Question Difficulty Prediction for Multiple Choice Problems in Medical Exams

Conference Paper · November 2019

DOI: 10.1145/3357384.3358013

3 authors, including:

Zhaopeng Qiu
Peking University
24 PUBLICATIONS **507** CITATIONS

SEE PROFILE

Question Difficulty Prediction for Multiple Choice Problems in Medical Exams

Zhaopeng Qiu
Tencent Medical AI Lab
zhaopengqiu@tencent.com

Xian Wu
Tencent Medical AI Lab
kevinxwu@tencent.com

Wei Fan
Tencent Medical AI Lab
davidwfan@tencent.com

ABSTRACT

In the ITS (Intelligent Tutoring System) services, personalized question recommendation is a critical function in which the key challenge is to predict the difficulty of each question. Given the difficulty of each question, ITS can allocate suitable questions for students with varied knowledge proficiency. Existing approaches mainly relied on expert labeling, which is both subjective and labor intensive. In this paper, we propose a Document enhanced Attention based neural Network(DAN) framework to predict the difficulty of multiple choice problems in medical exams. DAN consists of three major steps: (1) In addition to stem and options, DAN retrieves relevant medical documents to enrich the content of each question; (2) DAN breaks down the question's difficulty into two parts: the hardness for recalling the knowledge assessed by the question and the confusion degree to exclude distractors. For each part, DAN introduces corresponding attention layers to model it; (3) DAN combines two parts of difficulties together to predict the overall difficulty. We collect a real-world data set from one of the largest medical online education websites in China. And the experimental results demonstrate the effectiveness of the proposed framework.

CCS CONCEPTS

• **Applied computing → Education**.

KEYWORDS

educational mining, question difficulty prediction, document retrieval, attention

ACM Reference Format:
Zhaopeng Qiu, Xian Wu, and Wei Fan. 2019. Question Difficulty Prediction for Multiple Choice Problems in Medical Exams. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM'19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3358013

1 INTRODUCTION

Intelligent Tutoring System(ITS) services are widely adopted in a broad range of domains. For example, Duolingo (a platform for learning English) attracted 300 million active users. Among all application domains, medical ITS is one of the most appealing ones where doctors can consolidate knowledge and develop expertise.

In ITS services, personalized exercise recommendation is an essential function which can help students improve study efficiency and enhance user experience. A key challenge in personalized exercise recommendation is to predict the difficulty of each question. The difficulty of a question refers to the percentage of students who answer this question wrongly. Given the difficulties of questions, ITS can recommend suitable questions for students with varied knowledge proficiency. Furthermore, ITS can automatically compose a discrimination test by selecting questions at different difficulty levels.

However, the question difficulty is not known before students actually take it. To estimate the difficulty in advance, a straightforward way is to ask teachers or experts to label according to their experience. In practice, such manual estimation is not feasible. First, this manner is labor intensive and hard to scale, especially for ITS services with massive questions; Second, teachers or experts label according to their subjective opinions, which may lead to the biased and misleading results[6]. Another manner is to sample a small number of students and use their error rate to estimate the difficulty. Although such a manner is less biased than the first one, it is also labor intensive and brings challenges in the student sample strategy.

Recently, non-human based and data-driven solutions emerge. For example, [6] focuses on reading comprehension problems in standard English tests, which is similar to the problems in SQuAD contest[1]. [6] utilizes the reading passage, question, and options together to predict the question's difficulty. However, it is non-trivial to directly apply [6] to the multiple choice problems (MCP) in medical exams. As in reading comprehension problems, the answers to questions can be inferred from the given passages, which also means that the given passages are vital for this difficulty prediction solution. In other words, the required knowledge is self-contained. While in medical exams, except for the stem and options, no background knowledge is given.

The MCP is a typical problem style in medical exams, and the National Medical Licensing Examination in China uses the MCP as the primary question type. Figure 1 shows two examples. Each MCP only contains a question text and five candidate answers with four distractors.

To estimate the question difficulty for MCPs in medical exams, we propose a novel **D**ocument enhanced **A**ttention based neural **N**etwork (DAN) framework. DAN consists of three major components: (1) We build a database of medical papers and textbooks. Given a question, we use its stem and options to compose queries and retrieve relevant medical documents to enrich the context. Then

[1] https://rajpurkar.github.io/SQuAD-explorer/

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3358013

1. The preferred test for diagnosis of heart failure is:
A. Chest X-ray
B. Echocardiogram
C. Left ventricular angiography
D. Stress test
E. ECG

Answer: C

2. Female, 62 years old. Hypertensive patients, suddenly palpitations, shortness of breath, coughing up pink tinged foam sputum. Physical examination: BP 200/126 mmHg, heart rate 146 beats/min. In addition to other treatments, which of the following drugs should be used:
A. Lanatoside C, Nitroglycerin, ISOprenaline
B. Strophanthink, Sodium nitroprusside, Propranolol
C. Guanethidine, Phentolamine, Lanatoside C
D. Sodium nitroprusside, Lanatoside C, Furosemide
E. Nitroglycerin, Lanatoside C, Dopamine

Answer: D

Figure 1: Two examples of multiple choice problems in medical exams.

we leverage a BiLSTM-based architecture to generate the semantic representations for all text materials (i.e., the stem, option, and retrieved medical text); (2) We break down the question's difficulty into two parts: the confusion and recall. Confusion refers to the difficulty to separate the correct answer from the distractors, while recall refers to the difficulty to recall the knowledge assessed by the question. Next, we introduce two attention layers to model two types of difficulties, respectively; (3) We combine two types of difficulties to predict the overall difficulty. We collect a real-world data set from one of the largest medical online education websites in China. Moreover, the experimental results demonstrate the effectiveness of our proposed framework. As far as we know, this is the first comprehensive data-driven solution to question difficulty prediction task for multiple choice problems in medical exams.

2 RELATED WORKS

Question difficulty has been widely studied in educational psychology. In recent years, some research efforts have devoted to machine learning based question difficulty prediction. In this section, we discuss two categories of related work.

Question Difficulty in Educational Psychology. On the one hand, Susanti[16] investigated the relations between several factors of questions and the corresponding question difficulty. On another hand, question difficulty prediction has been widely studied in some theories. *Classical test theory (CTT)* is a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers. *CTT* utilizes statistical methods to predict question difficulty. *Item response theory (IRT)* is another theory, based on the application of related mathematical models, evaluating the latent traits of the people and questions. The latent traits of questions contain the question difficulty. The Rasch model, a particular case of *IRT*, is a probabilistic model and evaluates question difficulty from examinees' responses modeled by a logistic-like function. All of Gajjar[2], Rao[15], and Luger[10] utilize the Rasch model and student feedback to predict question difficulty.

However, the common limitation of these works is that they are all labor intensive. Hence none of these works apply to the ITS services, which have massive questions. Differently, our work is an entirely data-driven solution.

Text-based Question Difficulty Prediction. Considerable research efforts[8, 9, 11, 13, 25] have been devoted recently to using NLP (Natural Language Processing) methods to predict question difficulty. Loukina et al.[9] have revealed that a system based on multiple text complexity features, such as word unfamiliarity and the average frequency of long sentences, can predict question difficulty. Ulrike et al.[13] approximated the question difficulty by the amount of variation in student answers. They measured the answer variation as the average similarity of student answers among themselves or their average similarity with the reference answer.

These works all required the manual design of textual features, which are vital issues for these solutions. However, not all these features are suitable for other applications. Differently, our work is an end-to-end framework, which needs no other manual designed features.

The work closest to ours is that of Huang et al.[6]. They proposed a non-human based and data-driven solution focusing on question difficulty prediction for reading comprehension problems in standard English tests. This work utilized the reading passage, the question, and the options to predict the question difficulty. This solution first utilized a CNN-based architecture to extract sentence representations for the questions. Then, it used an attention strategy to qualify the difficulty contribution of each sentence in reading passage and options. Finally, it aggregated the semantic representations of the documents, the questions, and the options to predict the question difficulty.

In reading comprehension problems, the answers to questions can be inferred from the given passages, which also means that the given passages are vital for this difficulty prediction solution. While in medical exams, except for the stem and options, no reading passage is given. Hence, it is non-trivial to directly apply this solution to difficulty prediction for the MCPs in medical exams. Nevertheless, this solution used CNN, which is hard to capture the information in long-range contexts, to encode the text.

3 DAN FRAMEWORK

In this section, we first introduce several basic concepts used in this paper. Then we formally define the problem of question difficulty prediction. Finally, we present the technical details of our question difficulty prediction framework DAN.

3.1 Problem Definition and Framework Overview

In this paper, we focus on the question difficulty prediction for MCPs (see Figure 1) in medical exams. Let $\mathbb{Q}$ denote the set of medical questions. Each question $Q \in \mathbb{Q}$ has a difficulty attribute P obtained from student test logs, a correct answer A, and four distractors $\{C_1, C_2, C_3, C_4\}$.

Problem Definition. *Formally, given the question set $\mathbb{Q}$, the goal is to leverage all questions $Q \in \mathbb{Q}$ to train a model $\mathcal{M}$ (i.e., DAN) which can be used to estimate the difficulties for questions in the newly-conducted recommendation.*

As shown in Figure 2, DAN is a two-stage framework. The first stage is a document retrieval system, which is used to retrieve medical documents possibly related to the questions. The second stage

is a neural network based question difficulty prediction model. We will address each component in detail in the following subsections.

3.2 Document Retrieval

When the students try to solve the medical questions, they will first recall the medical documents related to the knowledge assessed by the questions. Then they can infer the correct answers according to the remembered medical documents. DAN also first uses the document retrieval module to retrieve the medical documents related to the questions to enrich the contexts of questions to simulate human behaviors. Considering that the candidate answer is relatively short and may not contain enough information for question difficulty prediction, we concatenate the question and each candidate answer as a statement. Given a question Q and one of its candidate answers (A or C_i), we append the correct answer to the question to form a question-answer statement $S_a = Q + A$ or append a distractor to the question to form a question-distractor statement $S_i = Q + C_i (i \in [1, 4])$. Then we use a statement as a query to perform an Elastic Search[3] based retrieval over the medical materials. The retrieved documents related to S_a are denoted as $D^a = \{D_1^a, D_2^a, ..., D_N^a\}$ and the retrieved documents related to S_i are denoted as $D^{c_i} = \{D_1^{c_i}, D_2^{c_i}, ..., D_N^{c_i}\}$, where N stands for the number of retrieved documents. The retrieved documents D^a and D^{c_i} are the N most related to S_a and S_i documents, respectively.

Inspired by some machine reading comprehension works [4, 22, 24], we use BM25 to measure the relevance between the question-answer statement and the medical material. The BM25 measures the relevance scores according to common words between the queries and documents, with the consideration of the word importance. Moreover, the most important words in questions and medical documents are medical entities, which also always have larger relevance scores for queries. Hence, the retrieved documents in general contain some common medical entities with the questions, which guarantee the semantic relevance between the questions and the retrieved documents.

3.3 The Prediction Model

The prediction model is designed to predict the question difficulty given a question Q with its candidate answers (A and $\{C_i\}_{i=1}^{i=4}$) and the relevant medical materials (D^a and $\{D^{c_i}\}_{i=1}^{i=4}$).

As shown in Figure 2, the model contains two key modules which evaluate the following two types of difficulties, respectively.

- *Recall difficulty*: When a student starts to solve a medical question, she/he will first probe into her/his memory to recall the knowledge related to this question. If the medical books or the teachers have mentioned the knowledge assessed by this question multiple times, the student will be familiar with the assessed knowledge and further can evoke a large amount of relevant knowledge, which can help her/him to solve this question easily. In contrast, if the medical books or the teachers rarely mention the knowledge assessed by this question, it will be hard for the student to solve this question. We call the difficulty caused by the rareness of the assessed knowledge as recall difficulty. To measure the recall difficulty of a question Q, we evaluate the semantic relevance between the retrieved medical documents D^a (like the knowledge recalled by the students) and the question-answer statement S_a in our proposed framework.
- *Confusion difficulty*: In each MCP, there are four distractors. In order to examine the students, question writers are instructed to make their distractors plausible yet clearly incorrect. The surface plausibility that the question writer has intentionally built into distractors will confuse the students and hinder them from excluding the distractors. We call the difficulty caused by the distractors as confusion difficulty. To measure the confusion difficulty of a question Q, we evaluate the semantic similarity between the question-distractor statements $\{S_i\}_{i=1}^{i=4}$ and the question-answer statement S_a in our proposed framework.

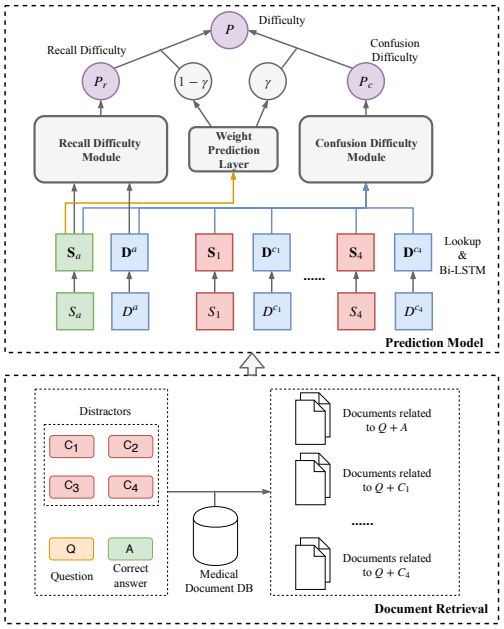


Figure 2: DAN framework.

The model first uses Bi-LSTM to encode each text sequence. Then the model will encode two types of difficulties by two modules: 1) confusion difficulty module; 2) recall difficulty module. Finally, the model aggregates two types of difficulties to predict the overall question difficulty.

Encoding Layer. Given a question-answer statement $S_a = \{w_t^a\}_{t=1}^{t=L_Q}$ and a medical document $D_j^a = \{w_t^m\}_{t=1}^{t=L_D}$ of N relevant documents, we first convert every word w to its d-dimensional vector $\mathbf{e}$ via an embedding matrix $\mathbf{E} \in \mathbb{R}^{|V| \times d}$, where V is the vocabulary. Then we use the Bi-LSTM to extract the contextual representation for each word.

The outputs of the bi-directional LSTM are two matrices: $\mathbf{S}_a \in \mathbb{R}^{L_Q \times d}$ and $\mathbf{D}_j^a \in \mathbb{R}^{L_D \times d}$. For each question-distractor statement S_i and its relevant medical document $D_j^{c_i}$, we also can obtain two matrices $\mathbf{S}_i \in \mathbb{R}^{L_Q \times d}$ and $\mathbf{D}_j^{c_i} \in \mathbb{R}^{L_D \times d}$ via the same encoding process. L_Q and L_D are the maximum lengths of $S_a(S_i)$ and $D_j^a(D_j^{c_i})$, respectively, and d is the dimension of word embedding.

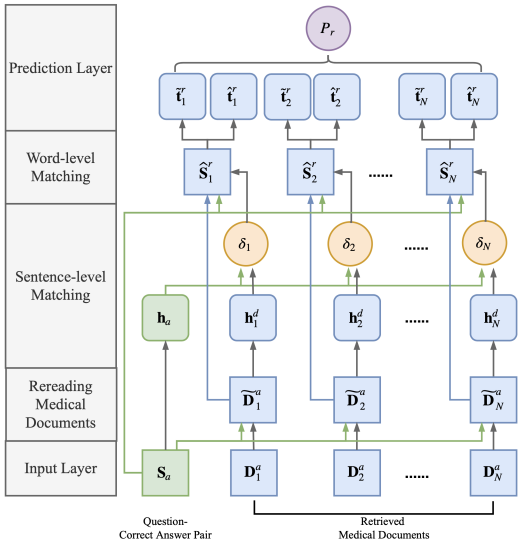


Figure 3: The recall difficulty module.

Recall Difficulty Module. As mentioned above, if the medical books rarely mention the knowledge assessed by a question, it will be hard for the students to solve this question. In contrast, if the retrieved medical documents have a strong correlation with the question, it will mean that the question is mentioned by the medical books multiple times, which will also mean that the students can recall easily the knowledge assessed by the question. Hence, we evaluate the semantic relevance between the retrieved medical documents $\mathbf{D}^a$ (like the knowledge recalled by the student) and the question-answer statement $\mathbf{S}_a$ to measure the recall difficulty. Figure 3 shows the recall difficulty module we proposed.

The inputs of this module are the contextual representations of the question-answer statement and its retrieved documents, i.e., $\mathbf{S}_a$ and $\{\mathbf{D}_j^a\}_{j=1}^{j=N}$. The similarity between $\mathbf{S}_a$ and $\{\mathbf{D}_j^a\}_{j=1}^{j=N}$ is calculated by the following four steps.

(1) Rereading medical documents

There exists a semantic gap between the medical documents and the question because they are collected from different sources. Hence, the medical documents first are reread with the help of the question-answer statement to gain a deeper understanding. Inspired by some machine reading comprehension works[14, 17, 19, 22, 24], we use the attention mechanism to incorporate the question-answer information into the medical documents and fuse the question-aware representations and original representations for better semantic understanding.

Given input matrices $\mathbf{U} \in \mathbb{R}^{L_U \times d}$ and $\mathbf{V} \in \mathbb{R}^{L_V \times d}$, the attention function is defined as

$$\alpha_{i,j} = \frac{\mathbf{U}_{i:} \circ \mathbf{V}_{:j}}{\sqrt{d}}, \mathbf{M} = \texttt{Attn}(\mathbf{U}, \mathbf{V}) = \left[\frac{\exp(\alpha_{ij})}{\sum_i \exp(\alpha_{ij})}\right]_{i,j} \quad (1)$$

where $\circ$ denotes the element-wise multiplication operation, and $\mathbf{M} \in \mathbb{R}^{L_U \times L_V}$ denotes the attention weight matrix.

We first introduce the attention layer to align the question-answer statement representation $\mathbf{S}_a$ against each medical document representation $\mathbf{D}_j^a$. Then we introduce the fusion layer to combine the original document representation and the question-aware representation together to form a new semantic representation, i.e., $\widetilde{\mathbf{D}}_j^a$. We adopt the fusion kernel used in recent works[1, 12] for better semantic understanding.

$$\overline{\mathbf{D}}_j^a = \mathbf{M}^{d_j}\mathbf{S}_a, \mathbf{M}^{d_j} = \texttt{Attn}(\mathbf{D}_j^a, \mathbf{S}_a) \quad (2)$$

$$\widetilde{\mathbf{D}}_j^a = \texttt{Fuse}(\mathbf{D}_j^a, \overline{\mathbf{D}}_j^a) = tanh([\mathbf{D}_j^a; \overline{\mathbf{D}}_j^a; \mathbf{D}_j^a \circ \overline{\mathbf{D}}_j^a; \mathbf{D}_j^a - \overline{\mathbf{D}}_j^a]\mathbf{W}_f + \mathbf{b}_f) \quad (3)$$

where $\mathbf{W}_f \in \mathbb{R}^{4d \times d}$ and $\mathbf{b}_f \in \mathbb{R}^d$ are the parameters to learn. $[;]$ denotes the column-wise concatenation and $-$ denotes the element-wise subtraction between two matrices.

In order to predict the recall difficulty, we introduce two layers to collect the sentence-level and word-level similarity information, respectively. Intuitively, the retrieved medical documents may only contain some keywords related to the question, but not be relevant to the question at the sentence level. Hence, in addition to the fine-grained word-level similarity information, we compare the sentence representations of the medical documents and the question-answer statement to collect the sentence-level similarity information in order to void producing biased results misdirected by some words.

(2) Sentence-level matching

For each medical document $\widetilde{\mathbf{D}}_j^a$, we use two following steps to calculate the sentence-level semantic similarity between it and the question-answer statement.

First, both the document's representation $\widetilde{\mathbf{D}}_j^a$ and the question-answer statement's representation $\mathbf{S}_a$ are self-aligned to obtain sentence-level representations $\mathbf{h}_j^d \in \mathbb{R}^d$ and $\mathbf{h}_a \in \mathbb{R}^d$, respectively.

$$r_k^a = ReLU(\mathbf{S}_{a,k:}\mathbf{W}_g^a + b_g^a), r_k^a = \frac{\exp(r_k^a)}{\sum_k \exp(r_k^a)} \quad (4)$$

$$\mathbf{h}_a = \sum_k r_k^a \mathbf{S}_{a,k:} \quad (5)$$

$$r_{j,k}^d = ReLU(\widetilde{\mathbf{D}}_{j,k:}^a \mathbf{W}_g^d + b_g^d), r_{j,k}^d = \frac{\exp(r_{j,k}^d)}{\sum_k \exp(r_{j,k}^d)} \quad (6)$$

$$\mathbf{h}_j^d = \sum_k r_{j,k}^d \widetilde{\mathbf{D}}_{j,k:}^a \quad (7)$$

where $\mathbf{W}_g^a \in \mathbb{R}^{d \times 1}$, $\mathbf{W}_g^d \in \mathbb{R}^{d \times 1}$, $b_g^a \in \mathbb{R}$, and $b_g^d \in \mathbb{R}$ are trainable parameters. $\mathbf{S}_{a,k:}$ and $\widetilde{\mathbf{D}}_{j,k:}^a$ denote the k-th row of $\mathbf{S}_a$ and the k-th row of $\widetilde{\mathbf{D}}_j^a$, respectively.

After that, the retrieved medical documents are considered one by one to capture the sentence-level correlation information with the question-answer statement.

$$\delta_j = \mathbf{h}_a \mathbf{W}_p \mathbf{h}_j^d + b_p \quad (8)$$

where $\mathbf{W}_p$ is a trainable bilinear projection matrix, and $b_p \in \mathbb{R}$ is also a parameter to learn. δ_j is the sentence-level semantic similarity between the question-answer statement and its j-th retrieved medical document.

(3) Word-level matching

In this layer, the retrieved medical documents $\{\widetilde{\mathbf{D}}_j^a\}_{j=1}^{j=N}$ are one-by-one compared with the question-answer statement $\mathbf{S}_a$ at the word level to collect the fine-grained semantic similarity information. Specifically, for each medical document $\widetilde{\mathbf{D}}_j^a$, we perform

the following three operations to calculate the semantic similarity between it and the question-answer statement $\mathbf{S}_a$ at the word-level.

First, we use the attention function $\texttt{Attn}(\cdot,\cdot)$ to align the retrieved medical document $\widetilde{\mathbf{D}}_j^a$ to the question-answer statement $\mathbf{S}_a$.

$$\mathbf{F}_j^d = \texttt{Attn}(\mathbf{S}_a, \widetilde{\mathbf{D}}_j^a)\widetilde{\mathbf{D}}_j^a \tag{9}$$

Secondly, we collect the similarity information at each position of the question-answer statement.

$$\overline{\mathbf{S}}_j^r = ReLU(\overline{\mathbf{S}}_j^r \mathbf{W}_m^r + \mathbf{b}_m^r), \overline{\mathbf{S}}_j^r = [\mathbf{S}_a - \mathbf{F}_j^d; \mathbf{S}_a \circ \mathbf{F}_j^d] \tag{10}$$

where $\mathbf{W}_m^r \in \mathbb{R}^{2d\times d}$ and $\mathbf{b}_m^r \in \mathbb{R}^d$ are trainable parameters.

Finally, we combine the sentence-level similarity information and word-level similarity information to obtain an overall comprehensive representation of the semantic similarity between the question-answer statement and its j-th retrieved medical document.

$$\widehat{\mathbf{S}}_j^r = \delta_j \widetilde{\mathbf{S}}_j^r \tag{11}$$

(4) Prediction layer

In this layer, we first use mean-pooling and max-pooling to fuse the similarity information of all positions of the question-answer statement. For $\widehat{\mathbf{S}}_j^d$,

$$\tilde{\mathbf{t}}_j^r = \texttt{MeanPooling}(\widehat{\mathbf{S}}_j^r), \hat{\mathbf{t}}_j^r = \texttt{MaxPooling}(\widehat{\mathbf{S}}_j^r) \tag{12}$$

Then we concatenate the similarity information from all retrieved medical documents (i.e., $\{\tilde{\mathbf{t}}_j^r\}_{j=1}^{j=N}$ and $\{\hat{\mathbf{t}}_j^r\}_{j=1}^{j=N}$) to predict the recall difficulty P_r.

$$\mathbf{t}^r = ReLU([\{\tilde{\mathbf{t}}_j^r\}_{j=1}^{j=N}; \{\hat{\mathbf{t}}_j^r\}_{j=1}^{j=N}]\mathbf{W}_1^r + \mathbf{b}_1^r) \tag{13}$$

$$P_r = Sigmoid(\mathbf{t}^r \mathbf{W}_2^r + b_2^r) \tag{14}$$

where $\mathbf{W}_1^r \in \mathbb{R}^{2Nd\times h}$, $\mathbf{W}_2^r \in \mathbb{R}^{h\times 1}$, $\mathbf{b}_1^r \in \mathbb{R}^h$, and $b_2^r \in \mathbb{R}$ are trainable parameters.

Confusion Difficulty Module. The confusion difficulty represents the degree of interference caused by four distractors. Rather than merely utilizing the low level "literal similarities" to measure the degree of interference, we first use the relevant medical documents to enrich the semantic representations of the candidate answers and then measure the semantic similarities among the candidate answers to measure the degree of interference. Figure 4 shows the confusion difficulty module.

As the recall difficulty module, the confusion difficulty module also collect the semantic similarity information to measure a part of question difficulty and use the $\mathbf{S}_a$ as an input. However, there are some differences between other inputs of two modules, i.e., the four question-distractor statements $\{\mathbf{S}_i\}_{i=1}^{i=4}$ and the retrieved medical documents $\mathbf{D}^a$. Hence, there are also some differences between the two modules' architectures.

- There exists a semantic gap between the medical documents and the question-answer statement because they are gathered from different sources. In contrast, the question-distractor statements have the same source as the question-answer statement. Nevertheless, the surface plausibility has guaranteed that each distractor has close semantic relevance to the correct answer. Hence, we do not need to leverage the question-answer statement to reread the question-distractor statements in the confusion difficulty module.

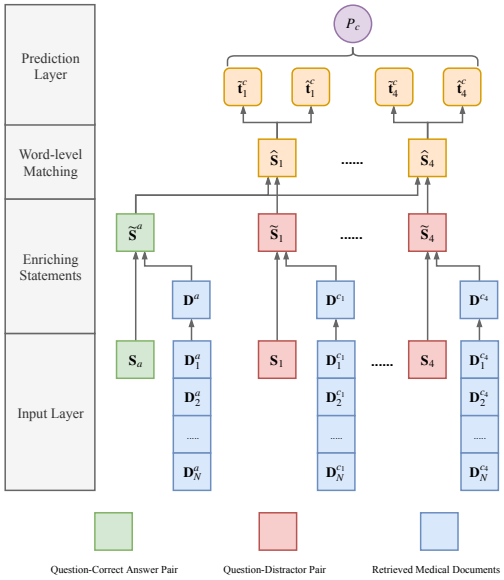


Figure 4: The confusion difficulty module.

- Each distractor has nature semantic relevance to the correct answer as mentioned above. Meanwhile, most of the distractors only have a short sequence, as the two examples shown in Figure 1. Hence, we only collect semantic similarity information at the word-level in confusion difficulty module.

The inputs of this module are the contextual representations of question-candidate answer statements and retrieved medical documents, i.e., $\mathbf{S}_a$, $\{\mathbf{S}_i\}_{i=1}^{i=4}$, $\mathbf{D}^a$, and $\{\mathbf{D}^{c_i}\}_{i=1}^{i=4}$. As the recall difficulty module, we compare all question-distractor statements one by one with the question-answer statement to collect the semantic similarity information. For simplicity, we consider the question-distractor statement $\mathbf{S}_i$ as a case in the following details. The similarity between $\mathbf{S}_a$ and $\mathbf{S}_i$ is calculated by the following three steps.

(1) Enriching statements

In order to enrich the semantic representation of each candidate answer, we first use the attention mechanism to attend the relevant medical documents to candidate answers. We use the question-answer statement $\mathbf{S}_a$ and its relevant documents $\mathbf{D}^a$ as an example to demonstrate the attention process:

First, we join all documents $\{\mathbf{D}_i^a\}_{i=1}^{i=N}$ together to form a large matrix.

$$\mathbf{D}^a = [\mathbf{D}_1^a|\mathbf{D}_2^a|...|\mathbf{D}_N^a] \tag{15}$$

where $[|]$ denotes the row-wise concatenation.

We secondly leverage the attention function $\texttt{Attn}(\cdot,\cdot)$ to incorporate the document information into the question-answer context, and obtain the document-aware statement representation $\overline{\mathbf{S}}_a$.

$$\overline{\mathbf{S}}_a = \mathbf{M}^a\mathbf{D}^a, \mathbf{M}^a = \texttt{Attn}(\mathbf{S}_a, \mathbf{D}^a) \tag{16}$$

Finally, we combine the original vectors $\mathbf{S}_a$ and the document-aware vectors $\overline{\mathbf{S}}_a$ together to form an enriched semantic representation of the question-answer statement.

$$\widetilde{\mathbf{S}}_a = \texttt{Fuse}(\mathbf{S}_a, \overline{\mathbf{S}}_a) \tag{17}$$

where $\texttt{Fuse}(\cdot,\cdot)$ is the same as the fusion kernel used by the recall difficulty module.

Likewise, we also can attend the knowledge of relevant documents to the question-distractor statements $\{\mathbf{S}_i\}_{i=1}^{i=4}$ to obtain enriched representations $\{\widetilde{\mathbf{S}}_i\}_{i=1}^{i=4}$.

(2) Word-level matching

Inspired by previous works[14, 23], the correct answer is compared with all distractors one-by-one to collect the information describing the extent of pairwise interference. Specifically, for the question-answer statement $\widetilde{\mathbf{S}}_a$, the interference information gathered from the question-distractor statement $\widetilde{\mathbf{S}}_i$ is computed as:

$$\dot{\mathbf{S}}_i = [\widetilde{\mathbf{S}}_a - \mathbf{F}_i^c; \widetilde{\mathbf{S}}_a \circ \mathbf{F}_i^c], \mathbf{F}_i^c = \texttt{Attn}(\widetilde{\mathbf{S}}_a, \widetilde{\mathbf{S}}_i)\widetilde{\mathbf{S}}_i \tag{18}$$

$$\widehat{\mathbf{S}}_i = ReLU(\dot{\mathbf{S}}_i \mathbf{W}_m^c + \mathbf{b}_m^c) \tag{19}$$

where $\mathbf{W}_m^c \in \mathbb{R}^{2d\times d}$ and $\mathbf{b}_m^c \in \mathbb{R}^d$ are trainable parameters. $\widehat{\mathbf{S}}_i$ represents the interference information gathered from the i-th distractor.

(3) Prediction layer

In this layer, we first use row-wise mean-pooling and row-wise max-pooling to fuse the interference information at all positions of the question-answer statement to get the final comprehensive representation for each distractor. For $\widehat{\mathbf{S}}_i$,

$$\tilde{\mathbf{t}}_i^c = \texttt{MeanPooling}(\widehat{\mathbf{S}}_i^c), \hat{\mathbf{t}}_i^c = \texttt{MaxPooling}(\widehat{\mathbf{S}}_i^c) \tag{20}$$

Then we aggregate the interference information collected from all distractors together.

$$\mathbf{t}^c = ReLU([\{\tilde{\mathbf{t}}_i^c\}_{i=1}^{i=4}; \{\hat{\mathbf{t}}_i^c\}_{i=1}^{i=4}]\mathbf{W}_1^c + \mathbf{b}_1^c) \tag{21}$$

where $\mathbf{W}_1^c \in \mathbb{R}^{8d\times h}$ and $\mathbf{b}_1^c \in \mathbb{R}^h$ are the parameters to learn.

Finally, the interference information is input into a feed-forward network to predict a scalar P_c, which is the confusion difficulty.

$$P_c = Sigmoid(\mathbf{t}^c \mathbf{W}_2^c + b_2^c) \tag{22}$$

where $\mathbf{W}_2^c \in \mathbb{R}^{h\times 1}$ and $b_2^c \in \mathbb{R}$ are trainable parameters.

Prediction Layer. Intuitively, different types of questions have different focuses and investigate the different abilities of students. The "*knowledge*" type of questions (like the first example in Figure 1) require that students remember accurately the knowledge assessed by the questions. Hence, the distractors of this type of questions may be more confused, which will undoubtedly increase the confusion difficulty. The "*inference*" type of questions (like the second example in Figure 1) examine more knowledge points than the "*knowledge*" type of questions. For example, the first question in Figure 1 only examines the knowledge about the diagnosis of heart failure, but the knowledge points assessed by the second question contain hypertension, acute heart failure, antihypertensive drugs, diuretics and so on. Hence, it is hard for students to recall all knowledge relevant to the "*inference*" type of questions.

Based on this intuition, we take the weighted sum of the confusion difficulty P_c and the recall difficulty P_r as the overall difficulty $\widetilde{P}$ and use the semantic representation of the question to calculate the weights. To get the comprehensive semantic representation of the question, we conduct row-wise pooling operations and concatenation operation. Then we input the question representation into a fully connected layer to predict the weight of the confusion difficulty. Finally, the question difficulty is predicted as:

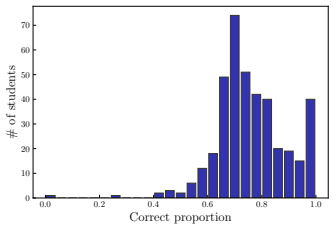


Figure 5: The distribution of the correct proportions of students. The correct proportion of a student is the division between the number of questions she/he answered correctly and the number of questions she/he answered.

$$\widetilde{\mathbf{q}} = \texttt{MeanPooling}(\mathbf{S}_a), \hat{\mathbf{q}} = \texttt{MaxPooling}(\mathbf{S}_a) \tag{23}$$

$$\gamma = Sigmoid([\widetilde{\mathbf{q}}; \hat{\mathbf{q}}]\mathbf{W}^\gamma + b^\gamma) \tag{24}$$

$$\widetilde{P} = \gamma P_c + (1-\gamma)P_r \tag{25}$$

Training. The question difficulty cannot be directly observed. Hence, we obtain the real difficulty of each question from the test logs, followed the previous works[5, 6]. Figure 6 shows a toy example of test logs. There are two ways to obtain the real difficulty.

(1) Proportion incorrect

A simple approach is to calculate the proportion of incorrect answers by dividing the number of students who have answered the question incorrectly by the number of students who have responded to the question[5, 6, 21]. The real difficulty of $Q_i \in \mathbb{Q}$ can be computed as follows:

$$P_i = g_i / G_i \tag{26}$$

where g_i represents the number of students who have answered question Q_i incorrectly, and G_i represents the number of students who have responded to question Q_i. For example, in Figure 6, the real difficulty of the question Q_1 is $P_1 = 1/3 = 0.333$.

(2) Rasch model

If all of the students have answered all of the questions, the first way could estimate the difficulty accurately. However, not all of the students answered all of the questions in our test logs. Meanwhile, as shown in Figure 5, the abilities of the students were also different.

The existence of these two facts will result in that the first way has some bias in the estimation process. For example, as shown in Figure 6, the difficulties, estimated by the first way, of Q_3 and Q_4 are the same. However, student U_3 who answered Q_3 incorrectly performs better than student U_4 who answered Q_4 incorrectly, which means that Q_3 may be more difficult than Q_4.

The Rasch model[18, 21] in *Item Response Theory* can be used to estimate the question difficulty with the consideration of the student ability. The Rasch model describes the probability of answering a question correctly by a logistic-like function.

$$\pi_{ij} = Probability(Y_{ij} = 1) = \frac{exp(\beta_j - P_i)}{1 + exp(\beta_j - P_i)} \tag{27}$$

where π_{ij} denotes the probability that student U_j will answer question Q_i correctly, β_j denotes the ability level of student U_j, and Y_{ij} denotes the score of student U_j on question Q_i (1 denotes U_j answer

StudentID	QuestionID	Student Answer	Correct Answer	Score
U_1	Q_1	A	B	0
U_2	Q_1	B	B	1
U_3	Q_1	B	B	1
U_4	Q_2	A	C	0
U_3	Q_3	D	C	0
U_4	Q_4	A	D	0
...	...	...	...	...

Figure 6: A toy example of test logs.

Q_i correctly, and 0 indicates U_j answer Q_i incorrectly). By fitting our test logs with the Rasch model, we can estimate the student ability β and the question difficulty P iteratively. After estimating, we use Min-Max normalization to scale the difficulty between 0 and 1, and the larger the difficulty is, the more difficult the question is.

We define the training loss (to be minimized) as the sum of the least square loss and a $l2$-regularization term.

$$L(\theta) = \sum_{i=1}^{i=|\mathbb{Q}|} (P_i - \widetilde{P}_i)^2 + \lambda ||\theta_{\mathcal{M}}|| \tag{28}$$

where P_i is the real difficulty, estimated by *proportion incorrect* or *the Rasch model*, of question Q_i, $\widetilde{P}_i$ is the predicted difficulty (see Eq.(25)), $\theta_{\mathcal{M}}$ denotes all trainable parameters in DAN, and λ is the regularization weight.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of DAN on a real-world dataset. First, we introduce our experiment setup. Then, we demonstrate that DAN outperforms all baseline methods. We further conduct the ablation analysis and qualitative analysis to provide further insight into how different difficulty components affect the integrated system.

4.1 Experiment Setup

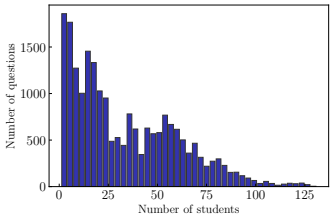
4.1.1 Dataset. The real-world dataset is collected from one of the largest medical online education websites in China. This dataset consists of more than 800,000 test logs (see Figure 6) and is collected from September 2017 to July 2018. After receiving the dataset, we preprocess it in two steps:

- *Drop duplicates*: In this data set, some students solve the same question multiple times. These students may repeat answer the same question to consolidate their knowledge. However, since we use the ratio of students who answered a question incorrectly to obtain the real difficulty, such duplicate attempts could lower the acquired difficulty. Hence, we drop the students' duplicate test logs and only reserve the first attempt. After dropping duplicates, a test log represents a unique <student, question> pair.
- *Filter*: If only a small count of students have tried to solve a question, the obtained difficulty of this question will have severe randomness. Hence, we filter the questions having no more than 10 test logs.

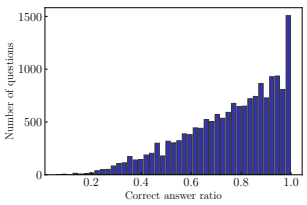
After pruning, we conduct the statistics on the dataset. Table 1 shows some statistics results. Figure 7(a) shows the distribution of the counts of students answered the questions. We can see that most questions are answered by more than 25 students. Figure 7(b) shows the distribution of correct answer ratios of questions. We can see that the correct answer ratios of questions are nearly a linear distribution. The high value at 1.0 suggests that there are some easy questions answered correctly by all students. Figure 7(c) shows the distribution of normalized correct answer ratios of questions. The correct answer ratios, normalized by the Rasch model, of questions are nearly a normal distribution.

Table 1: The statistics of the dataset.

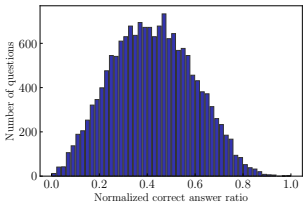
Statistics	Values
# of test logs	691,680
# of students	394
# of questions	16,342
Average test logs per question	43.325
Average test logs per student	1755.533



(a) The student count distribution.



(b) Number of questions w.r.t correct answer ratio.



(c) Number of questions w.r.t correct answer ratio normalized by the Rasch model.

Figure 7: Statistics of question set.

The unstructured medical materials (used by the document retrieval) consist of 2,130,128 published paper in the medical domain and 518 professional medical textbooks.

4.1.2 Model Details. Word embedding is pretrained using the whole unstructured medical materials with a vector dimension of 200 (i.e., $d = 200$), and the learned vector representations are shared across different components of the proposed method. All text materials used by DAN are truncated to no more than 100 words. Both each distractor and the correct answer have five relevant medical documents, i.e., $N = 5$. The Bi-LSTM in the encoding layer also have a dimension of 200. The parameters of the Bi-LSTM are shared between the processing of questions and documents. The hidden layer in the prediction layer has a dimension of 200, i.e., $h = 200$.

4.1.3 Training Settings. The model is trained with a mini-batch size of 32. We use Adam optimizer with a learning rate of 0.0005.

The dropout rate is set to 0.2 to reduce overfitting. The L2 weight decay λ is set to 0.001.

4.2 Baseline Approaches

Since DAN is an end-to-end difficulty prediction model, we only select a few end-to-end models as baselines with different considerations. Moreover, we also introduce two variants of DAN to highlight the effectiveness of each module of our framework. There are a total of five baseline approaches, including two variants of DAN.

SVM+TF-IDF: SVM is a commonly used machine learning method. TF-IDF is also a frequently used feature in many NLP tasks. We choose this method to help to understand the overall difficulty of this prediction task.

Bi-LSTM: Bi-LSTM is also a commonly used method in NLP. In order to help in understanding the overall difficulty of this task and demonstrate the effectiveness of two difficulty modules, we choose it as another baseline approach. We use a Bi-LSTM to encode the question-candidate answer statements to obtain the semantic representation and use a fully connected layer to predict the difficulty.

TACNN+Document Retrieval[2]: TACNN is a question difficulty prediction model for reading comprehension problems in standard English tests[6]. To the best of our knowledge, TACNN is the only data-driven end-to-end solution to question difficulty prediction task. Hence, we select this model to compare with our model. To apply it to our medical dataset, we need to make some changes as follows:

- Since the medical dataset does not contain the reading passages utilized by TACNN, we integrate the document retrieval module of DAN into it.
- The sentence length in the medical dataset is different from the sentence length in English test dataset. Hence, we need to change the pooling window sizes in CNN layer of TACNN to fit the lengths of the medical documents. Specifically, we set the pooling window sizes as (6, 6, 2, 4).

DANC: a framework which only has the confusion difficulty module. We choose this framework to highlight the effectiveness of confusion difficulty module.

DANR: a framework which only has the recall difficulty module. We choose this framework to highlight the effectiveness of recall difficulty module.

Both DAN and all baseline methods are implemented by PyTorch, and all experiments are run on a Tesla P40 GPU.

4.3 Performance Metrics

We evaluate all the models from both regression and rank perspectives. We omit the calculation details of all metrics for lack of space.

- *Root Mean Squared Error(RMSE) and Mean Absolute Error(MAE).* We use RMSE and MAE to measure the distance between the predicted difficulty and the real difficulty. Values closer to zero indicate better performances.
- *Spearman Rank Correlation Coefficient(SCC).* We use SCC to measure the correlation between predicted difficulties and real difficulties. The larger the SCC is, the better performance the model has.
- *Kendall Rank Correlation Coefficient(KCC).* Kendall Rank Correlation Coefficient[7] is also widely used in many regression problems in NLP[20]. We use it as another rank-based performance metric.

We perform 5-fold cross-validation and use paired two-tailed t-test to measure the statistical significance between DAN and the best baseline model ablating either DANR or DANC.

4.4 Performance Comparison

The experimental results of all the models, averaged over all five folds, are summarized in Table 2.

There are several observations: First, our proposed complete model, DAN, outperforms all baseline models significantly with all metrics. Note that not only are the difficulties predicted by DAN closer to real difficulties but also they have better agreement with real difficulties. Second, DANR and DANC outperform other baseline models, which indicates the effectiveness of the confusion difficulty module and recall difficulty module. Third, SVM and Bi-LSTM perform worse than other models. This observation suggests that the question difficulty prediction for multiple choice problems in medical exams is a challenging task, which is difficult to solve only with simple models and shallow semantic information.

In summary, all the above observations demonstrate that DAN has an excellent ability to predict the difficulties of multiple choice problems by incorporating the confusion difficulty module and the recall difficulty module.

4.5 Varying the Amount of Training Data

To investigate the learning curve of all models under two kinds of difficulty settings when varying the amount of training data, we evaluate all models by using the 25%, 50%, and 75% subsets from the training data of each fold. Note that the testing data in each fold remains unchanged. The experimental results of all models, averaged over all five folds, are summarized in Figure 8. There are two observations: First, under two kinds of real difficulty definitions, we can see that when the amount of training data is small(25%), the performance on RMSE of all models is similar. However, when increasing the amount of training data to 50%, DAN and its two variants outperform other baseline methods and lead them at 75% and 100% training data. This observation suggests that DAN has a stronger learning ability which can catch more semantic information. Second, when varying the amount of training data, DAN and its variants always have a better performance on two rank-based metrics. This observation indicates that the difficulties predicted by DAN have better agreement with the real difficulties.

4.6 Ablation Analysis

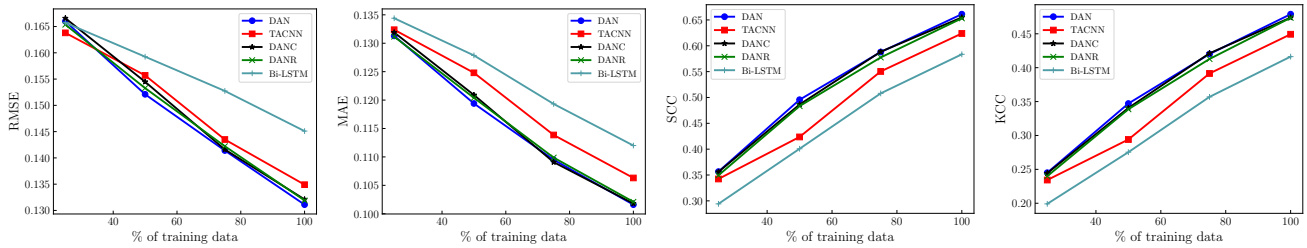
In order to highlight the individual contribution of each difficulty module, we run an ablation study. As shown in Table 2, we can observe a performance decrease by removing the confusion difficulty module or the recall difficulty module. Besides, we also can see that DANC performs better than DANR. This observation suggests that the confusion difficulty can reflect the question difficulty more accurately than the recall difficulty, which may be decided by the

[2]Because TACNN is not open source, we implemented our version based on PyTorch.

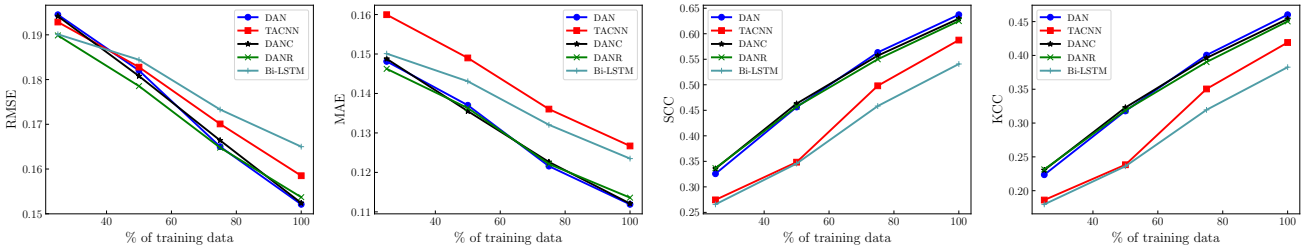
Table 2: The performance results

Models	Rasch model difficulty				Proportion incorrect difficulty			
	RMSE	MAE	SCC	KCC	RMSE	MAE	SCC	KCC
SVM	0.1716	0.1413	0.3026	0.2043	0.1913	0.1539	0.2926	0.1981
Bi-LSTM	0.1451	0.112	0.5835	0.4163	0.165	0.1235	0.5407	0.3826
TACNN	0.1349	0.1063	0.6238	0.4493	0.1585	0.1267	0.5876	0.4191
DANR	0.1319	0.1021	0.6531	0.4736	0.1537	0.1136	0.6249	0.4503
DANC	0.1321	0.1018	0.6548	0.4739	0.1524	0.1121	0.6295	0.4539
DAN	**0.1311***	**0.1016***	**0.6611****	**0.4789****	**0.1521****	**0.1119****	**0.6373****	**0.4602****

$^{**}p < 0.01$, $^{*}p < 0.05$



(a) The Rasch model difficulty



(b) The proportion incorrect difficulty

Figure 8: Varying the amount of training data. Since SVM performs far worse than other models, we do not illustrate its results.

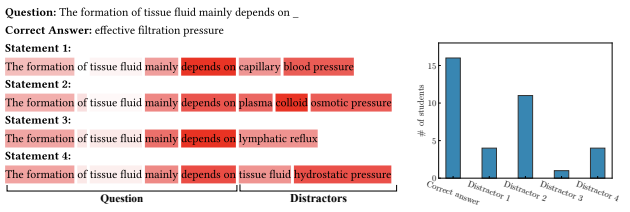
character of the multiple choice problem. The recall difficulty is caused by the rareness of the knowledge assessed by the question. However, in multiple choice problems, even though that the students are unfamiliar with the correct answer, they still can use the method of exclusion to select out the correct answer if they are familiar with four distractors. In other words, the comparison between the correct answer and the distractors can decrease the recall difficulty. Hence, it indicates the comparison structure in confusion difficulty module (as shown in Figure 4) can reflect the recall difficulty to some extent.

4.7 Qualitative Analysis

The design of confusion difficulty module and recall difficulty module enables convenient interpretation of the source of the question difficulty.

Intuitively, we think that the four distractors create the confusion difficulty of the question. Based on this intuition, in the confusion difficulty module, we use the co-attention strategy to extract the confusion information between the correct answer and each distractor (see Eq.(18)). The co-attention weights represent the semantic similarity between the correct answer and the distractor. The more similar a distractor and the correct answer are, the more contribution the distractor makes to the confusion difficulty. Figure 9(a) shows an example. We can see that in a statement the part most relevant to the correct answer is its distractor part. Moreover, the most relevant to the correct answer is the second statement. This observation means the second distractor makes the most significant contribution to the confusion difficulty. Figure 9(b) shows the count distribution of answers provided by students to the example question. We can see most students who answered this question incorrectly select the second distractor. This observation also means the second distractor is the most confusing candidate, which agrees with the analysis result from the attention heat map.

In the recall difficulty module, we capture the sentence-level similarity information between the question-answer statement and



(a) Attention heat map.

(b) The distribution of student answers.

Figure 9: Left: attention heat map from confusion difficulty module. A darker color indicates larger attention weight. In order to enable relevance comparisons among different statements, we apply a softmax over the attention weights of all words in all statements. Right: the distribution of student answers.

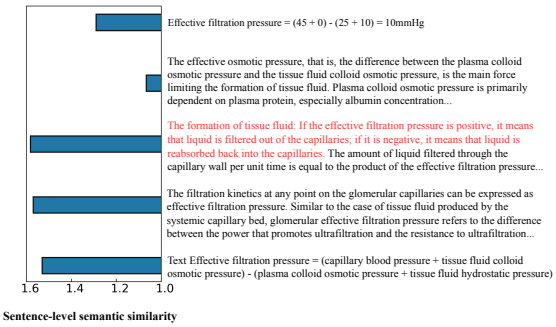


Figure 10: Sentence-level semantic similarity visualization.

the retrieved medical documents. Figure 10 shows the sentence-level similarities between the question-answer statement (shown in Figure 9) and its retrieved medical documents. We can see that the third document is most related to the question, indicating that it makes the most significant contribution to alleviating the recall difficulty of the question.

In summary, two visualization results hint that DAN provides a good way for the interpretation of the difficulty source of a question.

5 CONCLUSIONS

In this paper, we propose a novel Document enhanced Attention based neural Network(DAN) framework to automatically predict question difficulty for multiple choice problems in medical exams. For the proposed DAN, we design two methods based on different attention strategies to model two types of question difficulties, which are combined together to predict the overall difficulty. Experimental results on a real world dataset demonstrate the superiority of our proposed model and the effectiveness of two question difficulty components. Now, we only apply the DAN in medical exams, but it can be easily adapted to the multiple choice problems of other domain if we have the textbooks used by the retrieval module and the exam records used by obtaining the ground truth because the prediction model is agnostic to language and domain.

REFERENCES

[1] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2017. Neural natural language inference models enhanced with external knowledge. *arXiv preprint arXiv:1711.04289* (2017).

[2] Sanju Gajjar, Rashmi Sharma, Pradeep Kumar, and Manish Rana. 2014. Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine* 39, 1 (2014), 17.

[3] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine.* " O'Reilly Media, Inc.".

[4] Yu Hao, Xien Liu, Ji Wu, and Ping Lv. 2018. Exploiting Sentence Embedding for Medical Question Answering. In *AAAI*.

[5] Pedro Hontangas, Vicente Ponsoda, Julio Olea, and Steven L Wise. 2000. The choice of item difficulty in self-adapted testing. *European Journal of Psychological Assessment* 16, 1 (2000), 3.

[6] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests.. In *AAAI*. 1352–1359.

[7] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.

[8] Qi Liu, Zai Huang, Zhenya Huang, Chuanren Liu, Enhong Chen, Yu Su, and Guoping Hu. 2018. Finding Similar Exercises in Online Education Systems. In *KDD*.

[9] Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 3245–3253.

[10] Sarah KK Luger and Jeff Bowles. 2013. Two methods for measuring question difficulty and discrimination in incomplete crowdsourced data. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[11] Josiane Mothe and Ludovic Tanguy. 2005. Linguistic features to predict query difficulty. In *ACM Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop*. 7–10.

[12] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 130–136.

[13] Ulrike Padó. 2017. Question Difficulty–How to Estimate Without Norming, How to Use for Automated Grading. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. 1–10.

[14] Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. 2019. Option Comparison Network for Multiple-choice Reading Comprehension. *CoRR* abs/1903.03033 (2019). arXiv:1903.03033 http://arxiv.org/abs/1903.03033

[15] Chandrika Rao, HL Kishan Prasad, K Sajitha, Harish Permi, Jayaprakash Shetty, et al. 2016. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *International Journal of Educational and Psychological Researches* 2, 4 (2016), 201.

[16] Yuni Susanti, Hitoshi Nishikawa, Takenobu Tokunaga, and Obari Hiroyuki. 2016. Item difficulty analysis of english vocabulary questions.. In *CSEDU (1)*. 267–274.

[17] Min Tang, Jiaran Cai, and Hankz Hankui Zhuo. 2019. Multi-Matching Network for Multiple Choice Reading Comprehension. In *AAAI 2019*.

[18] Wim J van der Linden and Ronald K Hambleton. 2013. *Handbook of modern item response theory*. Springer Science & Business Media.

[19] Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A Co-Matching Model for Multi-choice Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

[20] William Yang Wang and Zhenhao Hua. 2014. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1155–1165.

[21] Kelly Wauters, Piet Desmet, and Wim Van Den Noortgate. 2012. Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education* 58, 4 (2012), 1183–1193.

[22] Ming Yan, Jiangnan Xia, Chen Wu, Bin Bi, Zhongzhou Zhao, Ji Zhang, Luo Si, Rui Wang, Wei Wang, and Haiqing Chen. 2019. A Deep Cascade Model for Multi-Document Reading Comprehension. *CoRR* abs/1811.11374 (2019).

[23] Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dual Co-Matching Network for Multi-choice Reading Comprehension. *CoRR* abs/1901.09381 (2019).

[24] Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[25] Wayne Xin Zhao, Wenhui Zhang, Yulan He, Xing Xie, and Ji-Rong Wen. 2018. Automatically Learning Topics and Difficulty Levels of Problems in Online Judge Systems. *ACM Trans. Inf. Syst.* 36 (2018), 27:1–27:33.