating Natural Language Question-Answer Pairs from a Knowledge Graph Using a RNN Based Question Generation Model. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Association for Computational Linguistics, Valencia, pp 376–385. https://www.aclweb.org/anthology/E17-1036
36. Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, pp 1–5. Vancouver. https://www.emc2-ai.org/assets/docs/neurips-19/emc2-neurips19-paper-33.pdf
37. Serdyuk D, Ke NR, Sordoni A, Trischler A, Pal C, Bengio Y (2018) Twin networks: Matching the future for sequence generation. In: Proceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, pp 1–12
38. Song J, Liu F, Ding K, Du K, Zhang X (2020) Semantic comprehension of questions in q& a system for chinese language based on semantic element combination. IEEE Access 8:102971–102981. https://doi.org/10.1109/ACCESS.2020.2997958
39. Sun Y, Tang D, Duan N, Qin T, Liu S, Yan Z, Zhou M, Lv Y, Yin W, Feng X, Qin B, Liu T (2020) Joint learning of question answering and question generation. IEEE Trans Knowl Data Eng 32(5):971–982
40. Weizenbaum J (1966) ELIZA-A computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45
41. Winograd T (1972) Understanding natural language. Cogn Psychol 3(1):1–191. https://doi.org/10.1016/0010-0285(72)90002-3
42. Woods WA, Kaplan R (1977) Lunar rocks in natural English: Explorations in natural language question answering. In: Zampolli A (ed) linguistic structures processing, fundamental studies in computer science. North-holland publishing company, pp 266–290
43. Yang W, Xie Y, Lin A, Li X, Tan L, Xiong K, Li M, Lin J (2019) End-to-end Open-Domain Question Answering with BERTserini. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Association for Computational Linguistics, Minneapolis, pp 72–77. https://doi.org/10.18653/v1/N19-4013
44. Ye Y, Zhang S, Li Y, Qian X, Tang S, Pu S, Xiao J (2020) Video question answering via grounded cross-attention network learning. Inf Process Manag 57(4):102265. https://doi.org/10.1016/j.ipm.2020.102265
45. Zahedi M, Rahgozar M, Zoroofi R (2020) HCA: Hierarchical Compare Aggregate model for question retrieval in community question answering. Inf Process sManag 57(6):102318. https://doi.org/10.1016/j.ipm.2020.102318
46. Zhu H, Dong L, Wei F, Wang W, Qin B, Liu T (2019) Learning to Ask Unanswerable Questions for Machine Reading Comprehension. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, pp 4238–4248. https://doi.org/10.18653/v1/P19-1415, https://www.aclweb.org/anthology/P19-1415