applied sciences

MDPI

Article

Question Difficulty Estimation Based on Attention Model for Question Answering

Hyun-Je Song [1,†], Su-Hwan Yoon [2,†] and Seong-Bae Park [2,*]

[1] Department of Information Technology, Jeonbuk National University, Jeonju 54896, Korea; hyunje.song@jbnu.ac.kr
[2] Department of Computer Science and Engineering, Kyung Hee University, Youngin 17104, Korea; yunsh3432@khu.ac.kr
* Correspondence: sbpark71@khu.ac.kr
† These authors contributed equally to this work.

Abstract: This paper addresses a question difficulty estimation of which goal is to estimate the difficulty level of a given question in question-answering (QA) tasks. Since a question in the tasks is composed of a questionary sentence and a set of information components such as a description and candidate answers, it is important to model the relationship among the information components to estimate the difficulty level of the question. However, existing approaches to this task modeled a simple relationship such as a relationship between a questionary sentence and a description, but such simple relationships are insufficient to predict the difficulty level accurately. Therefore, this paper proposes an attention-based model to consider the complicated relationship among the information components. The proposed model first represents bi-directional relationships between a questionary sentence and each information component using a dual multi-head co-attention, since the questionary sentence is a key factor in the QA questions and it affects and is affected by information components. Then, the proposed model considers inter-information relationship over the bi-directional representations through a self-attention model. The inter-information relationship helps predict the difficulty of the questions accurately which require reasoning over multiple kinds of information components. The experimental results from three well-known and real-world QA data sets prove that the proposed model outperforms the previous state-of-the-art and pre-trained language model baselines. It is also shown that the proposed model is robust against the increase of the number of information components.

Keywords: attention model; dual multi-head attention; inter-information relationship; question answering; question difficult estimation

Citation: Song, H.-J.; Yoon, S.H.; Park, S.-B. Question Difficulty Estimation Based on Attention Model for Question Answering. *Appl. Sci.* **2021**, *11*, 12023. https://doi.org/10.3390/app112412023

Academic Editors: Arturo Montejo-Ráez and Salud María Jiménez-Zafra

Received: 7 November 2021
Accepted: 14 December 2021
Published: 17 December 2021

1. Introduction

Question-Answering (QA) is an important natural language processing task in which a model understands questions and answers them based on its understanding of the questions. Several QA tasks such as ARC [1], SQuAD [2], and HotpotQA [3] were recently proposed, and many QA models based on a pre-trained language model have been developed to solve these QA tasks [4–7]. In these QA tasks, the questions are in general prepared without consideration of difficulty. Therefore, the QA models attacking the tasks do not recognize the difficulty of each question even though the difficulty is important information to answer the questions [8]. As a result, a difficulty level is tagged in new QA tasks such as DramaQA [9] in conjunction with Piaget's theory [10].

All QA tasks do not contain information about question difficulty, but the difficulty exists latently in their questions. The questions in a QA task can be regarded as easy if they are correctly answered by many answering models, and they can be considered as difficult if few models give a correct answer for them. When investigating (This investigation was done on 10 November 2020) the questions in the QuAC task with top three single models

from the leaderboard of the task, we found out that only 10% of the questions are answered correctly by all three models while about 50% are not answered correctly by any of the models. Besides, some QA tasks have intrinsic question difficulty. For instance, the RACE data set was collected by two subgroups of middle school examinations and high school examinations, respectively. Thus, the questions from middle school examinations are easier than those from high school examinations.

This paper deals with a question difficulty estimation of which goal is to estimate the difficulty level of a given question. Predicting the difficulty level of the question helps create adversarial QA datasets [11] or identify the way in which QA models challenges. Most previous studies on this task extracted some difficulty features from questions and then predicted the difficulty level of the questions with the features using machine learning algorithms [12–16]. These features were designed to model the relationship between a questionary sentence and associated information components such as a passage or candidate answers. However, some recent QA studies have shown that inter-information relationship is vital since many difficult questions can be answered through reasoning over multiple kinds of information components [17]. For such an example, Figure 1 shows a question in the RACE task. To answer this question, an answering model has to identify the relationship between the passage and a candidate answer (marked in cyan) as well as the relationship between the questionary sentence and the passage (marked in green). As in the question answering, these relationships are important factors also in estimating the question difficulty. Especially, the inter-information relationship should be considered explicitly because they are directly related to the question difficulty, but no previous studies made many efforts to consider the relationship.

questionary sentence		Why did Mami experience culture shock in Japan?
Information components	passage	A Japanese student called Mami told me about her own experiences in British. She spent 10 months in the UK last year, studying English at a language school. She really enjoyed her first two weeks in the UK. But soon she started to miss things of her own country. (…) To comfort herself Mami began to spend many hours on the Internet chatting with her friends back home. She spent a couple of weeks in the countryside in Kent. She went to a social club for British people who were interested in Japan and started to make some friends there. In addition, she took a short course in calligraphy to get an opportunity of mixing with local people. A few months later, Mami's impression of the UK had greatly changed. She found that most of the British were friendly, witty and fun. However, once Mami was back in Japan, she experienced "culture shock" again. She said, "I missed the friends I had made in England. My way of thinking had changed. Sometimes I was annoyed by the views of people in my country—for example, about the value of money and time. I thought people around me lived in such a small world." Mami noticed some changes in her behaviour: "I kept the habit of always carrying an umbrella with me, even on a fine day-—my friends thought I was crazy!"
	candidate answers	1. She didn't like Japanese culture any more. 2. The Japanese behaviour had changed a lot. 3. The world in Japan was too small for her. 4. She had got used to British culture and life.

Figure 1. An example question in the RACE data set that is difficult to answer without inter-information inference. The inter-information clues are marked in green and cyan, and the underline in the candidate answers implies a correct answer. (best view in color).

This paper proposes an attention-based model that estimates the difficulty of a question. The proposed attention model is designed to consider the inter-information relationship as well as the relationships between a questionary sentence and each information component. To be specific, the proposed model represents each type of the relationships consecutively and adopts the attention mechanism to capture both types of relationships. That is, the relationships between a questionary sentence and each information component

are first identified by the dual multi-head attention designed to capture a bi-directional relationship with two multi-head attentions. Since a single directional relationship is not sufficient for QA tasks [18], the proposed model captures bi-directional relationships between a questionary sentence and information components through the dual multi-head attention. Note that the bi-directional relationships do not reflect an inter-relationship among various information components fully. Thus, the proposed model represents the inter-information relationship by applying a multi-head attention again to the outputs of the dual multi-head attentions. That is, it first expresses the bi-directional relationships between a questionary sentence and each information component, and then accumulates the inter-information relationship onto the concatenation of the bi-directional relationship representations using the transformer encoder. Finally, it determines the difficulty of the question from the accumulated representation since the representation contains all information about the question components and their relationships.

The proposed model is verified with three QA data sets of RACE, QuAC, and DramaQA. Note that not all datasets are attached with the difficulty levels. DramaQA is manually tagged with four difficulty levels but RACE and QuAC are not tagged. For RACE dataset, we regard the middle school examinations as easy questions and high school examinations as hard questions. For QuAC, the difficulty levels are tagged using the results of multiple QA models [19]. The experimental results show the effectiveness of the proposed model in two folds. One is that the proposed model outperforms current state-of-the-art and pre-trained language models, and the other is that the performance of question difficulty estimation is improved by considering inter-information relationship. In particular, the proposed model achieves 68.37 of F1-score in QuAC. This is 8.5 higher than the F1-scores of the state-of-the-art pre-trained language models. It is also shown that the performance of the proposed model improves monotonically as the number of information components increases. The major performance improvement of the proposed model is made from difficult questions, since the proposed model is robust against the increase of the number of information components.

The major contributions of this paper can be summarized as follows:

- We formally define the question difficult estimation as estimating the difficulty level of a given question in question-answering tasks. The question difficult estimation for any question answering tasks can be formulated using the proposed definition.
- We design an attention-based model for question difficulty estimation. The proposed attention-based model captures the relationship among the information components as well as the inter-relationships between a questionary sentence and each information component.
- We examine the performance of the proposed model with intensive experiments on three real-world QA data sets. The intensive experiments validate the effectiveness of the proposed model.
- We empirically show that the performance of question answering is improved by adding the difficulty level.

The rest of this paper is organized as follows. Section 2 reviews related studies on question difficult estimation, and Section 3 presents the proposed model, the attention-based question difficulty estimator. The experimental results and discussions are given in Section 4. Finally, Section 5 draws some conclusions.

2. Related Work

Question answering is a task of answering a question where the question consists of a questionary sentence written in the natural language and a set of information components. Depending on the domain of the main information component, QA tasks are categorized into text-based [2,3], table-based [20,21], image-based [22,23], video-based [8,24], and so on. All QA tasks require an understanding of a question to answer it regardless of QA types. One key factor for the question understanding is the difficulty of the question [8], so that there have been many efforts to measure the difficulty of questions [14,16].

The efforts for the question difficulty estimation can be clustered into two types. The first type defines hand-crafted features from given QA materials. For instance, a question and its associated passage are usually given in the reading comprehension, where the passage provides background information of the question. Thus, the question difficulty is estimated with the information residing in the question and the passage. Desai and Moldovan defined, as such information, six features that are question length, cosine similarity between a question and a passage, the nature of a question and its answer, the number of clauses and prepositional phrases in a question, and existence of discourse connectives in a question [12]. On the other hand, Ha et al. defined the features for multiple-choice examinations [13]. Since they focused on medical examinations, they do not include only lexical, syntactic, and semantic features from a question and candidate answers, but also some cognitively-motivated features from a medical database. The main problem of these studies is that it is extremely difficult to design the features without profound knowledge about the reading materials.

The other type is to adopt a machine learning method to predict question difficulty without manual features. Since every QA task has its own idiosyncratic circumstances, the previous studies attacked question difficulty estimation by focusing on a specific task. Huang et al. estimated question difficulty for standard English tests in which each problem consists of a question, a reading passage, and candidate answers [14]. They proposed a CNN-like architecture to represent all sentences in the question, the passage, and the candidate answers as vectors, and adopted an attention mechanism to reflect the relevancy of the sentences in the passage and candidate answers to the question. Qiu et al. estimated question difficulty for multiple-choice problems at medical examinations [16]. Unlike English tests, the problems of medical examinations do not have a passage, but a set of documents related to a question. Thus, they measured two kinds of difficulties: the difficulty of searching the documents for potential answers of a question and the confusion difficulty among candidate answers. Then, the final difficulty of a question is determined by their weighted sum. Xue et al. expressed a question and candidate answers as embedding vectors by a pre-trained language model, ELMo, and then predicted the difficulty of the question using a simple linear regression of which input is the embedding vectors [25].

Note that many QA tasks provide some information components of a question as well as the question itself. Thus, the studies about representing inter-information have been performed [26], and they are grouped into two types according to the approach to expressing inter-information. One is to adopt a graph of which nodes are the entities appearing at information components and edges are a relation between the entities. Cao et al. expressed the relations among supporting documents in a multi-hop QA as a graph [27]. The nodes of this graph are the named entities in the documents and the edges are the co-reference or same-matching relation between entities. Then, they represented the graph as a vector reflecting the relations using the graph convolutional network. Song et al. also expressed the named entities as the nodes of a graph [28], but they added the window relation for the edges where two entities are regarded to have a window relation if they both appear within a word window. After that, they represented the graph as a vector for solving a multi-hop QA with the graph recurrent network.

The other approach to expressing inter-information is to obtain attention among information components. In the multi-passage reading comprehension, the candidate answers as well as the multiple passages can be regarded as information components. Thus, Wang et al. represented the candidate answers as vectors and expressed the relationship among all candidate answers as an attention matrix by applying an attention mechanism to the vectors [29]. On the other hand, Zhuang and Wang represented the relationships between a questionary sentence and its associated passages as vectors using Bi-DAF [17]. Then, they expressed the relationship among the passage vectors with the proposed dynamic self-attention. In the open-domain QA, Dehghani et al. used the universal transformer to represent the inter-information among the documents related to a question [30]. In the multi-evidence QA, Zhong et al. expressed the inter-information among a questionary

sentence, candidate answers, and associated documents [31]. In this work, they adopted the co-attention to express the relationship among the information components since the co-attention allows the representation of bidirectional relationships.

3. Attention-Based Question Difficulty Estimation

This paper defines a question difficulty estimation as determining the optimal difficulty level $y^* \in \mathcal{Y}$ of a given question, where $\mathcal{Y}$ is a set of difficulty levels. It assumes that a question consists of a questionary sentence q and a set of information components $A = \{a_1, \ldots, a_n\}$. An information component can be a passage associated with q, candidate answers in multiple-choice QAs, or a video-clip description in video QAs. Then, the difficulty estimation becomes a classification problem in which a classifier $f(\cdot;\theta)$ parameterized by θ determines y^* given q and A. According to Figure 1, q is "Why did Mami experience culture shock in Japan?" and the passage "A Japanese student ..." and five candidate answers become the elements of information component set, A. Then, the classifier f determines the question difficulty given q and A.

The proposed model of which architecture is given in Figure 2 implements $f(\cdot;\theta)$ with two kinds of attention modules. It takes q and A as its input and encodes them using a pre-trained language model. Then, it represents the bi-directional relationships between q and every $a_i \in A$ with the dual multi-head attention and the relationship among a_i's with the transformer encoder. Indeed, the representation of the relationships are accomplished in two steps, since the relationship among a_i's can be expressed after the relationships between q and every $a_i \in A$ are all represented. After that, it predicts the difficult level of q using the relationships.

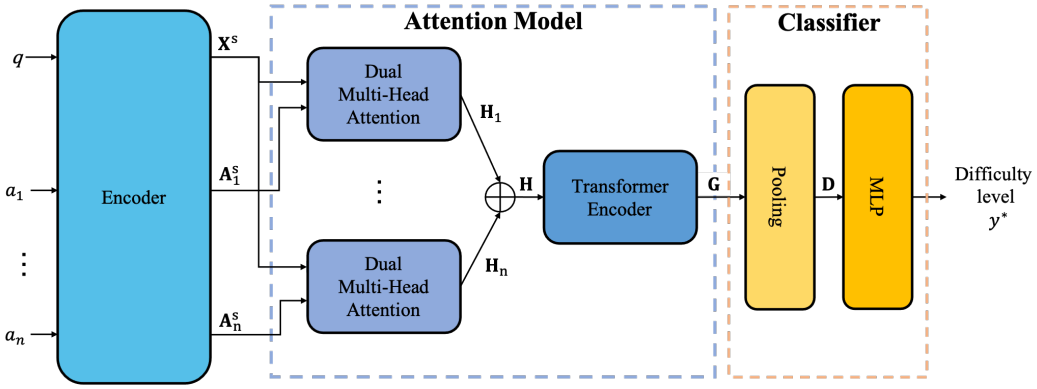


Figure 2. The overall architecture of the proposed model for question difficulty estimation.

3.1. Encoding Question Components

The proposed model first encodes the questionary sentence q and a set of information components $A = \langle a_1, \ldots, a_n \rangle$ into vector representations. As the first step of vector representation, q and all a_i's are expressed in the standard format for BERT [32] using special tokens of [CLS] and [SEP] (This paper assumes that all components in a question are represented in a text form. The question difficulty estimation for the QAs that require analysis of a video or audio stream is out of the scope of this paper). For instance, when q is "Why did Mami experience culture shock in Japan?", it is expressed as "[CLS] why did ma ##mi experience culture shock in japan ? [SEP]". Then, the formatted q and a_i's are encoded into vector representations using the BERT-Base. That is,

$$\begin{aligned} \mathbf{X}^p, \mathbf{X}^s &= BERT(q), \\ \mathbf{A}_i^p, \mathbf{A}_i^s &= BERT(a_i), \end{aligned} \tag{1}$$

where $\mathbf{X}^p$ and $\mathbf{A}_i^p$ are the pooled representations corresponding to the [CLS] token of q and a_i respectively, while $\mathbf{X}^s$ and $\mathbf{A}_i^s$ represent the sequence representations of the whole tokens

in q and a_i. This paper uses only $\mathbf{X}^s$ and $\mathbf{A}_i^s$ in the following steps because the individual tokens deliver more information than the special token in solving QA tasks.

3.2. Representing Relationships Using Attention Model

The attention model is responsible for capturing the relationships between q and A, and the model consists of two attention modules: a dual multi-head co-attention and a transformer encoder based on the multi-head attention. The proposed model first represents the relationships between q and every $a_i \in A$ directly since the questionary sentence q is a key factor in the question-and-answering. Thus, all information components should be represented in accordance with the questionary sentence. However, these representations do not express the relationship among a_i's sufficiently. Although the inter-information among a_i's is reflected indirectly and slightly through the relationships between q and a_i's, a direct inter-information relationship plays an important role in estimating the question difficulty and thus the second attention module is designed to consider the inter-information relationship directly.

In order to identify the bi-directional relationship between q and a_i ($1 \leq i \leq n$), the proposed model adopts the dual multi-head co-attention (DUMA) [18]. DUMA is composed of two multi-head attentions where each multi-head attention captures a single directional attention representation. Thus, it captures both representations from q to a_i and from a_i to q. Then, it fuses these two representations to obtain a final unified representation. That is, the relationship between q and a_i, denoted as $\mathbf{H}_i$, is obtained by applying DUMA to the representations of $\mathbf{X}^s$ and $\mathbf{A}_i^s$ in Equation (1).

$$\begin{aligned} \mathbf{H}_i &= \text{DUMA}(\mathbf{X}^s, \mathbf{A}_i^s) && (2) \\ &= \mathit{Fuse}(\text{MHA}(\mathbf{X}^s, \mathbf{A}_i^s, \mathbf{A}_i^s), \text{MHA}(\mathbf{A}_i^s, \mathbf{X}^s, \mathbf{X}^s)), && (3) \end{aligned}$$

where $\text{MHA}(\cdot,\cdot,\cdot)$ denotes a multi-head attention and $\mathit{Fuse}(\cdot,\cdot)$ is a function for fusing two representations dynamically.

The multi-head attention $\text{MHA}(\cdot,\cdot,\cdot)$ is an attention mechanism to obtain a representation by paying attention jointly to the information from different representations at different positions [33], where the attention is obtained by applying the scaled dot-product attention several times in parallel and then concatenating the results of the attention. Formally, the multi-head attention maps a sequence of query $\mathbf{Q}$ and a set of key-value pairs of $\mathbf{K}$ and $\mathbf{V}$ to a representation by

$$\begin{aligned} \text{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \text{Concat}(\mathit{head}_1, \ldots, \mathit{head}_h) W^O \\ \mathit{head}_i &= \text{Attention}(\mathbf{Q}W_i^{\mathbf{Q}}, \mathbf{K}W_i^{\mathbf{K}}, \mathbf{V}W_i^{\mathbf{V}}), \end{aligned}$$

where $W_i^{\mathbf{Q}}$, $W_i^{\mathbf{K}}$, $W_i^{\mathbf{V}}$, and W^O are all learnable parameters. Here, $\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})$ represents the scaled dot-product attention. It is a weighted sum of the values of which weight is determined by the dot product of the query with all the keys. Thus, it is defined as

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V},$$

where d_k is a key dimensionality that works for a scaling factor.

Among several candidates of $\mathit{Fuse}(\cdot,\cdot)$ function in Equation (3), the performance of using the concatenation is higher than that of using the element-wise summation according to our experiments below (see Section 4.2). This result complies with the results of the previous study by Zhu et al. [18], and thus the concatenation is used as a fuse function in this paper.

After obtaining n $\mathbf{H}_i$'s by applying Equation (3) to $\mathbf{X}^s$ and every $\mathbf{A}_i^s$, the proposed model applies a transformer encoder based on the multi-head attention [33] to them in order to capture inter-information relationship directly. For this, all $\mathbf{H}_i$'s are concatenated

as $\mathbf{H} = [\mathbf{H}_1; \ldots; \mathbf{H}_n]$, and then the transformer encoder is applied to $\mathbf{H}$ to produce the direct representation $\mathbf{G}$ of inter-information relationship. That is,

$$\mathbf{G} = \mathit{TransEncoder}(\mathbf{H}), \tag{4}$$

where *TransEncoder* denotes the transformer encoder. The transformer encoder is a stack of transformer blocks. The l-th transformer block is composed of two layers of a multi-head attention (MHA) and a feed-forward network (FFN). That is, the two layers of $\mathbf{g}^l$ and $\mathbf{h}^l$ are

$$\begin{aligned} \mathbf{g}^l &= \text{LayerNorm}(\text{MHA}(\mathbf{h}^{l-1}, \mathbf{h}^{l-1}, \mathbf{h}^{l-1}) + \mathbf{h}^{l-1}), \\ \mathbf{h}^l &= \text{LayerNorm}(\text{FFN}(\mathbf{g}^l) + \mathbf{g}^l), \end{aligned}$$

where LayerNorm$(\cdot)$ is a layer normalization [34], and $\mathbf{h}^l$ and $\mathbf{h}^{l-1}$ are the outputs of the l-th and $(l-1)$-th transformer block, respectively. The output of the 0-th transformer block is set as $\mathbf{H}$. That is, $\mathbf{h}^0 = \mathbf{H}$.

Note that *TransEncoder* forces every $\mathbf{H}_i$ to consider all other $\mathbf{H}_j$'s ($i \neq j$), since it is based on the self-attention of which query is $\mathbf{H}_i$, and both key and value are other $\mathbf{H}_j$'s. As a result, $\mathbf{G}$ gets able to reflect the inter-information relationship. Therefore, $\mathbf{G}$ becomes the representation that does not reflect only the relationships between the questionary sentence q and information components $a_i \in A$, but also the inter-relationship among all pairs of information components.

3.3. Difficulty Prediction and Implementation

After all relationships between q and A are represented as $\mathbf{G} \in \mathbb{R}^{|hidden| \times |n|}$ where *hidden* is the hidden dimension of *TransEncoder* in Equation (4), the difficulty of a question is determined by a MLP classifier of which input is $\mathbf{G}$. The classifier first summarizes $\mathbf{G}$ into a single dense representation $\mathbf{D}$. There are several operators for this summarization such as max-pooling, average-pooling, and attention. This paper adopts max-pooling for summarizing $\mathbf{G}$ because it is known to be effective in obtaining representative features [35] and shows higher performance than others in our preliminary experiments. After obtaining the final representation $\mathbf{D}$, the MLP predicts the final difficulty level y^* of q. The proposed model is trained to minimize the standard cross-entropy loss.

The proposed model can be applied to most well-known question answering tasks. In the machine reading comprehension tasks such as SQuAD, a question is composed of a questionary sentence, an associated passage, and an answer span. The tasks meet our problem formulation in that the questionary sentence is q, the associated passage is a_1, and the answer span is a_2. Thus, the proposed model can be applied to this type of tasks without any change. In the multiple-choice QAs such as RACE, a question is composed of a questionary sentence, an associated passage, and multiple answer candidates. The difference between the multiple-choice QAs and the machine reading comprehension is that the multiple-choice QAs have multiple answer candidates instead of a single answer. To encode the multiple candidate answers, the proposed model concatenates all candidate answers into one sentence. That is, it regards the multiple candidate answers as one information component. The rest is the same as the machine reading comprehension.

4. Experiments

4.1. Experimental Setting

Three QA tasks are used for the verification of the proposed model: RACE [36], QuAC [37], and DramaQA [9]. RACE is a data set for the multiple choice QA where a question is composed of a questionary sentence, an associated passage, and a set of candidate answers. This data set was collected from English examinations designed for 12∼15-year-old middle school students and those for 15∼18-year-old high school students in China. Thus, there are two subgroups in this data set with a difficulty gap: RACE-M and RACE-H. RACE-M includes middle school examinations and RACE-H contains high school ones. QuAC is a data set for the machine reading comprehension like SQuAD, and

is designed to model information-seeking dialogues. Given a section (in a text form) from a Wikipedia article, two annotators are involved to construct the data set as teacher-student interactions. That is, one annotator (student) asks a sequence of questions to learn about the article, and the other annotator (teacher) answers them by providing excerpts from the article. Since it follows an interactional form, the questions are context-dependent and open-ended so that it is more challenging than SQuAD. On the other hand, the DramaQA data set is constructed for a video QA task to measure the level of machine intelligence for video understanding. It is based on the South Korean television show '*Another Miss Oh*.' Each query in this data set consists of a sequence of video frames, a description of the video frames to deliver background information of the frames, character utterances, and a pair of a questionary sentence and candidate answers. Since this paper assumes that all components in a question are texts, the video frames are excluded from the information components. That is, a question in DramaQA is composed of a questionary sentence, candidate answers, a description of the video frames, and the utterances of the characters. Table 1 summarizes the simple statistics of these data sets.

Table 1. A simple statistics on the data sets used in the experiments.

Data Set	No. of Questions	No. of Information Components
RACE	97,687	2 (passage, candidate answers)
QuAC	7354	2 (passage, candidate answers)
DramaQA	16,191	3 (description, candidate answers, utterance)

DramaQA is manually tagged with four difficulty levels, but RACE and QuAC are not tagged with a difficulty level. Recall that the RACE data set consists of RACE-M and RACE-H. Since RACE-M is about middle school examinations, it is naturally regarded as easy (level 1) questions. RACE-H is then considered as difficult (level 2) questions. For QuAC, we followed the protocol by Gao et al. [19] to label the difficulty level of questions, where the protocol is to assess the difficulty of a question with multiple QA models. This paper employs top three single models (RoBERTa, BERT, and XLNet) from the leaderboard of the QuAC task. A question is labeled as level 1 if at least one model answers it correctly, and is labeled as level 2 if all models give a wrong answer for it. Figure 3 depicts the distributions of difficulty levels in these tasks. In QuAC and DramaQA, the level-1 questions account for about half of the whole questions. On the other hand, the ratio of the level-1 questions is just approximately 20% in RACE.

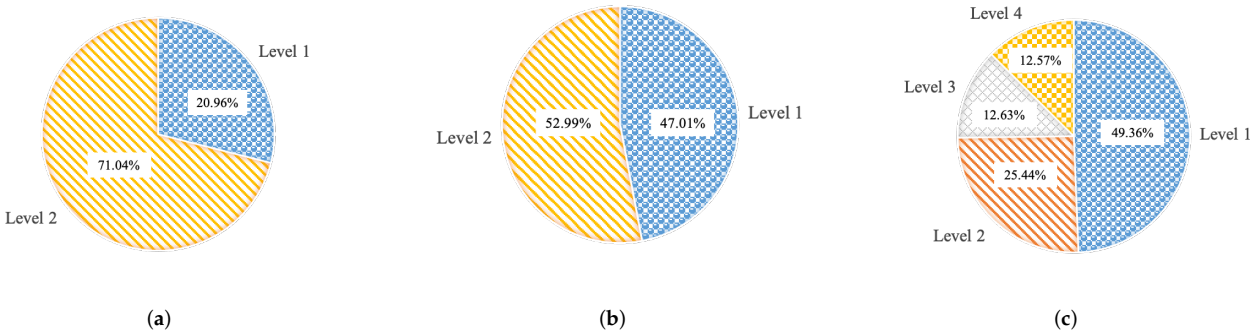


Figure 3. Distributions of difficulty levels in each data set. (**a**) RACE. (**b**) QuAC. (**c**) DramaQA.

For the evaluation of the proposed model, the official data split is used for RACE and DramaQA. In QuAC, the data set is split with the ratio of 80:10:10, where 80% are used for training, 10% are for validation, and the remaining 10% are for test. All hyper-parameters are searched using a grid search and the best hyper-parameters are selected over the validation set. The hyper-parameters used in the experiments are given at Table 2.

BERT-Base model is used for the encoder in Equation (1). The *Fuse* function in Equation (3) is set as the concatenation function. Adam optimizer [38] with default settings is used to train all models, and early stopping over the validation set is executed where 100 is the maximum number of epochs.

Table 2. Parameter values used in the experiments.

	Parameters	**RACE**	**QuAC**	**DramaQA**
Encoder	Model		BERT-Base	
DUMA	Hidden dim.	1536	1536	1536
	No. head	8	6	6
	dropout	0.2	0.2	0.2
	Fuse		*concat*	
TransEncoder	Hidden dim.	3072	3072	4608
	No. head	4	4	4
	No. layers	6	6	6
	dropout	0.2	0.2	0.2

The proposed model is mainly compared with TACNN [14] which is widely used as a main baseline for question difficulty estimation. TACNN uses CNN [39] to obtain the representations of a questionary sentence and a set of information components. Then, it constructs the relationships between the questionary sentence and each information component using a simple attention model, but does not consider the inter-information among the components. Some pre-trained language models are also adopted as baselines of the proposed model, since the language models achieve top performances in many QA tasks. The baseline language models adopted are BERT [32], RoBERTa [40], and XLNet [41]. They concatenate a questionary sentence and all information components with a special token [SEP] and then convert the concatenated sequence to the standard input format of each language model. After that, the formatted sequence is encoded to embedding vectors by each language model. Finally, the embedding vector for the [CLS] token is used to predict the difficulty level in BERT and RoBERTa, while the embedding vector for the last token is used in XLNet. All the models are evaluated with F1-score and accuracy.

4.2. Experimental Results

We first investigate the reliability of labeling the difficulty level on QuAC dataset. The reliability is measured by the agreement between the labels tagged by multiple QA models and the human-annotated labels. To do this, we first randomly sampled 50 data samples. Then, two annotators labeled the difficulty level manually for each sample. The Kappa coefficient [42] between the annotators is 0.52, which falls under the category of 'Moderate'. This implies that the annotators have an agreement to a degree. To obtain the final level of a question from human annotations, we performed an additional procedure as done in the automatic labeling protocol. That is, a question is labeled as level 2 if at least one annotator labels it as level 2 and is labeled as level 1 if both annotators label it as level 1. We have achieved 76% agreement which implies that the labeling of the difficulty level is reliable.

We also investigate the adequateness of the implementation options for $Fuse(\cdot,\cdot)$ in Equation (3) and the direction of the relationships between a questionary sentence and information components. Both have two options. That is, $Fuse(\cdot,\cdot)$ can be implemented by the concatenation or the element-wise summation, and the direction of the relationships can be single or dual. Table 3 summarizes the F1-scores according to the options. The F1-score of the concatenation is generally higher than that of the element-wise summation. Even if the F1-score of the concatenation is 0.26 lower in QuAC, it is much higher in both RACE and DramaQA. Thus, the concatenation is used for $Fuse(\cdot,\cdot)$ in all the experiments below for the sake of consistency.

Table 3. The F1-scores according to the implementation options for $Fuse(\cdot,\cdot)$ and the direction of relationships between a questionary sentence and information components.

Implementation Option		RACE	QuAC	DramaQA
$Fuse(\cdot,\cdot)$	*concat*	**89.56**	70.13	**89.15**
	summation	88.40	**70.39**	88.15
Relationship direction	single (MHA)	89.26	69.74	88.45
	dual (DUMA)	**89.56**	**70.13**	**89.15**

The effectiveness of the bi-directional relationships between a questionary sentence and information components is investigated by replacing DUMA in Equation (2) with a single directional multi-head attention (MHA). That is, $\mathbf{H}_i$, the relationship between a questionary sentence q and each information component a_i, is computed by

$$\mathbf{H}_i = \mathrm{MHA}(\mathbf{X}^s, \mathbf{A}_i^s, \mathbf{A}_i^s).$$

As shown in Table 3, the F1-score of DUMA is higher than that of MHA for all data sets, where the largest difference is 0.7 in DramaQA. This result implies that the bi-directional relationships between a questionary sentence and information components are helpful in improving the performance of question difficulty estimation.

Table 4 compares the performances of the proposed model and its baselines. The first thing to note is that TACNN shows the worst performance in RACE and DramaQA. This is because TACNN does not utilize any pre-trained contextual representation even if the contextual representation is one of the key factors to improve the performance of the natural language tasks. On the other hand, it achieves slightly higher F1-score and accuracy than other pre-trained language models in QuAC. This is due to the fact that TACNN considers the relationships between a questionary sentence and each information component using an attention model explicitly, while the language models do not.

Table 4. Performances of question difficulty estimation.

Data Set	RACE		QuAC		DramaQA	
	Acc. (%)	F1-Score	Acc. (%)	F1-Score	Acc. (%)	F1-Score
BERT	87.75	87.55	58.31	58.25	87.95	87.89
RoBERTa	**89.82**	**89.84**	58.72	58.34	88.81	88.73
XLNet	89.02	88.22	58.51	58.48	89.07	89.05
TACNN	87.27	87.12	60.71	59.87	84.46	84.72
Proposed model	89.81	**89.84**	**68.23**	**68.37**	**89.53**	**89.59**

Among the three pre-trained language models, BERT shows the worst performances for all data sets. RoBERTa and XLNet report similar performances on average. Especially, RoBERTa achieves the best performance in RACE with 89.82% of accuracy and 89.84 of F1-score, respectively. XLNet is the best baseline with 89.07% of accuracy and 89.05 of F1-score in DramaQA. However, the proposed model outperforms all the baselines in QuAC and DramaQA, and achieves a similar performance to RoBERTa in RACE. The F1-score of the proposed model is up to 8.5 higher than those of baselines in QuAC and up to 0.5 higher in DramaQA. These results prove that the proposed model is effective in estimating the difficulty of questions.

4.3. Ablation Study

We investigate the effectiveness of DUMA and *TransEncoder* in the proposed model. Table 5 shows the result of an ablation study over the validation set. The F1-scores of the proposed model over the validation sets of each task are 89.56 for RACE, 70.13 for

QuAC, and 89.15 for DramaQA. The '–' symbol in front of a module indicates exclusion of the module. Thus, '– DUMA' implies that DUMA is excluded from the proposed model. Without DUMA, the F1-score drops up to 4.91 from that of the proposed model, which implies that the bi-directional relationships between a questionary sentence and information components represented by DUMA helps improve the performance of the proposed model.

Table 5. Ablation study of the proposed over validation data.

Model Variations	RACE	QuAC	DramaQA
Proposed model	**89.56**	**70.13**	**89.15**
– DUMA	88.54	65.22	86.77
– TransEncoder	88.94	68.81	88.58
– DUMA and TransEncoder	87.81	63.53	85.38

A similar phenomenon is observed with *TransEncoder* in Equation (4). '– TransEncoder' implies that the concatenation $\mathbf{H}$ of bi-directional relationships $\mathbf{H}_i$'s is directly used as an input of the pooling layer of the final classifier. Its F1-score also drops up to 1.32 from that of the proposed model, which proves the consideration of inter-information relationship is helpful in boosting the performance of the proposed model. In order to take a close look at this result, the F1-scores of each difficulty level are further investigated. Figure 4 depicts the F1-scores for every difficulty level of the questions in DramaQA. When comparing F1-scores of the proposed model with those without *TransEncoder*, the improvement in difficult (level 3 and level 4) questions is larger than that in easy (level 1 and level 2) questions. Especially at level 4, the proposed model achieves 97.40 of F1-score, but the model without *TransEncoder* shows just 94.50. Finally, the model without both DUMA and *TransEncoder* demonstrates the worst performance for all data sets. From these results, we can conclude that the adoption of DUMA for bi-directional relationships and *TransEncoder* for inter-information relationship are effective to predicting the level of question difficulty.

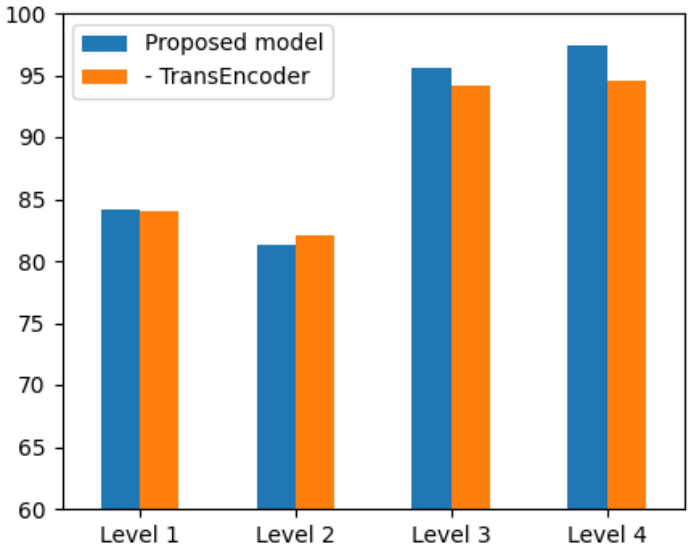
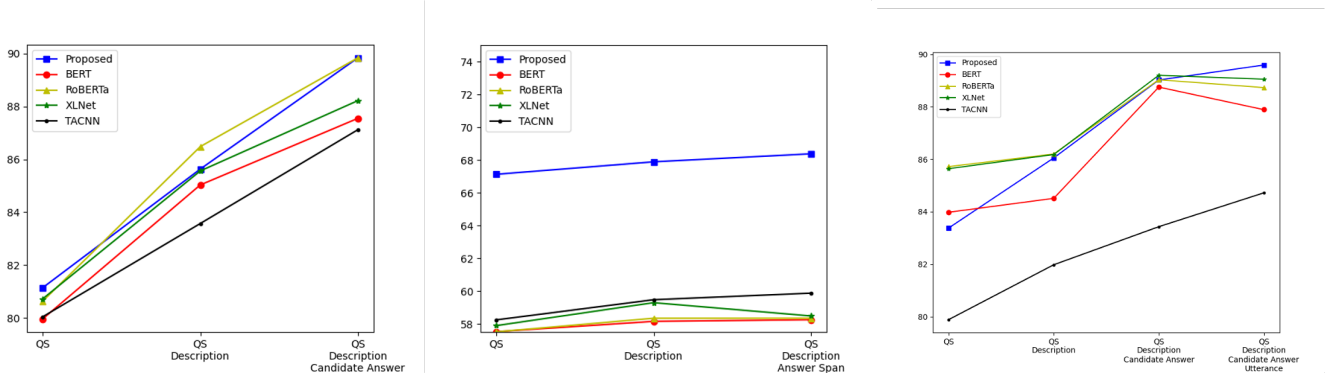


Figure 4. F1-scores for the difficulty levels of the questions in DramaQA.

4.4. Performance Change according to No. of Information Components

There are different numbers of information components depending on the QA tasks and the proposed model is designed to consider a various number of information components. Thus, one consequential question about the proposed model is how the performance

of the proposed model changes as the number of information components increases. Figure 5 depicts the performance changes according to the number of information components. The X-axis of this figure denotes the information components used and the Y-axis represents F1-score. In the QA tasks of our experiments, a description is the most common and important information component. Thus, it is used as a base information component for all models. The candidate answers and utterances are added consecutively in DramaQA, while only the candidate answers are added in RACE and an answer span is added in QuAC. The performances of all models in RACE increase monotonically as new components of candidate answers are added. This result seems natural because a questionary sentence (QS), a description, and candidate answers all provide somewhat information for predicting the level of question difficulty. On the other hand, the performances do not improve large in QuAC even though a new information component of an answer span is added. This is because the answer span is extracted from a description so that the information of the answer span might be already reflected by the description.



(**a**) (**b**) (**c**)

Figure 5. Performance change of the proposed model according to the number of information components. The X-axis denotes the information components used where QS stands for a questionary sentence. (**a**) RACE. (**b**) QuAC. (**c**) DramaQA.

An interesting fact is found in DramaQA. As in RACE, when a description, candidate answers, and utterances are added in order, the performances of the proposed model and TACNN increase monotonically but those of the pre-trained language models do not. Especially when utterances are newly added, the performances of the language models rather decrease. This is because the language models regard all information components as a single sequence, not as individual sequences. Although the sequence differentiates each information component with a special token [SEP], some individuality among the information components might be lost. Due to this loss of individuality, their performances decrease though the utterances are considered. On the other hand, the proposed model and TACNN treat every information component separately. Furthermore, the proposed model is superior to TACNN because it utilizes the pre-trained contextual representations and considers additional inter-information relationship. These results imply that the proposed model predicts the difficulty of questions well even when the number of information components increases.

4.5. Performance of Question Answering with Difficulty Level

In this section, we solve the question answering with a predicted difficulty level to verify that the performance of question answering is improved with the difficulty level. We choose the multi-level context matching model [9] as a question answering model, since is currently the state-of-the-art model for the DramaQA QA task. The multi-level context matching model is designed to understand the multimodal story of a drama. This QA model consists of two streams for a vision and a textual modality. Each stream of modality is combined with embeddings from a questionary sentence and information components

using a context matching module and then predicts a score for each answer. Since it does not adopt any difficulty level, we modify it to use the proposed difficulty level by regarding the difficulty level as an additional modality of the question answering (There will be several methods to utilize the difficulty level in the QA model. However, this experiment focuses on showing that the difficulty level helps the QA model to get better performance than the model without the level). That is, the modified QA model consists of three streams including the difficulty level information.

Table 6 shows the question answering on the drama QA dataset is improved by adding the difficulty level. The '+ Difficulty level' indicates the inclusion of the difficulty level to the multi-level context matching model. The QA model with the difficulty level achieves better performance than the QA model without the level. With the difficulty level, the accuracy rises to 73.83% which is higher up to 2.69% than that of the QA model. These results imply that the question difficulty estimation helps the performance of question answering tasks improved. Especially, the improvement in difficult (level 3 and level 4) questions is larger than that in easy (level 1 and level 2) questions. This is because the proposed method has achieved better performances on difficult questions than on easy questions (refer to Section 4.3 and Figure 4). From these results, we verify the usefulness of the question difficulty estimation.

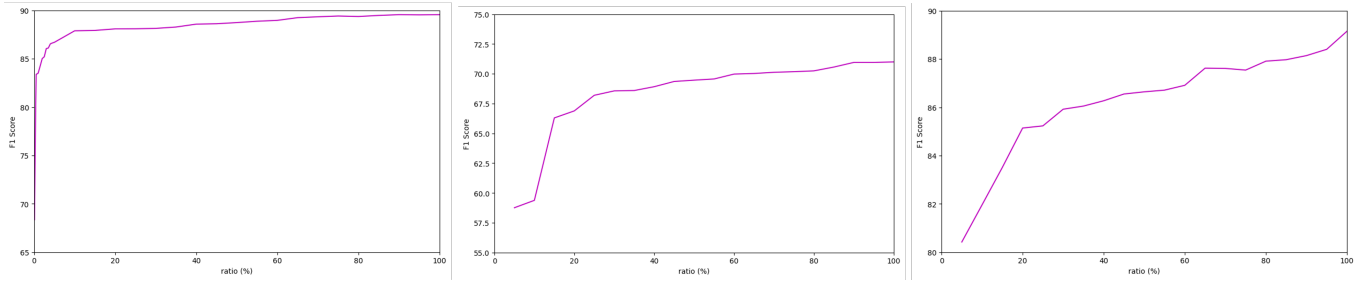
Table 6. Accuracy of question answering with the question difficulty estimation.

QA Model	Diff. 1	Diff. 2	Diff. 3	Diff. 4	Overall	Diff. Avg.
Multi-level context matching model [9]	75.96	74.65	57.36	56.63	71.14	66.15
+ Difficulty level	76.12	74.82	59.12	57.33	73.83	66.85

4.6. Performance Change according to Data Ratio

The proposed model is based on the transformer encoder designed to consider the relationships among all components in a question. It is known that a number of training examples are required to train the transformer encoder. Thus, one possible question about the proposed model is whether the training data in QA data sets are sufficient enough to train it. Since the proposed model adopts a pre-trained language model, BERT-Base, and fine-tunes it, it does not require too many training examples actually. This is proved empirically by showing the performance change according to the ratio of data used to train the proposed model.

Figure 6 depicts the performance changes, where the X-axis is the ratio of data used to train the proposed model and the Y-axis represents F1-score. In all QA data sets, the more the training data are used, the better the predictions of the proposed model are. In QuAC and RACE, the performances of the proposed model converge after 90% data are consumed. This implies that the proposed model is trained well for the data sets. However, a different phenomenon is observed in DramaQA data set with which the performance increases continually. This continual increase is believed to be affected by a larger number of information components in DramaQA. The number of information components in DramaQA is three, while it is two in other data sets. In addition, the F1-score is around 90 when 100% of data are used to train the proposed model. Thus, even if more data are provided, the improvement by them would not be great.



(**a**) (**b**) (**c**)

Figure 6. Performance change model according to the ratio of data sets. The X-axis denotes the ratio of data used to train the proposed model and the Y-axis is F1-score. (**a**) RACE. (**b**) QuAC. (**c**) DramaQA.

5. Conclusions

In this paper, we have proposed an attention model for question difficulty estimation. The proposed attention model first represents bi-directional relationships between a questionary sentence and information components, and then accumulates the inter-information relationship over the concatenated bi-directional relationships. As a result, the proposed method can model complicated relationships among the questionary sentence and information components.

The contributions of this paper are three folds. The first is that the proposed model achieves the state-of-the-art performance in this task. It outperforms the existing model and pre-trained language models. The second is that the proposed model predicts the difficulty of high-level questions accurately. It is required to reason over multiple kinds of information components to predict the difficulty of high-level questions. Since the proposed model is designed to consider the complicated relationships among information components, the reasoning is taken place properly in the proposed model. The last is that the proposed method works efficiently and can be applied to any text-based QA tasks. The proposed method is based on the simple attention model and does not require any other pre-training models except the BERT. Furthermore, it is free from the number of information components.

Through intensive experiments with three well-known QA data sets, it has been shown empirically that the proposed model achieves higher performances than all the previous study and pre-trained language models. Moreover, it is also shown that the proposed attention is essential for accurate prediction of the difficulty level for more difficult questions. Through these experiments, we have proven that the proposed model is plausible for predicting the question difficulty and helps to improve the performances of the question answering.

Author Contributions: Conceptualization, H.-J.S. and S.-H.Y.; methodology, H.-J.S. and S.-H.Y.; visualization, S.-H.Y.; validation, H.-J.S., S.-H.Y. and S.-B.P.; funding acquisition, S.-B.P.; writing—original draft preparation, H.-J.S. and S.-H.Y.; writing—review and editing, H.-J.S. and S.-B.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2017-0-01772, Development of QA systems for Video Story Understanding to pass the Video Turing Test) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A4A1018607).

Institutional Review Board Statement: Not applicable because this study is involved with neither humans nor animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The data can be found here: RACE: https://www.cs.cmu.edu/~glai1/data/race/ (accessed on 1 November

2021), QuAC: https://quac.ai (accessed on 1 November 2021), DramaQA: https://dramaqa.snu.ac.kr (accessed on 1 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv* **2018**, arXiv:1803.05457.
2. Rajpurkar, P.; Jia, R.; Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 784–789. [CrossRef]
3. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.W.; Salakhutdinov, R.; Manning, C.D. HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2369–2380. [CrossRef]
4. Cao, Q.; Trivedi, H.; Balasubramanian, A.; Balasubramanian, N. DeFormer: Decomposing Pre-trained Transformers for Faster Question Answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4487–4497. [CrossRef]
5. Saxena, A.; Tripathi, A.; Talukdar, P. Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4498–4507. [CrossRef]
6. Zhu, M.; Ahuja, A.; Juan, D.C.; Wei, W.; Reddy, C.K. Question Answering with Long Multiple-Span Answers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Online, 16–20 November 2020; pp. 3840–3849. [CrossRef]
7. He, Y.; Zhu, Z.; Zhang, Y.; Chen, Q.; Caverlee, J. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020; pp. 4604–4614. [CrossRef]
8. Heo, Y.J.; On, K.W.; Choi, S.; Lim, J.; Kim, J.; Ryu, J.K.; Bae, B.C.; Zhang, B.T. Constructing Hierarchical Q&A Datasets for Video Story Understanding. *arXiv* **2019**, arXiv:1904.00623.
9. Choi, S.; On, K.W.; Heo, Y.J.; Seo, A.; Jang, Y.; Lee, M.; Zhang, B.T. DramaQA: Character-Centered Video Story Understanding with Hierarchical QA. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; pp. 1166–1174.
10. Collis, K.F. *A Study of Concrete and Formal Operations in School Mathematics: A Piagetian Viewpoint*; Australian Council for Educational Research: Camberwell, Australia, 1975.
11. Bartolo, M.; Roberts, A.; Welbl, J.; Riedel, S.; Stenetorp, P. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 662–678. [CrossRef]
12. Desai, T.; Moldovan, D.I. Towards Predicting Difficulty of Reading Comprehension Questions. In Proceedings of the 32th International Flairs Conference, Melbourne, FL, USA, 21–23 May 2018; pp. 8–13.
13. Ha, L.A.; Yaneva, V.; Baldwin, P.; Mee, J. Predicting the Difficulty of Multiple Choice Questions in a High-stakes Medical Exam. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Florence, Italy, 2 August 2019; pp. 11–20. [CrossRef]
14. Huang, Z.; Liu, Q.; Chen, E.; Zhao, H. Question Difficulty Prediction for READING Problems in Standard Tests. In Proceedings of the 31th AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 1352–1359.
15. Liu, J.; Wang, Q.; Lin, C.Y.; Hon, H.W. Question Difficulty Estimation in Community Question Answering Services. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 85–90.
16. Qiu, Z.; Wu, X.; Fan, W. Question Difficulty Prediction for Multiple Choice Problems in Medical Exams. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 139–148. [CrossRef]
17. Zhuang, Y.; Wang, H. Token-level Dynamic Self-Attention Network for Multi-Passage Reading Comprehension. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2252–2262. [CrossRef]
18. Zhu, P.; Zhao, H.; Li, X. DUMA: Reading Comprehension with Transposition Thinking. *arXiv* **2020**, arXiv:2001.09415.
19. Gao, Y.; Bing, L.; Chen, W.; Lyu, M.R.; King, I. Difficulty Controllable Generation of Reading Comprehension Questions. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 4968–4974.
20. Pasupat, P.; Liang, P. Compositional Semantic Parsing on Semi-Structured Tables. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 1470–1480. [CrossRef]
21. Herzig, J.; Nowak, P.K.; Müller, T.; Piccinno, F.; Eisenschlos, J. TaPas: Weakly Supervised Table Parsing via Pre-training. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4320–4333. [CrossRef]
22. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.

23. Ionescu, B.; Müller, H.; Villegas, M.; de Herrera, A.G.S.; Eickhoff, C.; Andrearczyk, V.; Cid, Y.D.; Liauchuk, V.; Kovalev, V.; Hasan, S.A.; et al. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Avignon, France, 10–14 September 2018; pp. 309–334.
24. Ye, Y.; Zhang, S.; Li, Y.; Qian, X.; Tang, S.; Pu, S.; Xiao, J. Video question answering via grounded cross-attention network learning. *Inf. Process. Manag.* **2020**, *57*, 102265. doi: 10.1016/j.ipm.2020.102265. [CrossRef]
25. Xue, K.; Yaneva, V.; Christopher Runyon, P.B. Predicting the Difficulty and Response Time of Multiple Choice Questions Using Transfer Learning. In Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, WA, USA, 10 July 2020; pp. 193–197. [CrossRef]
26. Zheng, J.; Cai, F.; Chen, H.; de Rijke, M. Pre-train, Interact, Fine-tune: A novel interaction representation for text classification. *Inf. Process. Manag.* **2020**, *57*, 102215. doi: 10.1016/j.ipm.2020.102215. [CrossRef]
27. Cao, N.D.; Aziz, W.; Titov, I. Question answering by reasoning across documents with graph convolutional networks. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 2306–2317.
28. Song, L.; Wang, Z.; Yu, M.; Zhang, Y.; Florian, R.; Gildea, D. Exploring Graph-structured Passage Representation for Multi-hop Reading Comprehension with Graph Neural Networks. *arXiv* **2018**, arXiv:1809.02040.
29. Wang, Y.; Liu, K.; Liu, J.; He, W.; Lyu, Y.; Wu, H.; Li, S.; Wang, H. Multi-passage machine reading comprehension with cross-passage answer verification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1918–1927. [CrossRef]
30. Dehghani, M.; Azarbonyad, H.; Kamps, J.; de Rijke, M. 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, Melbourne, Australia, 11–15 February 2019; pp. 681–689. [CrossRef]
31. Zhong, V.; Xiong, C.; Keskar, N.S.; Socher, R. Coarse-grain fine-grain coattention network for multi-evidence question answering. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
32. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
34. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
35. Scherer, D.; Müller, A.; Behnke, S. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In Proceedings of the 20th International Conference on Artificial Neural Networks, Thessaloniki, Greece, 15–18 September 2010; pp. 92–101.
36. Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; Hovy, E. RACE: Large-scale Reading Comprehension Dataset From Examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 785–794. [CrossRef]
37. Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.T.; Choi, Y.; Liang, P.; Zettlemoyer, L. QuAC: Question Answering in Context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2174–2184. [CrossRef]
38. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
39. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
40. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
41. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32* , 5753–5763.
42. Carletta, J. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Comput. Linguist.* **1996**, *22*, 249–254.