Towards Robust Automatic Question Generation For Learning

Towards Robust Automatic Question Generation For Learning

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen
op Maandag 8 April 2024 om 12.30 uur

door

Peide ZHU

Master of Engineering in Computer Science and Technology,
University of Science and Technology of China, China,
geboren te Jinan, China.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. ir. G.J.P.M Houben
promotor: Dr. C. Hauff

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. G.J.P.M Houben,	Delft University of Technology
Dr. C. Hauff,	Delft University of Technology
Onafhankelijke leden:	
Prof. dr. A. Hanjalic,	Delft University of Technology
Prof. dr. S. Verberne,	Leiden University
Prof. dr. A. Jatowt,	University of Innsbruck
Dr. J. Yang,	Delft University of Technology
Prof. dr. K.G. Langendoen,	Delft University of Technology, reservelid

SIKS Dissertation Series No. 2024-12
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords:	Automatic Question Generation, Information Retrieval
Printed by:	Print Service EDE
Cover:	LeLe, Peide Zhu
Style:	TU Delft House Style, with modifications by Moritz Beller `https://github.com/Inventitech/phd-thesis-template`

Wonder is the beginning of wisdom.

Socrates

Contents

Acknowledgments

Achieving this PhD degree is a long and challenging journey, made possible only through the love, guidance, help, and encouragement of many. I am extremely grateful to everyone who has played a part in it. First, I would like to deliver my gratitude and love to my wife, Shengnan, for raising our children and backing me up through every challenge. Your support, encouragement, and confidence in me have been my backbone.

I would like to express my utmost gratitude to my promotors, Geert-Jan Houben and Claudia Hauff. Thank you for entrusting me with this opportunity and providing the perfect research environment I can imagine to learn and research. Thank you, Claudia, for your patience, guidance, countless meetings, and invaluable feedback.

I want to particularly thank Jie Yang, Avishek Anand, and Zhiming Zhao for all the discussions and help. I also want to thank Prof. Xiang-yang Li and Hao Zhou, who gave me immense help at the beginning of my research.

My gratitude extends to my colleagues and friends: Alessandro, Marcus, Nava, Maria, Asterios, Cristoph, Rihan, Ujwal, David, Lixia, Agathe, Alex, Alisa, Andra, Andrea, Arthur, Christos, Daan, Danning, Daphne, Dimitrios, Esra, Felipe, Gabriel, Gaole, Garrett, Georgios, Petros, Kyriakos, Lijun, Lorenzo, Manuel, Na, Nadia, Nirmal, Sara, Sepideh, Sihang, Shabnam, Shahin, Tim, Wenbo, Yuandou, Zhen, Ziyu, and of course many others whose contributions have been invaluable.

Finally, I want to thank my father and my parents-in-law. Thank you for your love, help, and support. I want to thank my mother, Mrs. Jiu-xiu Du, a brave great woman who loved me to the last moment. You will always live in my heart.

Peide Zhu
Delft, 2024

1

Introduction

Questions are critical for information seeking and learning. For example, it is widely accepted in education that some form of questions (quizzes, tests, examinations) should be used in the classroom. There is a long history of research on this topic, consistently demonstrating the advantages of active learning, which includes the use of questions, over the passive consumption of learning material [1–3]. In addition to classroom learning, there is an increasing trend towards learning online. The recent decade has witnessed a significant surge in enrollment for massive open online courses (MOOCs). By 2022, online learning platforms like edX, Coursera, Udemy, and Khan Academy collectively offered over 100,000 courses from more than 950 universities worldwide [4]. These platforms have garnered considerable attention, attracting over 220 million learners, with Coursera accounting for 118 million of them, thereby significantly enhancing access to higher education [5, 6]. These systems usually provide learners with assessments, quizzes, and discussion forums for testing their understanding of the material and tracking learning progress. MOOCs require a large pool of questions due to the extensive range of courses and the ongoing need for assessment. They also require the frequent replacement of these assessment questions to maintain their validity, as their effectiveness can diminish after multiple rounds of use, primarily due to sharing among participants. Designing a large number of high-quality questions is a time-consuming and cognitively demanding task that requires trained instructors' expertise and experience [7]. Besides, as suggested by Bloom *et al.* [8], instructors need to follow the learning objectives and create questions assessing varying types of cognitive complexity. Automatic Question Generation (AQG) can ease the burden of instructors and provide students with more opportunities for assessments and self-assessments [9]. In this case, the course material, such as textbooks, course video subtitles, and the learning objectives, are taken as the inputs to the question generator targeting for generating questions satisfying specific learning goals.

Beyond the MOOCs, informal learning by searching and reading online resources via search engines like Google and Bing has become a meaningful learning and information-seeking method. This process often involves only reading the results passively. As aforementioned, actively engaging participants with questions in their learning process would be helpful. However, active learning is not part of modern-day search engines. One lim-

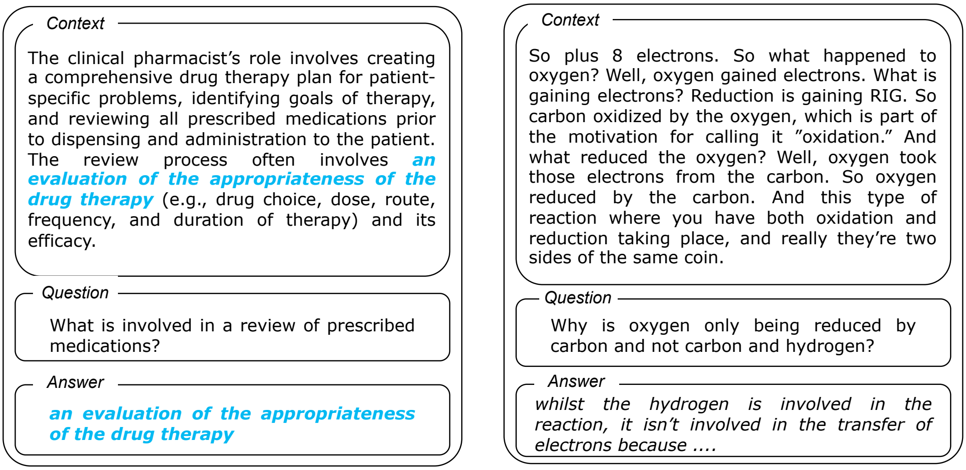


Figure 1.1: Examples of questions used for learning. The left example is taken from the SQuAD dataset where the context is passages from Wikipedia articles, and the question is close-ended regarding word spans extracted from the context (used in Chapter 2). The right example is taken from a MOOC discussion forum where the context is context course materials, and the question is open-ended regarding long-form answers (used in Chapter 3).

iting factor is that it requires trained experts to create good questions. Given the content available on the Web, the response time requirements of these platforms, and the ad-hoc information needs of users, the manually curated question banks can only cover a small part of the web corpus, and they need help to update. Therefore, creating high-quality, meaningful questions in a scalable way has become a critical challenge. Automatic Question Generation from text has become a promising solution [10–18].

Given these factors that motivate Automatic Question Generation (AQG) for learning, AQG has attracted considerable attention in the research community, and many aspects of AQG have been explored for developing robust, generalizable, and effective AQG systems. In this thesis, we focus on generating questions given a text passage. In this setting, the input to the AQG system, or the **question generator**, is the text **context** i.e., the grounded content for creating the questions, e.g., the textbook passage from the textbook and the transcripts from MOOC course video clips. Besides the text context, another optional input for the AQG systems is the **question requirements** i.e., the requirements from the users of the AQG systems, such as the instructors, regarding the types and target answers of the generated questions. In Figure 1.1, we present two types of QA examples to demonstrate the inputs, such as the text context and the question requirements, including the target answer and the question types to the question generator.

In recent years, research on AQG in this setting has witnessed significant progress, primarily attributed to the advancements in deep learning techniques, particularly pre-trained language models [19]. Despite the impressive progress, these methods can be brittle due to the diverse and complex properties of the input. The results are not entirely reliable in practice, which makes question quality evaluation critical. Designing reliable question quality evaluation metrics is also important to assess and train AQG systems and remains a challenging task in great need. Further, these systems are data-hungry and

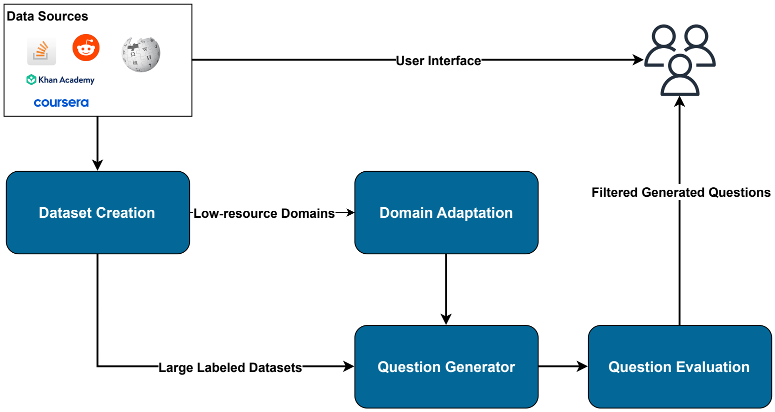


Figure 1.2: An illustration of the research scope of this thesis.

necessitate vast amounts of labeled data for training, highlighting the persistent challenge of data scarcity. To address this challenge, there has been a surge in the creation of labeled datasets facilitated by crowdsourcing and extracting question-and-answer data available on online platforms [20]. However, these data collection methods often entail significant labeling costs or result in datasets of suboptimal quality. Moreover, while the deep learning methods have made significant strides in AQG, their performance can exhibit performance inconsistencies when applied to domain-specific data that diverge in theme or genre from their training datasets.

As illustrated in Figure 1.2, we study the question generation for formal or informal learning, where the data sources range from MOOC platforms like Khan Academy and Coursera to Wikipedia and online discussion forums like Stackoverflow and Reddit. We study the dataset creation from diverse sources, the question generation and evaluation based on the collected data, and the adaptation of AQG models for different domains. Finally, we study the filtered automatically generated questions' effects on users' learning when applying them on the platforms.

Specifically, considering the challenges mentioned above, in this thesis, we investigate the following:

- the method and the automatic evaluation metrics of generating questions given a text passage (Chapter 2);
- the efficient question answering dataset creation via collaboration between crowdsourcing and deep neural models (Chapter 3&4);
- an unsupervised domain adaptation method for improving out-of-domain generalization of automatic question generation models (Chapter 5);
- the impacts and effectiveness of automatically generated questions on learning in the search-as-learning setting (Chapter 6).

Finally, in Chapter 7, we recap this thesis and discuss future research directions. In the following sections, we will introduce the research background and their limitations, including the question generation approaches, the related research on dataset creation and domain adaptation for question generation, and the questions' effects on human learning. Then, we will delve into several research questions, the corresponding research methodologies, and the contributions used in the thesis.

1.1 Question Generation Approaches

Prior question generation methods can be broadly classified into two high-level categories: the rule-based approaches and the sequence-to-sequence (Seq2Seq) generative neural models according to the generation approaches they employ.

1.1.1 Rule-based approaches

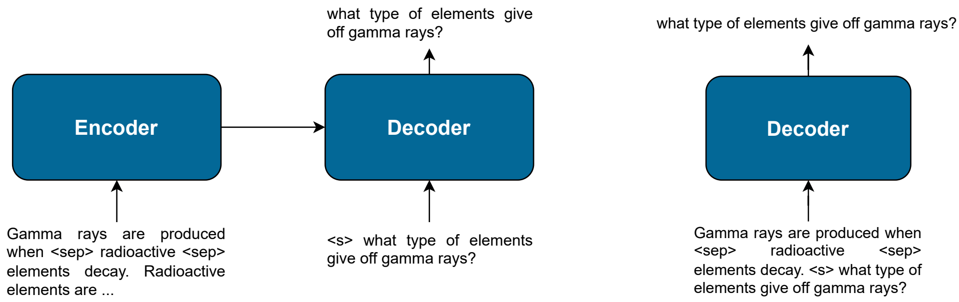
The rule-based approaches [10–12, 21–23] rely on well-designed, manually created templates and heuristic linguistic and semantic rules for question generation. Rule-based approaches are efficient and retain interpretability. Although the generated questions may fall short in quality compared with other methods, they are especially effective for question generation in low-resource domains where training/fine-tuning on these domains is impractical [24, 25]. The rule-based methods rely on high-quality rules and fall short in quality compared to human-curated questions. Therefore, recent research like Dhole and Manning [26] combines rule-based methods and neural models for AQG.

1.1.2 Sequence-to-sequence (Seq2Seq) Neural Models

With the advance in deep learning, various neural network models have been proposed for question generation [13–17, 27–31]. These models formulate the question generation task as a sequence-to-sequence (Seq2Seq) neural learning problem which takes the context passage as **input** and the questions as the **output**. The model architectures can be categorized into two types based on their architecture, including encoder-decoder models and decoder-only models, as shown in Figure 1.3.

Encoder-Decoder Models

In the models with the encoder-decoder architecture, the encoder is designed to encode the input text token sequence to hidden states while the decoder takes the hidden states as inputs and generates outputs autoregressively. A broad range of research that adopts this architecture has explored different types of encoders, decoders, and attention mechanisms, such as LSTM [32] and transformers based on multi-head self-attention [33]. Furthermore, as pre-trained language models (PLMs) have advanced the state-of-the-art across various natural language processing tasks [34], they have been introduced for AQG. For example, UniLM [18, 35], BART [36], and Prophetnet [37] are pre-trained encoder-decoder models for text generation and show the ability to generate high-quality questions after being fine-tuned for AQG; the encoder-decoder model T5 [38] is the backbone model for many approaches that create queries from documents such as docT5query [39] and InPars [40]. The encoder-decoder models are flexible to incorporate the bi-directional encoder that processes each token in the context both from the past and the future in order to capture intricate relationships and dependencies in the data effectively.



(a) (b)

Figure 1.3: An illustration of transformers-based question generation model in **(a)** encoder-decoder structure: the input to the encoder is the context passage, and the input to decoder is the target question started with the start-of-sentence <s> token and shifted right. **(b)** The Decoder-only generative language models: the input to the decoder is the context, and we use the special <s> token to indicate the start of the target question shifted right, and we only show the next token prediction outputs and ignore auxiliary prediction heads for clarity.

Decoder-only Generative Models

Unlike models with the encoder-decoder architecture, the decoder-only generative models utilize the generative decoder component of the traditional encoder-decoder architecture without the explicit encoder. GPT (Generative Pre-trained Transformers) [41] uses a multi-layer transformers decoder for the language model and is pre-trained on plain-text data to predict the following tokens with leftward tokens based on the left-to-right language model. It first demonstrates the effectiveness of this approach, and the successors GPT-2 [42] and GPT-3 [43] further show substantial improvements in a range of tasks. A lot of decoder-only models have been proposed [44]. These models have demonstrated remarkable generative capabilities by focusing solely on the decoder and leveraging large-scale pre-training, setting new standards for a wide range of language tasks.

Both the encoder-decoder and the decoder-only models for question generation are primarily based on autoregressive language generation. They are usually trained by the Maximum Likelihood Estimation (MLE) objective, which does not always align with the metrics to evaluate the generations' quality. Therefore, some research [45–47] argues that it is necessary to optimize directly for the eventual evaluation metrics via reinforcement learnings (RLs) in addition to the MLE loss. In this way, the generated questions are evaluated by external evaluation functions and use the measures as rewards to train the AQG model. In [48], we propose to evaluate various evaluation methods' impacts on the quality of the trained AQG model. PLMs in both architectures have achieved impressive results after fine-tuning for AQG or in a zero-shot or few-shot manner. However, their performance is not robust, and they suffer issues like language degeneration [47, 49], which results in both the data and the inherent structure of the models. To address the unreliability and the lack of robustness issue, it is necessary to conduct research on both data aspects, such as the creation and denoising of datasets, and the model aspects, such as domain adaptation and evaluation.

1

(a) (b)

Figure 1.4: An illustration of common QA dataset creation methods. **(a)**: Dataset creation by collecting user-generated questions and answers from online platforms. The collected data are further processed by fusion and validation with other online knowledge sources, such as Wikipedia, to formulate the final dataset. **(b)**: A common work pipeline of dataset creation by crowdsourcing where the crowd workers are tasked to follow the requirements and curate questions and answers, and their annotations would be further aggregated and validated to formulate the final dataset.

1.2 Dataset Creation and Domain Adaptation for Question Generation

1.2.1 Dataset Creation

The rapid advance in AQG has been driven by the aforementioned deep learning models and the access to massive labelled datasets that support training and evaluating these data-hungry models. AQG can be viewed as a dual task for Question Answering (QA) and shares the datasets initially created for QA. Questions can be categorized into two types based on their intent: i.e., for extracting specific knowledge, particularly the *exact spans* from the given context, as shown in the left example in Figure 1.1, or for seeking information [20], as shown in the right example in Figure 1.1. Figure 1.4 shows the two dataset creation pipelines. Datasets that contain probing questions are created mainly by crowdsourcing, where crowd workers already know the context passage and the target answer and then create these questions [50, 51]. Datasets that contain information-seeking questions are primarily created by crawling web platforms such as Amazon [52], Reddit [53], and MOOCs [54] where the questions are created by users on these platforms. These methods have been proven effective and efficient, and many benchmark datasets like SQuAD [50], HotpotQA [55], Natural Questions [56], and MS MARCO [57] have been widely used. However, given the inherent complexity associated with question generation and the extraction of relevant context, these methodologies often grapple with a trade-off between the financial implications of dataset creation and the ensuing data quality. As a result, the derived datasets either incur significant labeling costs or include incorrect labels. Therefore, in this thesis, we investigate data creation via both methods.

1.2.2 Domain Adaptation

Languages used in different applications often show unique linguistic characteristics, e.g. in-domain vocabularies, formal or in-formal style [58, 59], as we can observe in the two examples shown in Figure 1.1 where the left one is from the Wikipedia article on *Pharmacy* and the right example is collected from a MOOC in KhanAcademy platform. Question generation by fine-tuning PLMs heavily relies on the quantity and quality of available training data. However, data sources that contain well-formed questions are insufficient, especially in the educational domain, where much expertise is required to create questions geared toward human learning. To mitigate the lack of labeled training data, one solution is to pre-train models for AQG on a data-abundant labeled domain (source domain) and transfer the learned knowledge to the unlabeled target domain, which is known as *unsupervised domain adaptation* (UDA) [60]. It is a common challenge in machine learning research to learn knowledge in one domain and apply it in other domains with good generalization performance. One obstacle is the *domain shift* [61] between the source domain and the target domain, which violates the assumption that the training set and the test set are independent and identically distributed (i.i.d.). This, in turn, limits the model's generalization and portability.

1.2.3 Question Quality Evaluation

AQG plays an essential role in a wide range of critical tasks. It is critical to evaluate the quality of generated questions before applying them in real applications because the generated questions can drift away from the corresponding context semantics and the target answer or suffer from low grammar quality [45, 62, 63]. Human judgments are widely accepted as golden metrics to evaluate question quality. Generally, human evaluation in prior research scores the generated questions based on the following criteria [30, 63–65]: *Grammaticality, Relevance, and Answerability.* The grammaticality measures the grammatical fluency and the syntax structure coherence of the questions. The relevance measures the consistency between the generated question and both the context document and the target answer for factual questions. Different from other NLG tasks like translation and summarization, the evaluation of AQG should consider the answerability of the generated questions, which focuses on whether the question contains relevant information such as question words (Wh-types), necessary information like entities and relations, and whether the question is targeting on the correct answer span. However, human evaluation is applied on a small scale because it is expensive and time-consuming. Therefore, many automatic evaluation metrics have been proposed. n-grams-based metrics such as BLEU [66], ROUGE [67], and METEOR [68] are widely used for evaluating the generations' quality by comparing the lexical overlapping between the generations and the gold references because they are unsupervised. Furthermore, many metrics have been proposed to compare the similarity between the generations and the references using learned embeddings and PLMs for better evaluation than surface-level similarity [69–73]. Some recent research investigates using the learned models' ability to score the generated questions directly without references to mimic human judgments on question quality [74–79]. The model-based metrics rely on the learned models that depend on the data and methods used to train them. They may suffer from biases like spurious correlation and domain knowledge. Despite the numerous evaluation metrics proposed to evaluate one or some aspects of question quality, there is no comprehensive comparison, especially their effects on automatic question generation.

1.3 Questions' Effects on Human Learning

In both formal online learning and informal learning by online search, questions play a pivotal role, making it essential to investigate the quantity and qualitative study on the impacts of questions on learning. The effects of questions on human learning have been first studied in the classroom setting [1–3]. Since learning has increasingly transferred to digital media, some recent works [80–83] have begun to investigate questions' effects on learning through these media, although these studies tend to be on a small scale. In particular, the research [80] systematically analyzed the effectiveness of AQG on human learning compared to manually curated questions, as well as other impact factors such as learners' prior knowledge, the type of adjunct questions (factoid or synthesis), and the content that questions focused on. Similar to [80], Steuer *et al.* [83] studied automatically generated adjunct questions' effects on non-native speakers' English vocabulary learning. The effects were evaluated by the self-report of prior knowledge on the topic and the correctness of post-test questions. Van Campenhout *et al.* [84] used automatically generated questions in a university course as formative practice and evaluated the questions' effects by measuring the students' behaviour, such as engagement in the practice. However, the effects of automatically generated questions on human learning, including user behaviour and learning outcomes, are not well studied.

1.4 Main Research Questions

As mentioned above, despite the progress of AQG, it remains a challenging task to generate questions for long, unstructured documents and fairly evaluate the quality of generated questions. Deep neural models, when trained on limited labelled datasets, are susceptible to a range of issues that can compromise the quality and applicability of AQG. Deep neural models trained with limited labelled data suffer from *exposure bias*, and the generated questions can suffer from various types of errors, such as *the grammatical issue*, *semantic drift*, and *off-the-target answers*, which limits the application of AQG. This motivates our first main research question:

> **RQ1:** What metrics should be used for evaluating the quality of generated questions? How to compare the effectiveness of these metrics?

We then turn to dataset creation for the AQG task, considering both dataset creation by crawling online questions and by crowdsourcing. The two primary dataset creation methods suffer from noisy labels. For example, many user-created questions MOOC discussion forums do not contain specific references to the particular course content and limit the quality of question generation models trained on them [54]. The annotations created by crowd workers also often contain noisy annotations and disagreements among crowd workers. To address the issue of noisy annotations in crawled or crowdsourced data, we can extract the most relevant context related to questions, thereby enabling the training of more effective question-generation models. Further, we can develop models and strategies designed to handle labeling disagreements more efficiently and effectively. This motivates our second main research question:

> **RQ2:** How to use deep neural models to facilitate dataset creation for question generation and reduce the noise of created datasets?

There are various application domains of AQG. Although there are already abundant annotated questions in some domains, or it is easier to find competent crowd workers for creating questions, in some domains, there are no sufficient labeled data. Therefore, we further investigate the domain adaptation for AQG. Understanding and identifying the domain shift in datasets derived from different sources and devising strategies for mitigating its impacts is crucial for AQG. This leads to the third main research question:

> **RQ3:** How does domain shift in context affect AQG and how to improve AQG performance on out-of-distribution unlabeled target domains?

Finally, we investigate the effects of automatically generated questions on human learning, specifically in informal learning, by searching and reading online. Understanding the impacts of automatically generated questions on user learning behaviour and outcomes, characteristics of questions that can affect their impacts, and how the characters of users can impact their impacts are crucial for applying AQG in the platforms to facilitate users learning. This leads to our final research question:

> **RQ4:** How do automatically generated questions impact learners' behaviour and learning outcomes?

Guided by these research questions, we start by developing a question generation and evaluation pipeline. With this pipeline, we characterize the low-quality generated questions and provide an evaluation platform for evaluating the effects of automatic question quality evaluation metrics by comparing their effects on guiding the training of AQG model. Labeled datasets are essential for applying AQG for the target domains. Therefore, we explore the dataset creation for AQG with a preference on learning. We take a deep look at data collected from MOOC platforms and examine how well a deep neural ranking model can help improve label quality. We further explore how deep learning models can help aggregate crowdsourced disagreements in complex sequence annotations automatically. The understanding of automatic dataset denoising is crucial for training and evaluating AQG in broader application domains. Given the diversity in the datasets collected from different sources, we further investigate modeling the domain differences based on their deep contextual embeddings. With the knowledge of the differences in these domains, we research how to adapt AQG models trained on other domain data to the low-resource target domain. With the ever-improving performance of AQG systems, we take a user study to understand the generated questions' impacts on users' behaviour and learning outcomes, which is crucial for guiding the application of AQG systems.

1.5 Contributions

The main contribution of this thesis can be summarized as:

1. We propose three novel question quality evaluation metrics. These metrics provide a more accurate and multi-aspect evaluation of question quality beyond their similarity to gold references. We further provide a thorough empirical evaluation of the previously introduced metrics (Chapter 2).

2. We propose a novel dataset `MOOC-Clip` for MOOC discussion questions and evaluate the effectiveness of applying various neural ranking models for video clip recommendation for MOOC forum questions (Chapter 3).

3. We propose a novel aggregation approach for aggregating crowdsourced answer annotations for extractive question answering (EQA), which utilizes a simple yet effective method of collaboration between human and deep learning models for efficiently labeled dataset creation (Chapter 4).

4. We propose a two-stage unsupervised domain adaptation approach for AQG to make use of the labeled source domain data and abundant unlabeled data. In the first stage, The proposed two-stage approach mitigates noise and selects data close to the target application distributions with unsupervised domain clustering and data selection and achieves best adaptation performance (Chapter 5).

5. We conduct an empirical study on the automatically generated questions' effects on learners' behaviour and learning outcomes in the Search as Learning (SAL) setting (Chapter 6).

1.6 Thesis Origins

We now list the publications on which the research chapters were based.

Chapter 2 is based on the following paper:

- *Zhu, Peide, and Claudia Hauff. "Evaluating BERT-based Rewards for Question Generation with Reinforcement Learning." ICTIR'21.* [85]

Chapter 3 is based on the following paper:

- *Peide Zhu, Jie Yang, and Claudia Hauff. "MOOC-Rec: Instructional Video Clip Recommendation for MOOC Forum Questions." Poster@EDM'22.* [86]

Chapter 4 is based on the following paper:

- *Peide Zhu, Zhen Wang, Claudia Hauff, Jie Yang, and Avishek Anand. 2022. "Answer Quality Aware Aggregation for Extractive QA Crowdsourcing." Findings@EMNLP'22.* [87]

Chapter 5 is based on the following paper:

- *Peide Zhu, and Claudia Hauff. "Unsupervised Domain Adaptation for Question Generation with DomainData Selection and Self-training." Findings@NAACL'22.* [88]

Chapter 6 is based on the following paper:

- *Peide Zhu, Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff, "On the Effects of Automatically Generated Adjunct Questions for Search as Learning." To Appear@CHIIR'24*

2 Transformers-based Question Generation and Evaluation

In this chapter, we describe automatic question generation and evaluation methods. Automatic question generation systems aim to generate questions relevant to a given text, which can usually be answered by considering this text. Prior works have identified a range of shortcomings (including semantic drift and exposure bias) and thus have turned to the reinforcement learning paradigm to improve the effectiveness of question generation. As part of it, different evaluation metrics have been proposed to serve as rewards for the RL-based learning paradigm. Typically, these reward functions have been empirically investigated in different experimental settings (datasets, models, and parameters), but we lack a common framework to compare them. In this chapter, we first categorize existing rewards systematically. We then propose three new question evaluation metrics. Finally, we provide a fair empirical evaluation of different rewards in a common framework.

This chapter is based on the following conference paper: *Zhu, Peide, and Claudia Hauff. "Evaluating BERT-based Rewards for Question Generation with Reinforcement Learning." ICTIR'21.* [85]

2

2.1 Introduction

Automatic Question Generation (AQG) systems aim to generate natural language questions that are relevant to a given piece of text (the *context*) and can usually be answered by just considering the context. As an important natural language processing task, AQG can be used to improve question-answering [24, 45], conversational systems [89], and information retrieval (IR) [90, 91]. As a concrete example of the latter, AQG has been employed to improve the retrieval effectiveness of search systems by expanding documents with generated questions that the document might answer [39, 92]. The use of AQG has also recently been shown to be beneficial for learners in an interactive reading experiment [80], aiding learners' comprehension and learning. The natural next step is to employ question generation in the *search as learning* area [93], which consists of interactive reading, searching, and browsing activities [14–18].

The current state-of-the-art AQG systems are based on deep encoder-decoder neural networks, which take the context and the target *answer* as the input of the encoder and generate a question about the context (and the provided answer) with a decoder. Then, the models are trained with the objective of maximizing the log-likelihood of the ground-truth question paired with each input context. However, the evaluation of these models often involves not just the perplexity of the generated question but also other metrics like its relevance and fluency, adding complexity to the task.

Many datasets have been employed for AQG research, such as `SQuAD` [50, 94], `MS MARCO` [57] and `HotpotQA` [55]. In these datasets, only one ground-truth question is provided for each question-answer pair. However, for each context paragraph, there are usually several different facts related to the answer that questions can be generated about. In addition, even if there is only one fact contained in the answer, several syntactically very different questions may semantically be strongly related or even the same.

Based on these two observations, it is clear that the ground-truth questions provided in these datasets are not sufficient for high-quality question generation purposes. In fact, prior research has found that the likelihood-based training suffers from the problem of *exposure bias* [95], i.e., the model does not learn how to distribute probability mass over sequences that are valid but different from the ground truth. Because of exposure bias, many AQG models are not trained well enough to discover the relations between context and questions. In addition, AQG models trained in this manner can also suffer from the *semantic drift* problem, i.e., the models ask questions that are not relevant to the context and answer [45].

As a response to the training regime and dataset shortcomings, recently, the reinforcement learning (RL) paradigm has been taken up by the research community in order to optimize the AQG model during training with *rewards* that can directly evaluate question quality next to the available likelihood-based loss, so that questions with different forms from the ground-truth can be explored [45, 46, 62, 96, 97].

In the literature, a number of very different types of evaluation metrics have been proposed to evaluate question quality automatically and use them as rewards for the training process, such as the n-gram based metrics BLEU, Meteor and Rouge [46, 98, 99], the answerability metrics [27, 45], and fluency metrics [27, 100]. However, as Hosking and Riedel [101] reports, high scores in the metrics do not always equate to better questions when evaluated in a human evaluation setting. Undoubtedly, achieving a high score in a human

Table 2.1: Three examples of automatically *Generated* question, the *Context*, the ground-truth answer span, the question was generated for and the *Ground truth* question. The numeric columns represent the exact n-gram match metrics (BLEU-4), heuristic n-gram based metrics (Meteor), answerability (BERT-QA-loss), semantics-based similarity (QPP) and relevance based rewards (C-Rel, CA-Rel). These rewards are explained in more detail in Section 2.3.3. The scores range from 0 to 100.

Example 1		**B-4**	**Meteor**	**BERT-QA-loss**	**QPP**	**C-Rel**	**CA-Rel**
Context	At the end of World War I, the Rhineland was subject to the Treaty of Versailles.						
Ground truth	When was rhineland subject to the treaty of versailles ?						
Generated	The treaty of versailles was subject to the treaty of versailles?	53.32	77.82	0.02	1×10^{-5}	5×10^{-5}	0.87
Example 2							
Context	The clinical pharmacist's role involves creating a comprehensive drug therapy plan for patient-specific problems, identifying goals of therapy, and reviewing all prescribed medications prior to dispensing and administration to the patient. The review process often involves an evaluation of the appropriateness of the drug therapy (e.g., drug choice, dose, route, frequency, and duration of therapy) and its efficacy.						
Ground truth	What is involved in a review of prescribed medications?						
Generated	What does the review process often use?	0.00	14.31	100.00	99.94	99.98	99.91
	Example generated questions with issues (Section 4.2); assigned rating is shown in (brackets)						
Syntax issue (2)	what does the review process often involves ?	0.00	16.85	100.00	99.48	99.86	98.47
Non-answerable (0)	who does the review process involve ?	0.00	11.36	0.53	97.7	99.9	100.00
Relevance issue (1)	what is the dose of the drug ?	0.00	28.71	0.11	0.00	99.97	99.92
Example 3							
Context	This is the most common method of construction procurement and is well established and recognized. In this arrangement, the architect or engineer acts as the project coordinator.						
Ground truth	In the most common construction procurement, who acts as the project coordinator ?						
Generated	Who is the project coordinator?	14.16	36.60	100.00	97.83	99.98	99.98

evaluation is more important than an automatic evaluation metric.

The existing metrics are not sufficient for properly evaluating the quality of generated questions. First, n-gram-based metrics evaluate question quality by computing the exact match of n-grams in the generated and ground-truth questions. On the one hand, these metrics may give high scores for low-quality generated questions that repeat *n*-grams (such as shown in Example 1 in Table 2.1 where the term *versailles* appears twice in the generated question); on the other hand, since multiple questions are valid but only one ground-truth question is provided, these metrics can also fail to appropriately score question paraphrases and semantically equivalent questions (as shown in Examples 2 and 3 in Table 2.1). Second, there are several essential components involved in the generation and evaluation of a question: *the context*, *the answer*, and *the ground truth*. However, most of the proposed automatic metrics only consider one of them. For example, the n-gram based metrics compare the generated question with the ground-truth question; the answerability metric evaluates the possibility the question can be answered given the context, and the fluency metric computes the perplexity of the generated question. This chapter aims at further improving AQG with RL-based training with these metrics as the rewards. However, previously introduced metrics have been empirically investigated in different experimental settings (datasets, model, parameters), which does not enable us to compare their effectiveness directly. Therefore, we argue that further work is required to investigate how to *jointly* use these three components for question evaluation and evaluate the effectiveness of question evaluation metrics on serving as rewards for reinforced AQG training.

To sum up, we make two contributions: First, we categorize existing metrics and propose three novel question evaluation metrics and evaluate their effects of using them as rewards for reinforced AQG training; Second, we provide a thorough empirical evaluation of the previously introduced rewards employed inside a common base model. This in return allows us to compare the impact different rewards have on the model quality. Concretely, we use BERT [34] (because of its strong performance across a wide range of NLP tasks) as the base model to provide rewards for AQG. Overall, our main finding is that in such a fair comparison, the rewards that model *answerability* are the most effective, both in terms of an automatic evaluation as well as a human evaluation.

2.2 Background

In this section, we first discuss question generation applications and approaches, then turn to common evaluation metrics used to evaluate AQG approaches.

2.2.1 Question Generation

As an important natural language processing task, AQG has been applied to a wide range of applications. We here discuss three types of applications, including AQG for QA, conversational systems, and human learning, and then discuss the different types of existing automatic question generation approaches.

QG for QA

As the available information online and the requirement for quick access to information grows, question answering (QA) is playing an ever more important role. As a dual task of question-answering, AQG can be used to improve QA performance. Some works [14, 24, 45, 102, 103] take AQG as a generator to harvest question-answer pairs from passages and use this harvested data to pre-train QA models, which subsequently result in improved QA model effectiveness. AQG is also widely used in IR tasks, such as improving search system effectiveness by generating clarifying questions [91] or generating questions from e-commercial customers reviews [90].

QG for Conversational Systems
Conversational systems have become an important tool for information seeking. Asking good questions is significant for both providing user interaction and conversational QA training. Yao *et al.* [104] used AQG to create conversational characters. Wang *et al.* [89] and Ling *et al.* [105] proposed learning to ask questions in open-domain conversational systems with conversational context information. Gao *et al.* [106] and Gu *et al.* [107] proposed to use conversational question generation and conversation flow modeling as a means to generate synthetic conversations for training and evaluation purposes.

QG for Learning
Questions are a fundamental tool for a variety of educational purposes. Manual construction of good learning-oriented questions is a complex process that requires experience, resources and time. To reduce the expenses of manual construction of questions and satisfy the need for a continuous supply of new questions, AQG techniques are introduced. Kurdi *et al.* [9] provide a systematic review of AQG works for educational purposes. Besides, by conducting an interactive reading experiment and gaze tracking, Syed *et al.* [80] showed that the use of automatic AQG is indeed beneficial for learners as it aids learners' comprehension and learning.

QG Approaches
Past question generation research can be categorized as rule-based and neural network based on the generation approach employed. The rule-based approaches [10–12, 21–23] rely on well-designed manually created templates and heuristic linguistic and semantic rules for question generation. Labutov *et al.* [108] proposed a pipeline for question templates generation by crowdsourcing and ranking. Other works [24, 109, 110] proposed to generate factoid source question-answer triplets from passages, subtitles, or wiki knowledge graphs. Inspired by the advances in applying deep learning in natural language generation, various neural network models have been proposed for question generation [13–17, 27–29, 46, 46]. These models formulate the question generation task as a sequence-to-sequence (Seq2Seq) neural learning problem with different types of encoders, decoders and attention mechanisms.

2.2.2 RL-based Question Generation

To address the exposure bias and semantic drift problem, the reinforcement learning (RL) paradigm has been taken up by the research community in order to optimize the AQG model during training with metrics that can directly evaluate question quality next to the available likelihood-based loss so that questions with different forms from the ground-truth can be explored. To efficiently train AQG model with reinforcement learning, Rennie *et al.* [111] proposed the self-critical sequence training (SCST) algorithm that utilizes the test-time inference algorithm directly to obtain the output with the current model to normalize the rewards it experiences. In this way, SCST avoids estimating the reward signal and estimating normalization while at the same time harmonizing the model with respect to its test-time inference procedure. Because of its effectiveness, SCST is commonly used in follow-up RL-based AQG methods, while they use different evaluation metrics calculated with different methods and models as rewards [45, 46, 62, 96, 97]. We will discuss these metrics in the next section.

2.2.3 QG Evaluation Metrics

As a natural language text generation task, most previous AQG works use traditional metrics such as BLEU and Rouge to evaluate generated questions by comparing them with the ground-truth questions. However, Novikova *et al.* [112] and Nema and Khapra [65] pointed out that human ratings about question quality or answerability do not correlate well with these automatic evaluation metrics. Therefore, several different metrics have been proposed to evaluate different aspects of question quality, including fluency [27, 100], answerability [45, 65, 100], paraphrasing [45, 101], or discriminator-based relevance [100]. We broadly categorize question evaluation metrics used in prior works into lexical metrics and learned metrics, based on the underlying methods they use.

Lexical Metrics

Traditionally, lexical metrics like n-grams based metrics are widely used for evaluating the generations' quality by comparing the lexical overlapping between the generations and the gold references because they are unsupervised. Specifically, BLEU [66] is the most widely used metric in machine translation and AQG. It is a precision-based metric that measures the percentage of n-grams in the generations (here: the *generated question*) that overlap with references (here: the *ground-truth question*). ROUGE [67] is a recall-based metric that computes the percentage of n-grams in references that overlap with the generated questions, while Rouge$-L$ is a variant of Rouge-1, but uses the length of the longest common subsequence to compute the match rate. METEOR [68] extends BLEU and ROUGE by considering both precision and recall of overlapping n-grams by computing the harmonic mean. METEOR also extends exact n-gram matches to weighted matches of stemmed words, synonyms, and paraphrases.

Learned Metrics

Due to the ambiguity and diversity of natural languages, the lexical metrics may be misleading, especially for open-ended question generation. Therefore, many metrics have been proposed to mitigate this issue by comparing the similarity between the generations and the references using learned embeddings and PLMs for better evaluation. The learned word embeddings [113] or contextual embeddings [34] have been shown to provide better representations for capturing the lexical and semantic similarity. Various metrics have been proposed that use these learned embedding or neural models to optimize the correlation with human judgments, such as SMS [114] or BERTscore [71]. In addition to metrics that evaluate the lexical or semantic similarity of the generations and the ground truth, question quality evaluation requires a special focus on other dimensions, such as the answerability of the generations, and the relevance to both context and the ground truth. Many metrics have been proposed to fulfill these special requirements, such as Q-metrics [65], QPP, and QAP [45].

2.3 Methodology

We now present our methodology. In order to evaluate the different rewards, we designed a common framework that provides a fair testbed. This framework is visualized in Figure 2.1. It consists of two parts: the *QG model* and the *reward evaluator*. We see that beyond the reward computations (which are described in detail in this section), the remainder of the framework is the same, no matter the reward employed.

Generally, we use $\mathscr{C}$ and $\mathbf{A}$ to represent the context and answer span, respectively. Here, the context is composed of a sequence of words $\mathscr{C} = [w_i]_{i=1}^{M}$ with M being the size of the context. The answer span $\mathbf{A} = \{A^s, A^e\}$ indicates the start and end position of the answer in the context. Let $\hat{Q}$ represent the generated question, which is a sequence of predicted tokens $\hat{Q} = y_0, y_1, ..., y_N$. Then, the question generation task can be formalized as:

$$\hat{Q} = \arg\max_{Q} \Pr(Q|C, A) \tag{2.1}$$

We now describe our two framework components (QG model and reward evaluator) in turn before detailing the different rewards we implemented.

2.3.1 Question Generation Model

The AQG model uses the Seq2Seq framework with a maxout pointer mechanism and gated self-attention network similar to Zhao *et al.* [16] for paragraph-level question generation, as it is straightforward, and similar models have been widely employed in recent AQG research. To utilize the long-distance relation information at paragraph-level we add a multi-head attention mechanism in the encoder. We use the unsupervised pre-trained Glove [113] embeddings to initialize our word embeddings, as Glove embeddings have learned the substructure and statistical relation among words. In terms of word embeddings, besides word vectors, we also include word feature embeddings, including the part-of-speech (*POS*), named entity (*NE*), and answer tag. The answer tag vector is used to indicate whether a word is in the answer span. The *POS* and *NE* labels were extracted with Spacy[1].

2.3.2 Reward Evaluator

We use the self-critical sequence training (SCST) algorithm [111] for reinforcement learning training. SCST is an efficient reinforcement algorithm that directly utilizes the test-time inference output to normalize the rewards. In this setting, the evaluators are the *environment*, and the AQG model is the *agent* that interacts with it. The AQG model's parameters θ define a generation policy (i.e., the predicted token probability) $\mathbf{P}_\theta$ which makes the prediction of the next word, i.e., the *action*. After each action, the agent updates its state, i.e., updates hidden states, weights, etc., of the AQG model. Once the agent

[1] `https://spacy.io/usage/linguistic-features`

2

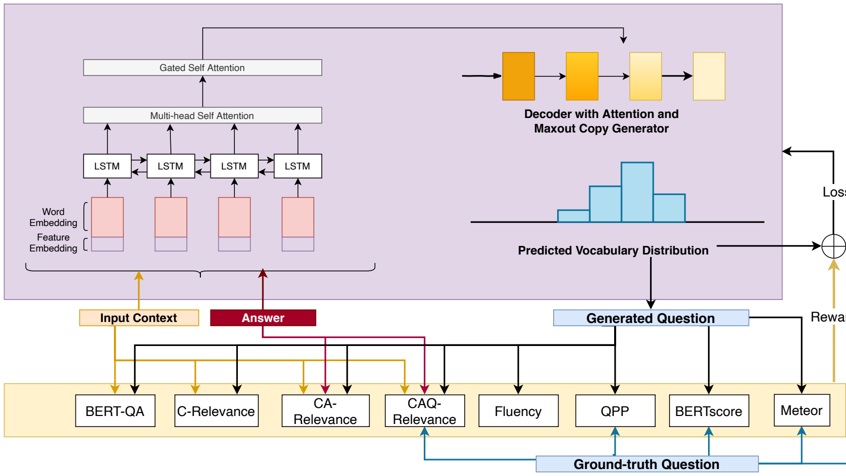


Figure 2.1: Architecture of our proposed question generation model.

finishes generating a sequence Q, it observes a *reward* $r(Q)$ computed with a given reward metric. Then, the RL loss function is defined as:

$$\mathbf{L}_{rl} = -\mathbf{E}_{Q^s \sim \mathbf{P}_\theta}(r(Q^s)) \tag{2.2}$$

where Q^s is the sampled output produced by multinomial sampling, that is, each word q_t^s is sampled according to the likelihood $\Pr(q_t|X, q_{<t})$ predicted by the generator. Because the sampling procedure is non-differentiable, the policy gradient $\nabla_\theta \mathbf{L}_{rl}$ is approximated using the baseline output Q^b obtained by greedy search, that is, by maximizing the output probability distribution at each decoding step. The loss function, when instantiated as just discussed, becomes thus:

$$\mathbf{L}_{rl} = (r(Q^b) - r(Q^s)) \sum_t \log \Pr(q_t^s|X, q_{<t}^s). \tag{2.3}$$

Using this reinforcement loss alone does not result in correctly learned word probabilities. For this reason, we follow the mixed objective approach [46], combining both cross-entropy loss (base model loss) and the RL loss:

$$\mathbf{L}_{mixed} = \lambda \mathbf{L}_{rl} + (1-\lambda)\mathbf{L}_{base}. \tag{2.4}$$

Here, λ is a mixing ratio to control the balancing between RL loss and the base model loss. In the following sections, we will explain the evaluation metrics used for rewards in detail.

2.3.3 Rewards

We categorized the metrics used for rewards from the literature into different types, as shown in Table 2.2. Importantly, in Table 2.2, we also provide insights into what information (context, answer, ground truth question, generated question) the evaluation metrics take as input. Naturally, all metrics used for rewards take the generated question into account. However, beyond that, there is little agreement as to what else to use. Based on the

Table 2.2: List of categorized reward functions employed in this chapter. Shown here are the inputs used to compute each reward. *GT* refers to the ground truth and *GQ* refers to the generated question. The novel rewards for AQG we propose in this chapter are labeled with ⋆.

Reward	BERT Task	Context	Answer	GT	GQ
Fluency category					
Fluency [100]	LM				☒
Similarity category					
⋆ **BERTscore**	LM			☒	☒
QPP [45]	Classifier			☒	☒
Answerability category					
BERT-QA-loss[45]	QA	☒	☒		☒
BERT-QA-geo[100]	QA	☒			☒
Relevance category					
C-Rel [100]	Classifier	☒			☒
⋆ **CA-Rel**	Classifier	☒	☒		☒
⋆ **CAQ-Rel**	Classifier	☒	☒	☒	☒

inputs to the evaluation metric and the downstream model type, we categorize the metrics into four types: (i) **fluency** indicates whether the generated question is a valid expression according to the language model; (ii) **similarity** indicates the similarity between the generated question and the ground-truth question; (iii) **answerability** indicates whether the generated question can be answered given the context; and (iv) **relevance** indicates how the generated question is relevant to the context, or the combination of the context, the answer and the ground truth.

The *BERT-Task* in Table 2.2 is the downstream task of BERT we use to compute the metric. For the Fluency metric and BERTscore, we use BERT as the language model to generate the contextual embeddings of the inputs. For metrics that are based on sequence classification, we add the BERT model transformer with a sequence classification head on top of the pooled output as the classifier. For metrics that rely on the QA task, we use the BERT model with a span classification head on top to predict the start and end positions of the answers. Lastly, we point out that we indicate in Table 2.2 also the three novel reward functions we contribute in this chapter: BERTscore, CA-Rel and CAQ-Rel.

We now discuss the different metrics that are used for rewards in the order of their appearance in Table 2.2.

Fluency Category

The perplexity of a sentence under a well-trained language model usually serves as a good indicator of its fluency [115]. We adopt the LM-based fluency metric as proposed by Xie *et al.* [100]. We first fine-tune the BERT language model with questions from the `SQuAD`

2

dataset. The fluency metric R_{flu} for question Q is calculated as follows:

$$R_{flu} = -\exp(-\frac{1}{|Q|}\sum_{i=1}^{|Q|}\log M_{flu}(Q_i|Q_{<i})) \tag{2.5}$$

We use the R_{flu} score as the reward.

Similarity Category

The n-gram based automatic evaluation metrics (BLEU, Meteor and Rouge) score the question similarity by computing the exact match of n-grams in the generated and ground-truth questions. As pointed out in Section 2.1, these metrics may yield a high score for low-quality generations that repeat n-grams in the generated question sequence. As there may be many valid questions with similar semantics, but only one ground truth question is provided, these metrics can also fail to appropriately score question paraphrases and semantically similar but syntactically very different questions. Therefore, we investigate two learned semantics-based question similarity metrics: *BERTscore* (the use as the rewards for AQG we propose) and *Question Paraphrasing Probability (QPP)*. These two metrics are based on BERT and compute the semantic similarity with high-level contextual representations instead of exact or heuristic *n*-gram matching.

BERTscore BERTscore [71] scores the similarity between the generated question (the *generation*) and the ground-truth question (the *reference*) by computing a similarity score for each token in the generation with each token in the reference. In contrast to n-gram-based metrics, BERTscore first represents contextualized token vectors with BERT and then uses greedy matching to maximize the matching similarity score, where each token is matched to the most similar token in the other sentence; subsequently, precision and recall are computed to yield the F1 measure. Given the generated question $\hat{Q}$ and the ground-truth question Q, the BERTscore can be computed as follows:

$$R_{BERT} = \frac{1}{|Q|}\sum_{y_i \in Q}\max_{\hat{y}_j \in \hat{Q}} \mathbf{y}_i^T \hat{\mathbf{y}}_j \tag{2.6}$$

$$P_{BERT} = \frac{1}{|\hat{Q}|}\sum_{\hat{y}_i \in \hat{Q}}\max_{y_j \in Q} \hat{\mathbf{y}}_i^T \mathbf{y}_j \tag{2.7}$$

$$F_{BERT} = 2\frac{P_{BERT}\dot{R}_{BERT}}{P_{BERT} + R_{BERT}.} \tag{2.8}$$

Here, y_i is the i^{th} token in the question sequence, and $\mathbf{y}_i$ is the pre-normalized contextual vector generated by BERT. We use the F_{BERT} score as our reward.

QPP Given one reference, n-grams based metrics sometimes fail to evaluate question paraphrases appropriately. Thus, inspired by the QPP metric proposed by Zhang and Bansal [45], we propose a BERT-based question paraphrasing classifier to provide paraphrasing probability as the reward. We pre-train this classifier model with the Quora

Question Pairs dataset[2]. As shown in Example 2 of Table 2.1, it scores question paraphrases more fairly: given the ground-truth question *What is involved in a review of prescribed medications?* and the generated question *What does the review process often use?*, we find BLEU-4 to assign these semantically similar questions a score of 0 while QPP assigns a score of 99.94. During the training of the AQG model, we use the QPP classifier to provide the probability of the generated questions and the ground-truth question being paraphrased as the reward.

Answerability Category

The answerability of a question measures the possibility that the generated question can be answered given the context. There are several reasons to consider answerability as a reward for AQG. First, for many AQG applications, such as generating questions for reading comprehension or question answering, it is a common requirement to ask questions that can be answered with the context information. Second, semantically drifted questions usually cannot be answered by the given context and answer, such as the *relevance issue* and the *non-answerable* question shown in Example2 in Table 2.1. Third, given the context, several valid questions are usually valid for the answer. Some contain information that is not used in the ground truth. The question similarity based metrics cannot evaluate this kind of novel generation fairly. Besides the ground-truth question, the answerability reward can take the context information into consideration. Therefore, we investigate two BERT-QA based answerability rewards. One is based on the QA loss *(BERT-QA-loss)*, and one is a heuristic reward based on the geometric average of the QA probability *(BERT-QA-geo)*. We use the BERT-QA model that is pre-trained on `SQuAD` to provide the QA probability, i.e., given the input context $\mathscr{C}$, the ground-truth answer $A = \{A^s, A^e\}$ and the generated question $\hat{Q}$, the question answering model outputs two probability distributions $P^s_{ans} = P(A^s|\mathscr{C}, \hat{Q})$ and $P^e_{ans} = P(A^e|\mathscr{C}, \hat{Q})$ over tokens in C, where $P^s_{ans}(i)/P^e_{ans}(i)$ is the probability that the i-th token is the start and end position of potential answer spans in the context.

BERT-QA-loss Given the ground-truth answer $A = \{A^s, A^e\}$, we evaluate the answerability by computing the cross-entropy loss of the QA predictions with the ground-truth answer:

$$loss(P^s_{ans}, P^e_{ans}, A) = CE(P^s_{ans}, A^s) + CE(P^e_{ans}, A^e) \tag{2.9}$$

$$R_{ans}(C, Q, A) = e^{-loss} \tag{2.10}$$

BERT-QA-geo As argued by Xie *et al.* [100], when the question is answerable, the model should be quite confident about the start/end span of the answer, so the distribution should peak for both P^s_{ans} and P^e_{ans}, i.e., the value of $max_i P^s_{ans}(i)$ and $max_j P^e_{ans}(j)$ are both large. Therefore, the geometric average of these start and end position probability distributions can be used as a heuristic answerability reward:

$$R_{ans}(C, Q) = \max_{1 \le i \le j \le T, j-i \le l} \sqrt{P^s_{ans}(i|C, Q) \cdot P^e_{ans}(j|C, Q)}. \tag{2.11}$$

[2]`https://www.kaggle.com/c/quora-question-pairs`

Here, l represents the maximum answer length.

Relevance Category

There are several essential components involved in the generation and evaluation of a question: *the context*, *the answer*, and *the ground truth*.

We investigate a series of binary classifier based discriminators to judge whether the generated question is relevant to the context *(C-Rel)*, the context and answer *(CA-Rel)*, and the context, answer, and reference*(CAQ-Rel)*. While the first reward (C-Rel) stems from prior work, we extended it and propose the just mentioned two novel rewards for AQG (which include more information than C-Rel in the input).

C-Rel This reward indicates whether a question is relevant to the context. We design a binary classifier based on BERT, inspired by Xie *et al.* [100]. It takes the context $\mathscr{C}$ and the generated question $\hat{Q}$ as inputs, and the output is the probability that $\hat{Q}$ is relevant to $\mathscr{C}$. To fine-tune the BERT classifier, we use the ground truth questions provided in the SQuAD dataset as the positive samples. We create negative samples in two ways: based on (i) question swapping and (ii) entity swapping. Negative sampling based on question swapping means randomly selecting ground-truth questions about a different context $\mathscr{C}$ as negative question samples for context C. In contrast, negative sampling based on entity swapping means replacing entities in the ground truth question with entities that do not occur in the context. We prefer to select entities that are of the same entity types, such as locations, dates, and names. Secondly, we create negative samples based on entity swapping by replacing entities in ground truth questions with entities from the same context though of different entity types.

CA-Rel We propose to use the probability that $\hat{Q}$ is relevant to the context $\mathscr{C}$ and the answer A pair as a reward. We design a BERT-based binary classifier that takes the context, the answer, and the generated question as inputs.

As there is only one ground-truth question for each context-answer pair, it is a challenge to create enough positive samples to train the classifier. We use three approaches to create positive samples: (i) back translation, (ii) information from a large paraphrase database, and (iii) a neural paraphrasing model. We now discuss each of these options in more detail. Paraphrases can be obtained by translating an English string into a foreign language and then back-translating it into English [116]. We select German as the pivot and use two pre-trained neural translation models: English-German and German-English, to generate question paraphrases. The PPDB [117] is a large-scale paraphrase database containing over a billion paraphrase pairs in 24 different languages. In our work, we employ bidirectionally entailing rules from PPDB, which replace single words or phrases with their paraphrases in PPDB. Finally, we train a seq2seq translation model with the Quora Question Pairs dataset and apply beam search to decode paraphrasing questions. Having created positive samples in these manners, we are left with creating negative samples for each question: we here employ the same manner as described for C-Rel.

CAQ-Rel Lastly, we propose a binary classifier that takes the context, answer, ground-truth Q_G and the generated question as input and outputs the probability that the generated question is relevant to the triplet $\{\mathscr{C}, A, Q_G\}$. We create the positive and negative samples in the same way as described for CA-Rel.

2.4 Experiments

We conduct our experiments on the SQuAD 1.0 [50] dataset which is widely used in AQG and QA research [13, 45, 91, 103]. It contains over 100K question-answer pairs generated by crowd-workers from 536 Wikipedia articles. The answers are selected word spans from Wikipedia article sentences. The dataset contains publicly accessible train and validation splits and a privately hosted test split. We split the public validation set into two parts: the development set and the test set. Thus, we have 87,598/5,285/5,285 samples for training, validation, and testing, respectively.

In the first step, we train all the proposed metrics with the huggingface's[3] uncased PyTorch BERT implementation. For the answerability metrics, we fine-tune BERT for the QA model with the SQuAD dataset. On the test set, the fine-tuned model obtains 80.28% exact match score and 87.89% F1 score. For the fluency metrics, we fine-tune the BERT language model with ground-truth questions in SQuAD and achieve 23.29 perplexity on the development set. For BERTscore, we use the available BERTscore implementation[4] provided by Zhang *et al.* [71]. This model does not require further fine-tuning.

We use the BERT model with a linear layer on top of the pooled output as the discriminator for the QPP metric and all three metrics in the relevance category. We train the model for all rewards with different datasets. For the QPP metric, we rely on the Quora Question Pairs dataset and spit the dataset as train/dev/test sets following the ratio of 70%, 15%, 15%, which expressed in numbers of samples amounts to 283K/60,643/60,643 respectively. For the C-Rel metric, based on the dataset creation strategy mentioned in Section 2.3.3, we harvest 297,980/17,322/17,954 samples for training, validation, and testing, respectively. For the CA-Rel metric, we harvest 1,137,052/68,649/68,703 samples as the training, development, and test set. Finally, for the CAQ-Rel metric, the size of the training, development, and test sets are 560,774/33,809/33,177. The performance of the trained models used as rewards is summarized in Table 2.3. Numbers are reported on each task's test set. In all cases, the accuracy reaches at least 90.98, indicating that our training regime yielded highly accurate models.

Before RL training with these metrics as rewards, we first train the basic AQG model by minimizing the cross-entropy loss and the copying loss. The encoder of the basic AQG model uses a two-layer bi-directional LSTM. The LSTM hidden cell size is 300. A dropout layer with a probability of 0.3 is applied between two bi-directional LSTM layers. We keep the 30K most frequent words in SQuAD as vocabulary. The word embedding size is 300. The decoder uses a 1-layer LSTM. We use SGD with momentum for optimization (momentum value is 0.8). The initial learning rate is 0.1 and decreases linearly after half of the training steps. We use beam search (beam size 10) for the decoding. We first train the basic AQG model for 16 iterations, and then we fine-tune the basic model with RL training, as described in Section 2.3.2. The mixing ratio (λ) in RL is set to 0.2. We use the basic AQG model as our baseline to compare the performance of all the rewards. To compare the

[3] https://huggingface.co/transformers/
[4] https://github.com/tiiiger/bert_score

Table 2.3: Performance of the fine-tuned BERT-based classification models that serve as the rewards.

Reward	Precision	Accuracy	Recall	F1
QPP	85.70	90.98	90.68	88.12
C-Rel	86.06	92.20	87.70	86.87
CA-Rel	93.27	92.62	92.99	93.13
CAQ-Rel	97.67	97.86	98.95	98.30

BERT-based rewards with n-gram-based metrics, we also train our AQG model with a Meteor-based reward. We choose Meteor as the representation of n-gram-based rewards as, based on our previous experience, Meteor usually outperforms other n-gram rewards.

2.4.1 Automatic Evaluation

We investigate the AQG models' performance along n-gram-based automatic evaluation metrics and the proposed rewards. The automatic metrics we use are BLEU, Meteor, and Rouge-L. They are based on the n-gram similarity between the generated questions and the ground truth, and are commonly used in text generation tasks. We calculate these metrics with the package released by Du *et al.* [13].

Table 2.4 summarizes our main results. We make the following key observations:

1. Training the AQG model with RL on every reward leads to better effectiveness with respect to the automatic metrics, except for the fluency reward on Meteor. **This result shows that it is effective to apply reinforcement learning on AQG model training in terms of automatic metrics.**

2. Optimizing one reward always leads to the improvement of the corresponding reward score. However, the improvement of each reward varies from each other, e.g., when optimizing the CAQ-Rel reward, the CAQ-Rel score improves by 5.02 compared to the baseline; however, optimizing the fluency reward only leads to a 0.02 improvement. This shows that the degrees of how rewards influence AQG training differ.

3. The rewards we use can be categorized into four types as already outlined in Section 2.3.3: fluency, answerability, similarity, and relevance. **We find optimizing one reward also leads to a score increase for other rewards of the same type.** This implies that rewards of the same type are correlated. We further investigate the correlation between them. The correlation matrix (expressed in Pearson correlation coefficient) is shown in Figure 2.2. We find the similarity-based rewards BERTscore and QPP are strongly correlated to each other, with a correlation coefficient 0.62. The relevance-based rewards are more related to the similarity rewards than each other. The BERT-QA-loss reward and the BERT-QA-geo reward are almost independent, which shows the heuristic reward BERT-QA-geo may not be a good indicator of whether a question can be answered by a QA model. This insight is helpful for designing an unsupervised QA system. The BERT-QA-loss reward and the fluency

Table 2.4: Performance evaluation along automatic metrics and rewards. The automatic metrics are BLEU-3 (B-3), BLEU-4 (B-4), Meteor (M) and Rouge-L (RGL).

Models	B-3	B-4	MT	RGL	QA-L	QA-G	QPP	BERTscore	C-Rel	CA-Rel	CAQ-Rel	Fluency
Baseline	23.98	18.44	21.79	45.95	**65.60**	72.23	26.90	67.62	85.23	89.59	51.19	-10.97
Meteor	25.56	19.84	22.80	47.23	65.52	72.81	28.84	68.36	86.34	90.70	52.54	-10.96
QA-loss	**25.88**	**20.13**	22.93	**47.51**	65.48	74.99	30.57	68.43	85.28	93.81	54.34	-10.95
QA-geo	24.82	19.23	22.23	46.49	65.52	**75.04**	28.06	67.81	86.51	90.39	51.46	-10.96
QPP	25.76	20.08	**22.99**	47.47	65.53	73.87	**31.68**	**68.53**	**88.93**	91.37	54.71	-11.02
BERTscore	24.84	19.27	22.25	46.89	65.59	72.33	28.00	68.18	85.29	90.38	52.97	-10.98
C-Rel	24.71	19.11	22.03	46.65	65.47	71.88	27.33	67.85	87.05	89.83	53.00	-11.00
CA-Rel	24.00	18.51	21.81	46.29	65.34	73.16	28.54	67.26	84.29	**94.55**	**56.22**	-11.04
CAQ-Rel	24.24	18.68	21.92	46.01	65.41	72.56	27.29	67.61	84.53	90.01	52.13	-10.97
fluency	24.19	18.67	21.77	46.11	65.59	71.86	26.47	67.59	84.95	89.61	51.75	**-10.95**

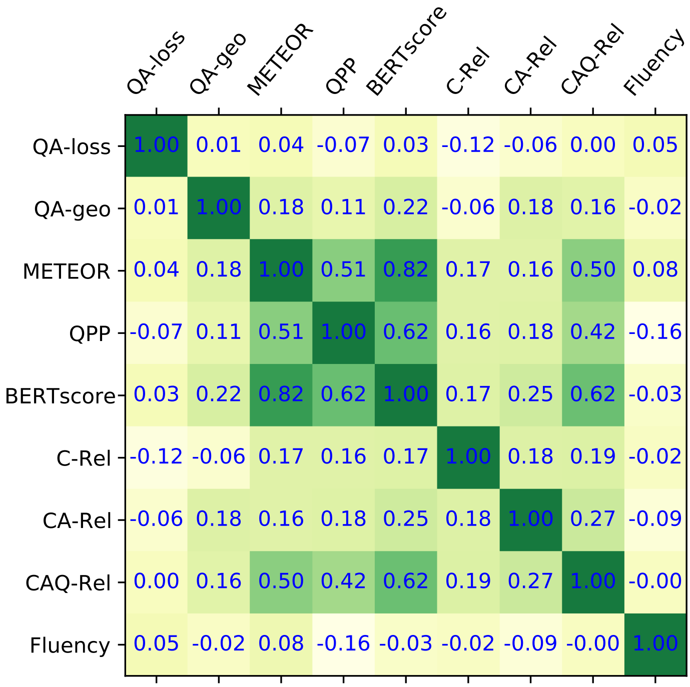


Figure 2.2: Pearson correlation coefficient matrix of the rewards.

reward are not correlated to other rewards, which shows the fluency and BERT-QA-loss rewards focus on different aspects of the generated questions' quality.

2.4.2 Human Evaluation

In addition to the automatic metrics, we further conduct a human evaluation of our test set to investigate whether optimizing the proposed rewards leads to the improvement in question quality by human standards.

To this end, we randomly sampled 100 testing documents, and three computer science students rated questions generated by 9 different models in a blind setup (i.e., they did not receive information on which question was generated by which model): the basic AQG model, and the models trained with our different reward functions. We also included the ground-truth question in the labeling process as a control setting, as we expect these questions to receive the highest scores in a human evaluation. In order to rate each sample, we provided the context, the ground truth answer span, and all the questions for each sample on one screen.

The rating was conducted along three criteria: the *Syntax* (on a scale of 1-3), the *Relevance* (on a scale of 1-3), and the *Answerability* (a boolean value). For syntax, score 1 means major syntax issues; score 2 means a small mistake (e.g., lacking an article or pronoun); and score 3 is correct. In the relevance category, score 1 means the question

Table 2.5: Human evaluation results. *GT* means *Ground Truth*. Shown in bold is the best measure for each of the three evaluation dimensions. The ground truth row is not included here.

Reward	Syntax (1/2/3)	Relevance (1/2/3)	Answerability (0/1)
GT	2.86	2.84	0.93
Baseline	2.49	2.32	0.67
Meteor	**2.55**	2.35	0.67
Fluency	2.47	2.18	0.63
QA-loss	2.50	**2.39**	**0.72**
QA-geo	2.41	2.20	0.66
BERTscore	2.50	2.24	0.69
QPP	2.36	2.31	0.68
C-Rel	2.39	2.19	0.61
CA-Rel	2.30	2.22	0.63
CAQ-Rel	2.40	2.22	0.63

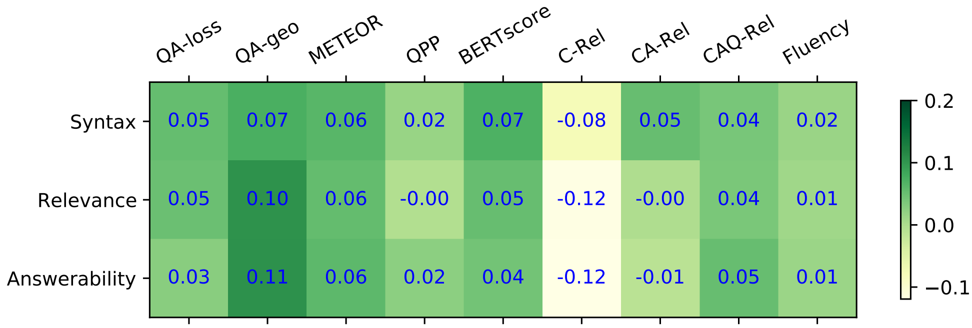


Figure 2.3: Pearson correlation coefficient matrix between reward scores and human ratings.

is not relevant to the context and the answer; score 2 means it is partially relevant (e.g., a question may be more general than what the answer is about); and score 3 means the question is relevant and relevant to the given answer. In terms of answerability, it needs to be rated whether the question can be answered with the context information and the provided answer. To provide the reader with intuition, we report three examples of generated questions with syntax/relevance/answerability issues in Table 2.1, **Example 2**. As all raters rated the same 100 samples, we considered their average rating for each dimension.

We report the human evaluation results in Table 2.5. In addition, in Figure 2.3 we present the correlation between the reward scores and the human ratings.

We make the following observations.

1. The baseline (i.e., no reward, just the likelihood loss) outperforms all relevance-

based rewards. Although optimizing on relevance-based rewards (C-Rel, CA-Rel, and CAQ-Rel) leads to improvement of the automatic rewards, it reduces the human rating with respect to syntax, relevance, and answerability.

2. We also add Meteor for the comparison of the performance of n-gram-based rewards. We find that optimizing on the Meteor rewards improves all of the three rating criteria. It achieves the best syntax score. As we show in Figure 2.2 for the automatic evaluation, Meteor is strongly correlated with BERTscore, QPP and CAQ-Rel rewards. This implies that Meteor can capture the lexical and semantics similarity in a way and can be used as a computation-efficient reward for AQG.

3. The BERT-QA based answerability reward BERT-QA-loss outperforms all other rewards regarding both Relevance and Answerability. This shows that the BERT-QA-loss metric is a good indicator that reflects the questions' relevance and answerability. This also shows that the AQG task is different from common text generation tasks like machine translation or summary generation; here, answerability is a critical criterion for question quality evaluation. Although BERT-QA-geo does not perform as well as BERT-QA-loss, as shown in Figure 2.3, BERT-QA-geo is most correlated to the human judgment on answerability, relevance, and syntax. As the BERT-QA-geo reward is a heuristic indicator for a question's answerability and it does not require answer information, this correlation between the BERT-QA-geo reward and the human judgments implies that it is possible to develop an indicator based on BERT-QA-geo for unsupervised or semi-supervised QA/QG training.

4. In general, the correlation between the human evaluation dimensions (syntax, relevance, answerability) and the reward scores is low: the linear correlation coefficient reaches 0.11 (between answerability and BERT-QA-geo) at best. One apparent reason is the different scoring system (binary or three levels for the human evaluation dimensions). At the same time, though, this lack of a high correlation between human ratings and reward scores shows that the reward functions we use are vastly different from the human rating dimensions.

2.5 Limitations

One limitation of this study is that we only focus on extractive close-domain question answering, which limits its application in the open-domain and targeting at long-form answering applications. Some of the evaluation functions designed in this research, such as the rewards in the answerability category, cannot apply on information-seeking questions. The more recent research, such as GPTScore [118] and G-Eval [119] that use LLMs such as ChatGPT and GPT-4 as evaluators, which may be better for evaluating quality of diverse types of questions. Another limitation is that we only one dataset is used for evaluating these rewards. Although it fits the requirement of a fair evaluation on the rewards, perform comparison along different extractive QA datasets constructed from different domains and applications would provide more insights on performance of different rewards.

2.6 Conclusions

In this chapter, we investigated the task of question generation. We systematically categorized the question evaluation metrics that past reinforcement learning proposed as rewards for training AQG model. We implemented all these rewards—as well as three we proposed ourselves—in a common framework to enable a fair evaluation. We performed both an automatic evaluation (with established metrics commonly employed for AQG evaluation) as well as a human evaluation, where human raters evaluated the generated questions along the dimensions of syntax, relevance, and answerability. We found that it is indeed effective to apply reinforcement learning on AQG model training in terms of the automatic metrics. Overall, the BERT-QA-loss and QPP rewards had the best effectiveness. Our human evaluation showed that BERT-QA-loss also achieves the highest relevance and answerability scores while using Meteor as the reward achieves the highest syntax rating.

3

MOOC-Rec: Instructional Video Clip Recommendation for MOOC Forum Questions

*In this chapter, we focus on the noisy data issue in the data collected from MOOC online discussion forums for addressing the information overload. To this end, we investigate the recommendation of one-minute-resolution video clips based on the textual similarity between the transcripts of clips and MOOC discussion forum entries. We first create a large-scale dataset from Khan Academy video transcripts and their forum discussions. We then investigate the effectiveness of applying pre-trained transformers-based neural retrieval models to rank video clips in response to a forum discussion. The retrieval models are trained with supervised learning and distant supervision to effectively leverage the unlabeled data—which accounts for more than 80% of all available data. Our experimental results demonstrate that the proposed method is effective for this task by outperforming a standard baseline by **0.208** on the absolute change in terms of precision.*

This chapter is based on the conference paper: *Peide Zhu, Jie Yang, Claudia Hauff. "MOOC-Rec: Instructional Video Clip Recommendation for MOOC Forum Questions". EDM'22.* [86]

3.1 Introduction

Massive Open Online Courses (MOOCs) provide open access to world-class courses for the public, which greatly improves the opportunities for online learning. The discussion forum is a major component of a MOOC as it is the primary communication tool among learners and instructors [120] to moderate the lack of physical access in MOOCs. It can help learners build a sense of belonging and learn from peers or help instructors monitor learner affect and academic progress [121]. However, since questions targeting the same video content are scattered among discussion threads, without supporting navigation facilities, learners cannot effectively retrieve valuable discussions for a particular piece of content. In addition, learners' posts seeking help may be drowned out by the many other competing posts, making it hard for learners to get attention from instructors and peers. The unstructured, unorganized forums with a large number of discussions (that can lead to information overload [122]) are hindering instructors and learners from benefiting from them, decrease community interaction, reduce responsiveness in forums, and in the end lead to low MOOC retention rates [123, 124].

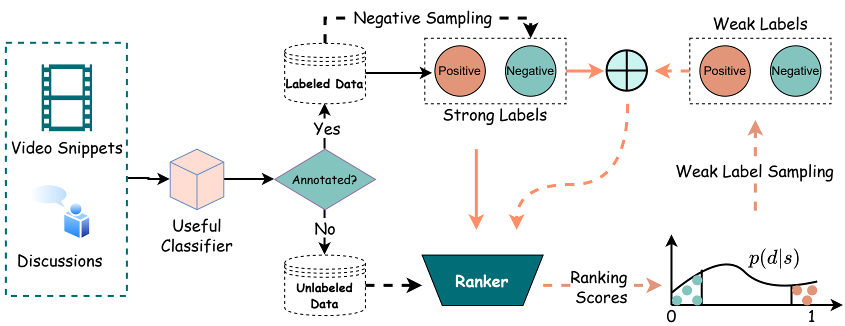


Figure 3.1: Overview of the MOOC-Rec system. The useful classifier first distinguishes the unuseful questions in the MOOC forum discussions. Then the discussions with annotated timestamps would be used for training the ranker as labeled data, and the unlabeled data would be used for distant supervision.

Existing works directed at addressing the information overload issue in MOOC forums have proposed more effective navigation tools to identify instructional video content and make recommendations of a ranked list of video clips. For example, Agrawal *et al.* [121] classify posts that need help and employ bag-of-words-based retrieval techniques to map those posts to minute-resolution course video clips. The clip recommendation algorithm is evaluated on posts from one course. Trirat *et al.* [125] built a recommender system to generate a ranked list of video clips giving a student's question with a deep neural network; they evaluate the system with 50 questions. Despite these attempts, we argue that prior works on video clip recommendation suffer from a lack of training data and, as a consequence, report evaluations only on small-scale data. It remains a challenge to develop and evaluate a system that can scale to thousands of MOOCs across different domains.

In our work, we first address the lack of training data issue by creating `MOOC-CLIP`, a

Table 3.1: Examples of MOOC video clip transcripts, labeled and unlabeled discussions from Khan Academy, and an overview of the creation of strong-labeled and weakly labeled items for BERT training.

Video Clips	**8:00 - 9:00**	Actually, let me write it like this. Let me move this part. So cut and paste. Let me move it over to the right a little bit because what I want to show is the gaining of the electrons. So plus 8 electrons. So what happened to oxygen? Well, oxygen gained electrons. What is gaining electrons? Reduction is gaining RIG
	9:00 - 10:00	So carbon oxidized by the oxygen, which is part of the motivation for calling it "oxidation." And what reduced the oxygen? Well, oxygen took those electrons from the carbon. So oxygen reduced by the carbon. And this type of reaction where you have both oxidation and reduction taking place, and really they're two sides of the same coin.
Labeled Discussion	**Q**	At **8:22**, why does oxygen gain 8electrons and not gain 2 electrons?
	A	one oxygen gains 2 electrons there were 4 oxygen's so it gained 4*2=8 electron's
Unlabeled Discussion	**Q**	Why is oxygen only being reduced by carbon and not carbon and hydrogen?
	A	whilst the hydrogen is involved in the reaction, it isn't involved in the transfer of electrons because its oxidation number remains the same throughout the reaction. if the hydrogen was reducing the oxygen as well, its oxidation number would also have to change (from 1+ to 0)
Content Irrelevant Questions		plz help me now. how new are these videos? dadhadhadhahdhadh nice smiley face.

novel large-scale dataset from Khan Academy [1], that includes video transcripts and forum posts (both questions and answers) using raw data available from `LearningQ` [54], an open source tool and dataset for educational question generation. We split a whole MOOC video into one-minute-resolution clips based on timestamps in video captions. We represent each video clip with its textural feature, i.e., the captions of each clip. We tread a thread of forum posts starting with one question and its following answers as one discussion. Table 3.1 shows examples of minute-resolution video captions and discussions. Second, we propose `MOOC-Rec`, a dense retrieval based instructional video clip recommendation system for MOOC forum questions. For each content-related thread, `MOOC-Rec` recommends a ranked list of video clips that are likely relevant and helpful for answering the question. Figure 3.1 demonstrates the overview of the system. Although dense retrievers have been applied in various retrieval tasks such as DPR [126] and ColBERT [127], it is unknown

[1] `https://www.khanacademy.org/`

whether they are an effective approach for MOOC video clip recommendation. Lastly, we point out that only 11.57% of all discussions in our dataset are labeled with a target video clip, which poses challenges for training MOOC-Rec with limited labeled data and abundant unlabeled resources.

We here first investigate the effectiveness of MOOC-Rec, and then we address the scarcity of labeled data by using distant supervision and in-batch negatives to train the ranker. The comprehensive experiments on our large-scale dataset, which consists of about 274K discussions, show that our systems significantly improve the clip recommendation performance by outperforming a standard baseline by 0.208 in terms of precision.

3.2 The MOOC-Clip Dataset

To address the lack of research data, we create a large-scale dataset using raw data crawled with LearningQ[2] from Khan Academy, a MOOC platform that allows learners to ask and answer questions about the learning materials during learning. We keep video transcripts, forum questions, and answers of MOOCs with both transcripts and discussions available.

Learners use discussion forums in different ways. Besides asking questions related to the course materials, they may also discuss irrelevant topics [128] for the purposes of socializing, spamming, or expressing their appreciation for the course materials. Some questions posted by learners also suffer from a lack of proper context or are too generic. Therefore, it is necessary to remove these relatively—for our purposes—low-quality questions. In line with LearningQ, we consider a user-generated question to be useful for learning when all of the following conditions hold: (i) the question is concept-relevant, i.e., it seeks for information on knowledge concepts taught in lecture videos; (ii) the question is context-complete, containing sufficient context information to enable other learners to answer the question; and (iii) the question is not generic. Besides labeled questions in LearningQ, we manually labeled the usefulness of 2K questions among topics that are not labeled in the useful question subset of LearningQ by adopting the same procedure. We also labeled 5K questions based on their lexical relevance to video transcripts (2.5K with the highest BM25 scores as useful, 2.5K with the lowest BM25 scores as negative) in order to exclude non-relevant questions. In total, there are 13,290 labeled questions over 8 topics. We found 60.9% of them to be *useful* and 39.1% of them to *not be useful.* We keep all items belonging to 3 topics (2,344 in total) as the unknown set for our cross-topics evaluation, 8,766 questions on the remaining five topics for training, and 2,186 questions as the known topic test set. We train a BERT-based text sequence classifier for useful question classification. Table 3.2 summarizes its performance.

During preprocessing, we first remove noisy discussions that contain only meaningless tokens, as well as videos that have no discussions. Then, we apply the useful question classifier on all items(522K) and retrain only items that are classified as useful. In the end, we retain 273,887 discussions from 7,349 videos of 6 topics. We use regular expressions to retrieve discussions where learners label posts with exact timestamps in questions or answers. We split the video transcripts into snippets with a one-minute length. The discussions and the snippets which cover the timestamp are labeled as *positive* items. The other discussions are treated as unlabeled. Table 3.2a and Figure 3.2a summarize the data

[2] https://github.com/AngusGLChen/LearningQ

Table 3.2: Useful question classifier results. The classifier is trained on three topics. Therefore, we report its performance on the same topics of training data, and the performance on other topics (Cross-Topics).

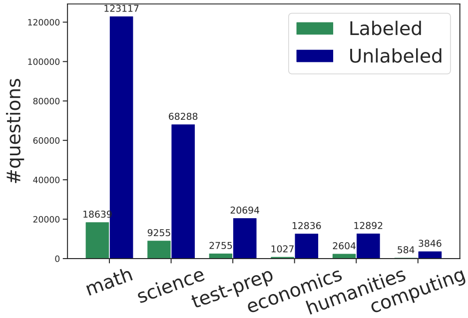
	Same Topics			Cross Topics		
Method	Acc	Rc	F1	Acc	Rc	F1
Q	89.40	96.68	92.90	77.20	74.49	75.82
Q+C	89.75	96.54	93.02	73.30	82.68	77.71

Table 3.3: Dataset overview, in terms of videos (#V), snippets (#S) per video, discussions (#D) per video, clip (#W), the number of words per question (Q) and the number of words per answer (A)

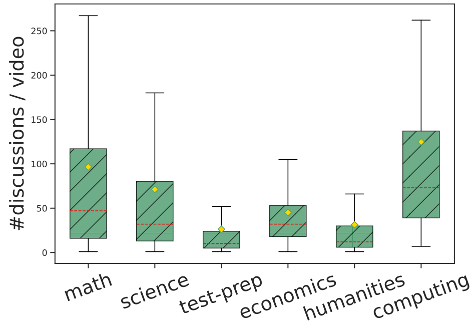
Split	**#V**	**#S/V**	**#W/S**	**#W/Q**	**#W/A**
Train	4590	7.91	198.51	39.96	80.89
Dev	895	8.37	199.04	40.02	79.26
Test	1126	8.14	198.64	39.67	81.92
Unlabeled	7283	7.70	197.96	38.46	78.58

statistics. In summary, there are 31,680 positive labeled items and 240,551 unlabeled items, i.e., 11.57% of all discussions are labeled.

This dataset also covers a series of educational topics, including math, science, careers, humanities, etc. We conduct an exploratory analysis along each topic dimension in Fig 3.2a. We observe a topic imbalance, e.g., discussions under math and science topics account for 78.88% of labeled items and 76.82% of all items. Figure 3.2b further demonstrates the imbalance within the number of discussions of each video. The number of discussions of each video under topic *test-prep* and *humanities* are significantly less than topics like *computing*. The labeled data is then split into 80% and 20% for training and test sets, respectively, based on the number of discussions in each set.



(a)



(b)

Figure 3.2: Dataset overview regarding the number of labeled and unlabeled questions and the average number of discussions for each video across all topics. We can see the unbalanced distribution of questions in each topic.

3.3 Methodology

The problem of MOOC video clip recommendation studied in this chapter can be described as follows. Given a forum discussion question, the system retrieves a ranked list of the most relevant video clips as represented by their transcripts. We assume the questions filtered by the useful question classifier are relevant to the course materials, and the most relevant video clips should be instructional for learners. Assume a MOOC video $\mathscr{V}$ lasts for T seconds, then we split it into s $t = 60$ seconds clips, where $s = \lceil \frac{T}{t} \rceil$. Then the video $\mathscr{C}$ contains clips $c_1, c_2, \cdots, c_s$. Each clip c_i is represented with its transcripts, which can be viewed as a sequence of tokens $w^i_1, w^i_2, \cdots, w^i_{|c_i|}$. We also formally define a discussion as $d_i = [q_i, \{a_i\}]$, where $\{a_i\}$ are the answers to the question q_i. Note that in some cases, the question has not been answered yet, which is common in MOOC forums. The task is retrieve a ranked list of clips $c_{i,1}, c_{i,2}, \cdots, c_{i,s}$ given each discussion d_i. Notice that the video clip recommender needs to work effectively for MOOCs in different domains that the corpus covers. Formally speaking, the recommender $\mathscr{R} : (d, \mathbf{C}) \longrightarrow \mathbf{C}_{\mathscr{R}}$ is a function that takes a discussion d and video clip list $\mathbf{C}$ as the input and returns a ranked list of clips $\mathbf{C}_{\mathscr{R}}$. We can also choose to return only the top-K relevant clips.

3.3.1 Dual-Encoder

We employ a standard neural IR architecture [126] for the ranker. It uses a dense encoder $E_C(\cdot)$, which encodes the video clip transcripts into m-dimensional real-valued vectors. At run-time, `MOOC-Rec` maps the input discussion $d = [q, a]$ to another m-dimensional vector using the query encoder $E_Q(\cdot)$, and retrieves the top-k closest video clip vectors from the same video. We use cosine similarity to model the similarity between the discussion and the clip vectors by the following function:

$$sim(d, c) = (1 - \iota(a)) \cos(E_Q(q), E_C(c)) + \iota(a) \cos(E_Q(d), E_C(c)). \tag{3.1}$$

where

$$\iota(a) = \begin{cases} 0, & \text{if } len(a) = 0 \\ \alpha, & \text{otherwises} \end{cases} \tag{3.2}$$

Some questions are not answered yet, but the `MOOC-Rec`'s performance on these questions is especially important since the recommendation can assist learners in resolving their confusion and finding the answers they need. Therefore, we define the $sim(d, c)$ in this way in order to improve the encoders' performance on unanswered discussions where the similarity is calculated based on the video clip and the question only. The parameter α is a real value between 0 and 1 to give different weights for similarities between the question, answers, and the clip.

3

The goal of training is to learn a better embedding function for both the clips and discussions which can map relevant pairs of discussions and clips to vectors with smaller distances, i.e., higher similarity, so that the similarity function $sim(d, c)$ becomes a good ranking function for the task of MOOC video clip recommendation. This is essentially a *metric learning* problem [126, 129, 130].

Let $\mathcal{M} = \{\langle d_i, c_i^+, c_{i,1}^-, \cdots, c_{i,n}^- \rangle\}_{i=1}^m$ be the training MOOC discussion corpus that contains m instances. Each example has one discussion $d_i = [q_i, a_i]$, one relevant (positive) video clip transcript c_i^+, and n irrelevant (negative) clips $c_{i,j}^-$. We train the retrieval model by optimizing the negative log-likelihood of the positive clip:

$$L(d_i, c_i^+, c_{i,1}^-, \cdots, c_{i,n}^-) = -\log \frac{e^{sim(d_i, c_i^+)}}{e^{sim(d_i, c_i^+)} + \sum_{j=1}^n e^{sim(d_i, c_{i,j}^-)}} \tag{3.3}$$

Positive and Negative Video Clips

For labeled discussions, positive and negative video examples are explicit. We use the video clip whose time duration contains the timestamp of the discussion as a positive example. All other video clips *from the same video* can be treated as negatives. As MOOC videos vary in the number of clips and to boost the model training and balance the number of positive and negative examples, we selected n of them as the training negative examples. We apply in-batch negatives [126, 131] for training. In this case, the positive clips for other questions are also treated as the negatives for the current question.

Distant Supervision with Unlabeled Data

As we show in Table 3.3, over 80% of all discussions are unlabeled (i.e., there is no video timestamp available). It would be labor-intensive and expensive to create human annotations. Thus, we adopt *distant supervision* [132] to effectively utilize the rich unlabeled data and train a better model with them. This process involves training the model with noisy, weakly labeled data. `MOOC-Rec` is able to achieve over 50% precision in top-1 prediction and over 70% in top-3 with a Recall@3 of over 80%. Therefore, we use the ranker trained on the labeled training set as the scorer, and clips with the highest $sim(d, c)$ are selected as positives, while the clips with the lowest $sim(d, c)$ (besides top-3) as negatives. The weakly labeled data are then used to train the ranker.

Inference

During inference time, we pre-compute all clip embedding v_c by applying the clip encoder E_C to all MOOC video clips offline. Given a discussion $d = [q, a]$ at run-time, we concatenate the question and answers if a is available and compute the discussion embedding $v_d = E_Q(d)$. The clips are then ranked by $sim(d, c)$, and the top-k are retrieved.

Although encoders can be implemented in many different ways [132], in this work, we use two independent **BERT** [34] variant models as encoders, and the mean value of all token embeddings is used as the final representation. We tokenize clip transcripts and truncate the token list to the maximum length of 512 (starting with `[CLS]` and ending with the `[SEP]` token). The discussion encoder works as a *query* encoder in typical neural IR systems. Instead of using separate encoders for questions and answers of the discussion, in our design, both of them share the same encoder. In this way, we train a better query encoder for questions by taking advantage of important answer information.

3.3.2 Cross-Encoder

Both the cross-encoder and dual-encoder are two common approaches for matching sentence pairs. While the dual-encoder produces sentence embedding vectors for clips and discussions independently, the cross-encoder treats the clip recommendation for discussions as a sequence classification task and performs full self-attention over the entire sequence. We concatenate the video clip transcripts and the discussions (question and answers) with the `[SEP]` token as the input to the transformer network. The `[CLS]` token embedding is then passed to a binary classifier to predict the binary relevance between them.

We train the cross-encoder model by optimizing the pointwise loss as negative log-likelihood of the positive clip:

$$L(d_i, c_i^+, c_{i,1}^-, \cdots, c_{i,n}^-) = -\log sim(d_i, c_i^+) - \sum_{j=1}^{n} \log sim(d_i, c_{i,j}^-) \tag{3.4}$$

We use the query and the video clip containing the query timestamp as a positive item labeled as 1 and select n other video clips with the lowest BM25 score to the query as the negative items labeled as 0.

3.4 Experiments and Results

3.4.1 Experimental Settings

Implementation

Two BERT variants: MPNet [133] (*abbrv.* `MP`, embedding size: 768) and MiniLM [134] (*abbrv.* `MP`, embedding size: 384) are used as text encoders. We implement dual-encoders using pre-trained weights provided by Sentence-Transformers library [3] [135]. Both models are pre-trained on a large and diverse dataset of over 1 billion training query-paragraph pairs for the semantic search task. The Adam optimizer [136] with warming-up and cosine schedule is used for training; we set the maximum learning rate to $lr = 2 \times 10^{-5}$, $\epsilon = 1 \times 10^{-8}$ and the warmup steps to 1000. For the **cross-encoder** baseline, we follow previous

[3] https://github.com/UKPLab/sentence-transformers

research [132, 137]. The **BM25** baseline is based on the Okapi BM25 implementation of the `rank_bm25` library [4]. We train our models using 8 GTX-1080 GPUs for 10 iterations with a batch size of 32. As Figure 3.3b shows, after one iteration, both clip recommendation systems outperform the BM25 baseline.

3.4.2 Evaluation Metrics

We evaluate the trained model on the test set. We investigate the performance of the trained model in terms of widely used rank-aware metrics [138, 139], including mean reciprocal rank(MRR), MRR of top-K items (MRR@K), normalized discounted cumulative gain(NDCG) and NDCG of top-K items (NDCG@K), using the implementation of the information retrieval evaluation toolkit Pytrec_eval [140] library. We focus on the performance of top predictions, therefore, we also use metrics like precision@1 (P@1). Higher MRR and $NDCG$ values suggest better ranking performance. MRR is the inverse harmonic mean of the rank of the first relevant clip in the MOOC video's clips list.

$$MRR@K = \frac{1}{n}\sum_{i=1}^{K}\frac{1}{r_i} \tag{3.5}$$

where r_i is the rank of the most relevant clip for discussion d. Since the first result of the ranking result matters for our application, MRR is a suitable metric.

NDCG@K is a precision-based metric that accounts for the predicted position of the ground truth instance. Different from MRR, rather than rewarding only the first relevant clip,

$$NDCG@K = \frac{1}{Z_K}\sum_{j=1}^{K}\frac{2^{rel_j}-1}{\log_2(1+j)} \tag{3.6}$$

where Z_K is a normalizer to ensure the perfect ranking has a value of 1. rel_j is the ground-truth rating of the item at position j.

3.4.3 Effectiveness of Dense Retrieval

Performance Comparison with Baseline

In this paper, we propose to jointly consider the similarity score of the answer and question encoding to the clip encoding vector, as shown in Equation 3.1. We conduct ablation experiments on the choice of α value. As shown in Figure 3.3a, we find when we set $\alpha = 0.4$, the system can achieve the best overall performance. We report results with $\alpha = 0.4$. After several iterations, the models' performance first improves gradually and then becomes steady, as illustrated in Figure 3.3b, which shows the effectiveness of the training system and the effectiveness of the proposed models. Table 3.4 summarizes the models' effectiveness on the test set. We use BM25 as our baseline. Sparse vector-space models and the probabilistic BM25 model have been widely used in instructional clip recommendation systems. BM25's effectiveness in terms of Precision@1 (P@1) and MRR is 0.417 and 0.60, respectively, which shows queries possess more lexical similarity to related MOOC clips than other clips in the course video and BM25 is an effective and strong baseline for this

[4] https://github.com/dorianbrown/rank_bm25

Table 3.4: Performance of the proposed MOOC-Rec ranker and baselines on the test set in terms of rank-aware metrics. MLM/MP$_{dual}$ represents the MiniLM or MPNet based dual-encoder and MLM/MP$_{cross}$ represents the MiniLM or MPNet based cross-encoder. "PT" represents ranker performance using pre-trained encoders without fine-tuning. "FT" means fine-tuned model performance. "WL" means the model performance after training with weakly labelled data.

Method		P@1	MRR	MRR@3	nDCG	nDCG@3
	BM25	0.417	0.600	0.550	0.696	0.593
PT	**MLM$_{cross}$**	0.132	0.346	0.254	0.497	0.297
	MLM$_{dual}$	0.422	0.614	0.568	0.707	0.617
	MP$_{cross}$	0.135	0.344	0.248	0.495	0.288
	MP$_{dual}$	0.386	0.583	0.529	0.683	0.576
FT	**MLM$_{cross}$**	0.511	0.677	0.641	0.755	0.683
	MLM$_{dual}$	0.529	0.692	0.658	0.767	0.700
	MP$_{cross}$	0.613	0.745	0.716	0.807	0.750
	MP$_{dual}$	0.570	0.720	0.690	0.788	0.730
WL	**MLM$_{cross}$**	0.540	0.696	0.661	0.770	0.700
	MLM$_{dual}$	0.520	0.683	0.646	0.760	0.687
	MP$_{cross}$	**0.625**	**0.751**	**0.722**	**0.812**	**0.754**
	MP$_{dual}$	0.557	0.711	0.680	0.782	0.720

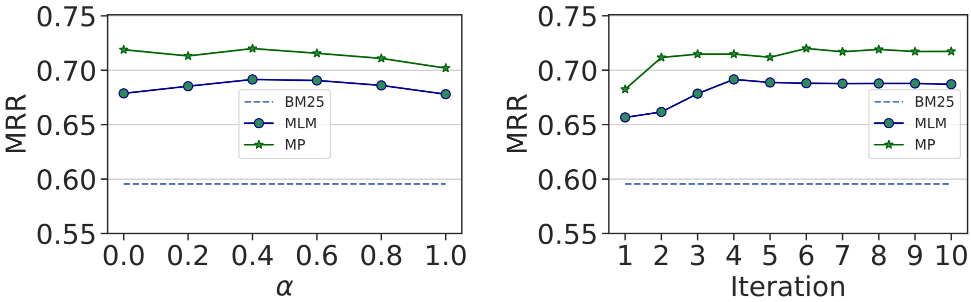
task. First, we find that without fine-tuning, the pre-trained dual-encoder can achieve similar (MPNet), or even better (MiniLM-L6) performance than the BM25 baseline, while the cross-encoders cannot make clip recommendations for discussions without training. Second, we observe significant gains ($p = 1.95 \times 10^{-7}$) when using the MOOC-Rec neural ranker after it has been trained on the data, with gains of over 0.15 in P@1 and over 0.19 in nDCG scores compared to the BM25 baseline. Thus, dense retrieval is an effective instructional MOOC clip recommendation approach for forum discussions which can model the relevance between discussions and clip transcripts.

Impact of Model Size

To compare the impacts of model size, we use one distilled transformer model MiniLM which contains **22M** parameters, and one BERT size model MPNet which contains **109M** parameters. As Table 3.4 shows, in both cross-encoder and dual-encoder settings, the larger model (i.e. MPNet) achieves better effectiveness after training, which shows that the transformer model with more parameters may have a better potential to model the relevance between clips and discussions.

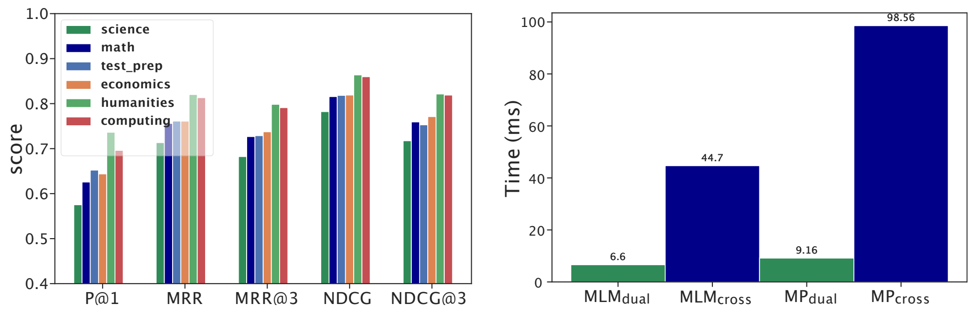
Comparison of Cross-Encoder and Dual-Encoder

Both cross-encoder and dual-encoder are commonly used for sentence pair matching problems. In Table 3.4, we observe that with the distilled transformer model, the dual-encoder outperforms the cross-encoder by 0.018 in terms P@1. However, with a large model, the cross-encoder outperforms the dual-encoder by 0.043 on P@1, and around 0.02 on other metrics. Despite the performance advantage of the cross-encoder with a large model, as



(a) (b)

Figure 3.3: (a) Ablation experiments on the α value. (b) System performance along each training iteration. The BM25 method is the baseline method. We report the performance of BM25 using the default setting. We compare the MRR scores of dual rankers based on `MiniLM` (MLM) and `MPNet` (MP).



(a) (b)

Figure 3.4: (a)The performance of `MOOC-Rec` on the MOOC-Clip dataset along all topics. (b) The average processing time of different rankers.

outlined in Section 3.3.2, we observe a massive computational overhead with the cross-encoder as illustrated in Figure 3.4b.

Effect of Distant Supervision

In the weakly-labeled data (WL) section of Table 3.4, we summarize the different models' performance after distant training with weakly labeled data. Compared with model trained with labeled data only, cross-encoders benefit from WL (+0.029 for `MiniLM` and +0.012 for `MPNet` in terms P@1), while dual-encoders perform gets worse (-0.009 for `MiniLM` and -0.013 for `MPNet` in terms P@1). Our hypothesis is that although `MOOC-Rec` achieves good effectiveness after the initial training, the weakly labeled data created with it still contains considerable noisy content.

3.5 Limitations

One limitation of this study is we assume the useful questions in discussion forums are related to the video clips' transcripts. However, the useful questions can also be related to concepts that are not explicitly mentioned the current video clip, such as prerequisite courses. Creating context for discussion forum questions with both the course materials and external knowledge might further alleviate the information overload, and downstream tasks such as question generation and answering.

3.6 Conclusions

We studied the task of video clip recommendation in the context of MOOC forums which has the eventual goal to reduce learners' information overload. We created a novel dataset `MOOC-Clip` which includes video transcripts and discussions. We systematically investigated how well the state-of-the-art pre-trained neural IR models work for the task of MOOC clip recommendation, and proposed a novel approach including data preparation, useful question classification, clip ranker, and weak supervision training for this task. We conducted the experiments with both cross-encoders and dual-encoders. The results on our dataset show that neural IR approaches are indeed effective—at the same time, a P@1 value of less than 0.63 (at best) shows that we are still far away from solving this task. In future work, we plan to further investigate the factors that affect `MOOC-Rec`'s effectiveness such as the clip duration and methods of creating weak labels.

4

Answer Quality Aware Aggregation for Extractive QA Crowdsourcing

In this chapter, we focus on quality control for creating datasets via crowdsourcing. Quality control is essential for creating extractive question answering (EQA) datasets via crowdsourcing. Aggregation across answers, i.e., word spans within passages annotated by different crowd workers, is one major focus for ensuring quality. However, crowd workers cannot reach a consensus answer on a considerable portion of questions. We introduce a simple yet effective answer aggregation method that takes into account the relations among the answer, question, and context passage. We evaluate answer quality from the perspectives of the QA model to determine how confident the QA model is about each answer annotation and the answer verification model to determine whether the answer annotation is correct. Then, we compute aggregation scores using the quality of each answer annotation and its contextual embedding produced by pre-trained language models. The experiments on a large real crowdsourced EQA dataset show that the proposed approach outperforms baselines by around 16% on precision and effectively conducts answer aggregation for EQA. The code is available at https://github.com/zpeide/Answer-Quality-Aware-Aggregation.

This chapter is based on the following conference paper: *Zhu, Zhen Wang, Claudia Hauff, Jie Yang, and Avishek Anand. 2022. "Answer Quality Aware Aggregation for Extractive QA Crowdsourcing.", findings@EMNLP'22.* [87]

4.1 Introduction

Extractive Question answering (EQA) is a fundamental task in natural language processing [141]. With access to large-scale datasets, deep neural models have achieved significant advances in the EQA task [34, 142, 143]. Creating large-scale, high-quality datasets is one of the essential factors driving this progress [144]. Currently, a prevalent method for creating EQA datasets is crowdsourcing [50, 55, 94, 145, 146] thanks to its efficiency and scalability due to the availability of crowd workers. Yet, answers collected from crowd workers often contain a substantial amount of noise due to the reliability issue of crowd workers affected by their varying expertise, skills, and motivation [147, 148].

Question	***What did the GOP leaders say?***	**Vote**	***Agreement Measure***	
Answer_1	*Newt Gingrich called Sotomayor a racist*	0	0.3433	✗
Answer_2	*he wants more than an explanation*	0	0.3118	✗
Answer_3	*they were discriminated against after a promotion test was thrown out, because critics said it discriminated against minority firefighters.*	2	0.5564	✓

WASHINGTON (CNN) -- During the presidential campaign, then-candidate Barack Obama said that he hoped his administration wouldn't get [...]ssue. Former Republican Speaker of the House *Newt Gingrich called Sotomayor a racist*. Conservative talk [...] a better conclusion than a white male who hasn't lived that life." One top GOP senator said *he wants more than an explanation*. "I think she should apologize, but I don't believe any American wants a judge on the bench that's going to use empathy or their background to punish someon. "She's been called the equivalent of the head of the Ku Klux Klan by Rush Limbaugh; [...] yor's appellate court decision against a mostly white group of firefighters who say *they were discriminated against after a promotion test was thrown out, because critics said it discriminated against minority firefighters.* But legal experts have said her full record on race isn't that controversial -- in 96 race-related cases decided by Sotomayor on the court of appeals, ...

Figure 4.1: An example of answer aggregation for QA crowdsourcing. In this example, three crowd workers are asked to select a word span in the passage as the answer to the question. The gold answer can be aggregated from the disagreed answers by asking another group of workers for answer selection (vote) or using answer aggregation models (aggregation measure).

To reduce noise in crowdsourced data, a widely-adopted solution in previous crowdsourcing research is to assign each instance to multiple crowd workers to crate redundant annotations [55, 145, 146]. Aggregation across answers provided by different crowd workers thus becomes one primary focus for crowdsourcing EQA datasets. Majority voting is

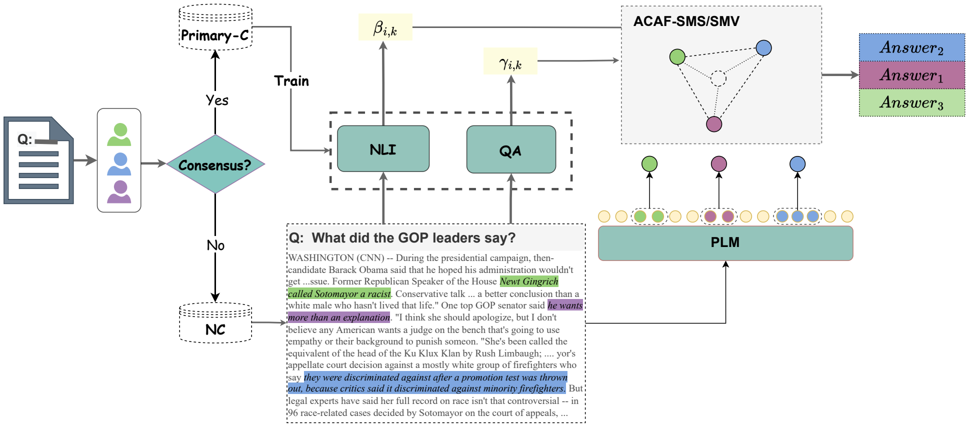


Figure 4.2: System overview and an example of automatic answer aggregation. Crowd workers are asked to label answer spans in passages for the given questions. If they achieve consensus, the QA pairs are used to fine-tune the natural language inference (NLI) based answer correctness evaluation model and the question answering (QA) model. Then we sort the non-consensus answers based on their encoding using a pre-trained language model (PLM), the answer correctness ($\beta_{i,k}$), and the question answering confidence ($\gamma_{i,k}$).

a simple and widely adopted aggregation method [149] which elects answers that most crowd workers agree with. However, most of these majority voting based methods are for categorical labels where the label space is small enough such that workers will more likely produce the same label [150–153]. They cannot apply to this EQA task where the answer candidates are word spans rather than a limited number of categorical labels, due to the huge number of words in the dictionary. There are some methods for automatically aggregating text sequences [154, 155], but they only apply to free text sequence tasks such as translation. Unlike free text sequence tasks, answer candidates are word spans within context passages, and their quality is related to both the question and the context passage. The previous methods do not consider these dependencies. Therefore, answer aggregation for EQA is commonly performed by having a second group of workers select and verify answers [145, 156]. As the example in Figure 4.1 shows, crowd workers provide three distinct answer spans for the same instance. Another three crowd workers are then asked to vote for each answer annotation. $Answer_3$ got 2 votes and is selected as the ground-truth answer for the question. This method requires more resources and human efforts.

In this chapter, we first model the candidate answer as a text sequence aggregation problem [155]. Previous methods aggregate the best answer based on inter-answer distances of their vector representation. As answers for EQA are word spans within context passages, we adapt previous methods by presenting answers using contextual vector embedding produced by pre-trained language models [157]. In previous research, answer quality is evaluated by estimating worker reliability. However, we argue that in EQA, answer quality can also be evaluated based on its relation to the context passage and the question. We investigate answer quality evaluation from both the view of question answering (*Answer Confidence* measure) by using QA models and from the view of answer verification (*Answer Correctness* measure) by using natural language inference (NLI) models. We further propose a novel joint approach to incorporate the answer quality measures with the inter-answer distances based answer aggregation methods for EQA, as shown in Figure 4.2.

With this chapter, we make the following contributions:

- We propose a simple yet effective novel aggregation approach for aggregating crowd-sourced answer annotations for EQA.
- We explore two answer quality measures *Answer Confidence* and *Answer Correctness* using weak heuristic question answering signal and NLI models and illustrate their effectiveness.
- The comprehensive experiments on a real large-scale crowdsourced QA dataset suggest the effectiveness of the proposed answer quality measures and the proposed answer aggregation methods. The results show that the proposed approach can effectively leverage the rich information of context passage, questions and answer candidates for answer aggregation and achieve an improvement of around 15% on precision to baseline methods.

4.2 Background

4.2.1 Crowdsourcing for QA Dataset Creation

Quality control in crowdsourcing has attracted intensive research [147, 148, 158–160]. To reduce the noises of crowdsourced data, each data instance is commonly assigned to multiple workers to create redundant annotations to infer the hidden ground truth by aggregation [55, 145, 146]. In contrast to classification or categorical crowdsourcing tasks [152, 153, 161–163] which have small label spaces, it is harder for crowd workers to achieve consensus on the answer for the same question.

What signals the disagreement contains and how to effectively use them is an interesting research question [164, 165]. Most existing work on this question focuses on classification problems. Some work [166, 167] found that it is possible to use noisy answers as weak supervision signals to improve QA performance, especially in low-resource domains. However, they still rely on the existence of ground-truth answers which are obtained by crowdsourcing. In practice, multistage methods are commonly adopted for answer aggregation in QA [56, 145, 156]. For example, a four-stage collection process is utilized for collecting NewsQA [145]. Each item is assigned to multiple crowd workers (avg. 2.73) to make answer annotations. Then another group (avg. group size is 2.48) is asked to validate distinct answer annotations collected in the previous stage). The Google Natural Questions dataset [56] evaluates non-null answer correctness with consensus judgments from 4 "experts" and the k-way annotations (with k = 25) on a subset. This approach leads to more cost of human efforts, time, and money.

4.2.2 Crowdsourced Text Sequence Aggregation

Majority Voting is the most common and simplest aggregation method. It assumes most workers have comparable accuracy and reliability on the task. Thus some workers will produce the same answer for the same question, especially for categorical label tasks where the label space is small enough. However, it can perform poorly on complex sequence labeling tasks such as translation, summarization, and question answering. The number of words in the dictionary is so huge that it is difficult for workers to produce the same answer so that the ground truth answer can be found. Therefore multi-stage crowdsourcing patterns are used to resolve disagreements by selecting, verifying, or correcting answers like the aforementioned methods in the last subsection. Several automation methods have been proposed to reduce human labor. Li [154], Li and Fukumoto [155] converted the answer texts into embeddings and extracted the potential optimal answer by estimating the embeddings of the true answer, considering both worker reliability and sequence representation. Braylan and Lease [168] proposed a single, general annotation and aggregation model by modeling label distances to support diverse tasks such as translation and sequence labels. Braylan and Lease [169] proposed to perform answer aggregation on complex annotations such as sequence labeling and multi-object image annotation by matching and merging different labels. Although the proposed methods have achieved great advantages in complex answer aggregation, little research focuses on the question answering crowdsourcing.

4.3 Method

4.3.1 Problem Definition

For the extractive answer labeling task, each instance D_i assigned to crowd workers is a tuple containing a *context passage* P_i and a *question* Q_i, i.e. $D_i = (P_i, Q_i)$. The worker k is asked to select a word span $A_{i,k}$ from the context passage $A_{i,k} = (A^s_{i,k}, A^e_{i,k})$, s, e indicates the start and end position of the answer in the passage, or `NULL` if no answer is present in the passage. Then we get a set of answers for question Q_i: $\mathscr{A}_i = \{A_{i,k}\}_1^K$ from K workers. The answer aggregation model aims to select one answer from $\mathscr{A}_i$ as the golden answer or reject all answers. In this chapter, we focus on designing an effective automation answer aggregation model to reduce human labor for multi-stage answer selection and verification, especially when none of them agree with each other. We achieve this goal by making a ranked list of all answers, so the answers with the highest evaluation score are ranked in front.

4.3.2 Text Sequence Aggregation for Answer Aggregation

As word spans from context passages, we first model the answer aggregation problem as a free text sequence aggregation problem and adopt the free text sequence aggregation methods *Sequence Majority Voting (SMV)* and *Sequence Maximum Similarity (SMS)* on it [155]. These methods perform text sequence aggregation based on answers' vector representations.

Answer Representation

Different from text sequence aggregation problems like translation, the answer correctness depends not only on the answer word span but also on its context. Therefore, to produce a single vector representation of each answer, instead of encoding the answer independently, we get the answer's contextual embedding by encoding the passage containing the answer with transformers-based pre-trained language models. Then we use the mean value of all answer token embeddings as the embedding of the answer. Formally, we define the passage which consists a sequence of words as $P_i = \{p_j\}_{j=1}^{|P_i|}$ (with $|P_i|$ being the length of the passage and p_j being the tokens in the passage), the language model as E and the token-wise encoding as:

$$\{\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{|P_i|}\} = E(\{p_1, p_2, \cdots, p_{|P_i|}\})$$

then the answer representation $\hat{a_{i,k}}$ is produced by: $\hat{a}_{i,k} = mean(\{\hat{p}_{A^s_{i,k}} : \hat{p}_{A^e_{i,k}}\})$

Sequence Majority Voting (SMV)

The SMV method proposed by Li and Fukumoto [155] is the direct adaptation of majority voting to the sequence label problem. *SMV* estimates the true answer embedding $\hat{e}_i$ as the mean vector of all answer vector representations:

4

$$\hat{e}_i = mean(\hat{a}_{i,1}, \hat{a}_{i,2}, \cdots, \hat{a}_{i,K}) \tag{4.1}$$

and ranks answer candidates according to their similarity to $\hat{e}_i$ and extracts the golden answer $\hat{z}_i$ as the answer candidate with the most semantic similarity to $\hat{e}_i$:

$$s_{i,k} = sim(\hat{a}_{i,k}, \hat{e}_i) \tag{4.2}$$

Sequence Maximum Similarity (SMS)

The SMS method was first proposed for the unsupervised ensemble of outputs of multiple text generation models [170]. It selects the gold output by selecting a majority-like output close to other outputs by using cosine similarity, which is an approximation of finding the maximum density point by kernel density estimation. Li and Fukumoto [155] adopts SMS for crowdsourcing translation data which are generated by crowd workers instead of text generation models. However, they only use it on free text sequences. In this chapter, we further adopt it to the extractive QA task. We produce answer representation as forementioned, and extract the golden answer $\hat{z}_i$ as the answer candidate with the largest sum of similarity $s_{i,k}$ with other answer annotations of the same question:

$$s_{i,k} = \frac{1}{|\mathscr{A}_i| - 1} \sum_{k_1 \neq k} sim(\hat{a}_{i,k_1}, \hat{a}_{i,k}) \tag{4.3}$$

4.3.3 Answer Quality Aware Answer Aggregation

The answer representations concentrate on answer contextual representation only, but the quality of each answer also depends on whether it can answer the question based on the context passage. The answer text sequence aggregation methods cannot fully utilize the rich information of both the context and question. Therefore, we further propose to aggregate crowdsourced answers in an answer quality-aware way. We first propose to evaluate answer quality from the view of the question answering model (*Answer Confidence*) and the view of the answer verification model (*Answer Correctness*). Due to the lack of labeled data for training the QA and NLI models, the prediction of these models is noisy and inaccurate. However, they can still provide hints on answer quality. Then we propose a novel aggregation method to strengthen the influences of possible high-quality answers (**ACAF-SMS/SMV**).

Answer Quality Evaluation

Answer Confidence (AF)

We use BERT-QA [34] as our QA model. It consists of two parts, the BERT encoder and the answer classifier. The answer classifier predicts the distributions of the start position and the end position separately based on the outputs of the BERT encoder. As argued by Zhu and Hauff [85], Xie *et al.* [100], the QA model should be quite confident about the prediction of the answer start/end span to the answerable question. Thus the prediction probability distribution should peak on both $A^s_{i,k}$ and $A^e_{i,k}$. Therefore, the geometric average of these start position probability ($\Pr_s(s|P_i, Q_i)$) and end position probability ($\Pr_e(e|P_i, Q_i)$) distributions can be used as a heuristic of the confidence of the answer prediction. Formally, We define the answer confidence $\gamma_{i,k}$ as follows:

$$\gamma_{i,k} = \max_{A^s_{i,k}-w \le b \le c \le A^e_{i,k}+w} \sqrt{\Pr_s(b|P_i, Q_i) \cdot \Pr_e(c|P_i, Q_i)}. \tag{4.4}$$

where w is search window size.

Answer Correctness (AC)

QA models often lack the ability to verify the correctness of the predicted answer [171]. One way to address this issue is to reformulate it to a textual entailment problem [171–173] by viewing the answer context as the premise and the QA pair as the hypothesis. Then we use a natural language inference (NLI) system to verify whether the candidate answer proposed by crowd workers satisfies the entailment criterion. We use the transformers-based pre-trained sequence classification model for answer correctness verification. We treat the answer candidate as a short text sequence (answer-text), and formulate the input to the model in the format " `[CLS]` *question* `[SEP]` *passage* `[SEP]` *answer-text* `[SEP]` ". We truncate passages longer than the maximum 512 tokens and only keep the sentences containing the answer span. The embedding of the `[CLS]` token is used as the pooling encoding of the sequence, and a linear classification layer has performed the encoding. Finally, according to the passage, we use the $softmax$ function to get the final probability that an answer candidate is correct.

$$\beta_{i,k} = \mathbf{V}(P_i, Q_i, A_{i,k}) \tag{4.5}$$

Above, $\mathbf{V}$ represents the NLI model to verify the answer's correctness. $\beta_{i,k}$ is the probability that the answer $A_{i,k}$ to question Q_i is correct.

We then propose to combine the answer confidence and the answer correctness probability for answer quality evaluation. Assuming these two measures are complementary, to make the method simple, we combine them as a simple sum:

$$v_{i,k} = \gamma_{i,k} + \beta_{i,k}. \tag{4.6}$$

The Joint Method (ACAF-SMS/SMV)

We propose to join the NLI model, QA model, and contextual answer vector representations for answer aggregation by incorporating the answer correctness probability and answer confidence with sequence aggregation methods **SMV** and **SMS** to strengthen the influence of high-quality answers further. The joint sequence majority voting (**ACAF-SMV**) method computes the answer aggregation measure $s_{i,k}$ as:

$$s_{i,k} = \frac{v_{i,k}}{\sum_k v_{i,k}} sim(\hat{a}_{i,k}, \hat{e}_i) \tag{4.7}$$

and the joint sequence maximum similarity (**ACAF-SMS**) method as:

$$s_{i,k} = v_{i,k} \frac{\sum_{k_1 \neq k} v_{i,k_1} \cdot sim(\hat{a}_{i,k_1}, \hat{a}_{i,k})}{\sum_{k_1 \neq k} v_{i,k_1}} \tag{4.8}$$

The **AF-SMS** algorithm and **AF-SMV** algorithms are similar to the methods mentioned above by replacing answer correctness probability $\beta_{i,k}$ with answer confidence $\gamma_{i,k}$ or $r_{i,k}$. Figure 4.2 illustrates the proposed method, and the pseudo-code in Algorithm 1.

4

Algorithm 1: ACAF-SMS/SMV answer aggregation algorithm.

Input : Passage: P_i;
Question: Q_i;
Answer candidates: $\{A_{i,k}\}_1^K$;
Answer verification model: **V**;
Sequence encoder: E
Question answer model $G(P_i, Q_i)$
Output : Ranked answer candidate list
for *Answer* $A_{i,k}$ **do**
  $\beta_{i,k} = \mathbf{V}(P_i, Q_i, A_{i,k})$;
  $e_i^k = E(A_{i,k})$
  Answer start position probability: $Pr_s(A^s_{i,k} \dots A^e_{i,k} | P_i, Q_i) = G(P_i, Q_i)$
  Answer end position probability $Pr_e(A^s_{i,k} \dots A^e_{i,k} | P_i, Q_i) = G(P_i, Q_i)$
  $\gamma_{i,k} = \max_{A^s_{i,k} - w \leq i \leq j \leq A^e_{i,k} + w} \sqrt{Pr_s(i|P_i, Q_i) \cdot Pr_e(j|P_i, Q_i)}$.;
end
if *Using SMV* **then**
  $\hat{e}_i = mean(E(\mathcal{A}_i))$;
  $\omega_{i,k} = \beta_{i,k} + \gamma_{i,k}$;
  $s_{i,k} = \frac{\omega_{i,k}}{\sum_k \omega_{i,k}} sim(E(A_{i,k}), \hat{e}_i)$
if *Using SMS* **then**
  $\omega_{i,k} = \beta_{i,k} + \gamma_{i,k}$;
  $s_{i,k} = \omega_{i,k} \frac{\sum_{k_1 \neq k} \omega_{i,k_1} \cdot sim(E(A_{i,k_1}), E(A_{i,k}))}{\sum_{k_1 \neq k} \omega_{i,k_1}}$;
Rank answer list according to $s_{i,k}$;

4.4 Experimental Setup

4.4.1 Dataset

We evaluate the proposed method with the NewsQA dataset because it provides all crowd-sourced raw answer annotations. The creation process of NewsQA demonstrates the chal-

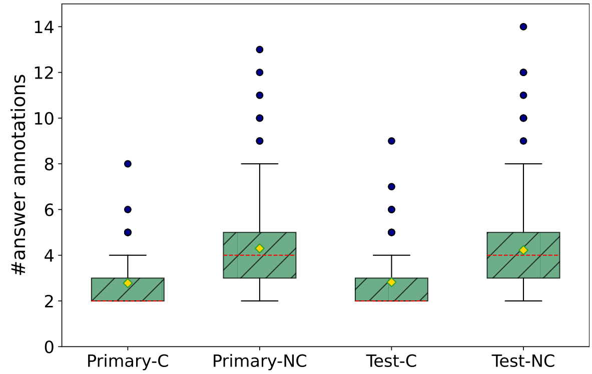


Figure 4.3: Number of answer annotations for questions in the four datasets we use, including the primary consensus (Primary-C) set, the primary non-consensus (Primary-NC), the test consensus (Test-C) set, and the test non-consensus (Test-NC) set.

lenges of QA dataset crowdsourcing and the importance and necessity of answer aggregation. Answers in the NewsQA are collected through a two-stage process: the primary stage (answer sourcing) and the validation stage. In the primary stage, each question solicits answers from avg. 2.73 crowdworkers. 56.8% of questions have consensus answers between at least two answers on the primary stage. 37.8% of questions got consensus answers after the validation stage. Crowdworkers do not come to a consensus for the rest 5.3% questions.

In this chapter, we split NewsQA into four subsets: the ***primary consensus (Primary-C)*** set, which contains all passages, questions and their answers from the training set that achieve answer agreement on the primary stage; the ***primary non-consensus (Primary-NC)*** which contains all passages, questions and answer candidates that only achieve agreement after an additional round of answer validation from the training set; ***test consensus (Test-C)*** set which contains passages, questions and answers that achieve consensus from the test set, and the ***test non-consensus (Test-NC)*** set which contains data items that only reach consensus after an additional round of answer validation from the test set. Figure 4.3 shows the boxplot of the number of crowdsourced answers for each question. There are more than four distinct answers per question in non-consensus sets. The Primary-C and Test-C sets are gold answers that can be used for training and evaluating the NLI and QA models used for answer aggregation. The Primary-NC and Test-NC sets are used for evaluating the proposed method. Passages in the training set do not contain passages in the test set, making our evaluation generative. Table 4.1 shows the statistics of our data.

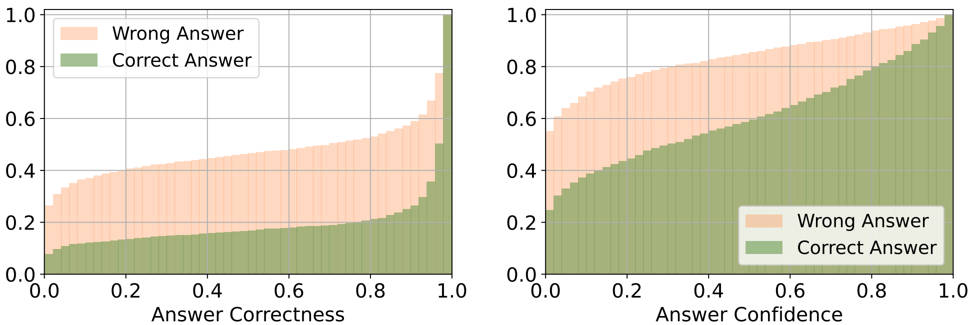
4.4.2 Hyper Parameters

Hyper-parameters for Training The NLI Model

Adam optimizer [136] with warming-up and linear schedule is used for fine-tuning the answer verification model. We set the maximum learning rate (lr) as $lr = 2 \times 10^{-5}$ and $\epsilon = 1 \times 10^{-8}$ and the warmup steps of 1000. The models are trained on a server using 4 GTX-1080 GPUs for 20,000 iterations, where each iteration is a batch size of 32 and uses the best-performing checkpoint.

Table 4.1: Statistics of the datasets; number of passages $|P|$; number of answerable questions $|Q_A|$; number of unanswerable questions $|Q_U|$; number of correct answers $|A_C|$ and number of wrong answers $|A_W|$.

Data	$\|P\|$	$\|Q\|$	$\|A_C\|$	$\|A_W\|$
Primary-C	11,469	61171	93,842	76,163
Primary-NC	11,469	40713	52,941	122,071
Test-C	634	3393	2,306	1,906
Test-NC	637	2273	2,980	6,620



(a) (b)

Figure 4.4: (a) Cumulative answer correctness distribution and (b) cumulative answer confidence distribution on correct answers and incorrect answers.

Hyper-parameters for Training The QA Model

The QA model is trained on the same server consisting of 4 GeForce GTX 1080 GPUs with a batch size of 32, the maximum learning rate of 1×10^{-5} with Adam as the optimizer for 10 epochs and take the epoch with the best validation accuracy as the final model.

4.4.3 Baselines

Random Selection (RS)

The baseline is to rank answer annotations randomly for each question. We report the RS performance as the average performance over five rounds of random answer ranking.

Context-Free (CF) SMS/SMV

This baseline is to produce answer representation by treating answers as free text sequences without considering the context passages, i.e., the original SMS/SMV methods proposed by Li and Fukumoto [155].

4.4.4 Evaluation

For each question, we sort the answers by the proposed aggregation methods. We evaluate the results in terms of widely used rank-aware metrics, including Precision@1 (P@1), Recall@1 (R@1), Mean Average Precision (MAP), and normalized discounted cumulative gain(NDCG). We choose the implementation of the information retrieval evaluation toolkit Pytrec_eval [140] library.

4.5 Results and Analysis

4.5.1 Effectiveness of Answer Quality Evaluation Methods

Performance of AC on Answer Classification

We train the NLI model for producing AC using the BERT for sequence classification implementation from the Huggingface Transformers library [157] on the Primary-C set. It achieves 80.65% in accuracy and 87.59% in F1 on the Test-C set. On the Test-NC set, its performance is 62.57% in accuracy and 64.52% in F1, which is much worse than its performance on the Test-C set. The results indicate answers to questions that achieve consensus in the first sourcing stage are relatively more distinguishable and show the difficulty of specifying the correctness of disagreed answers. Figure 4.4a and Figure 4.5 show that AC is an effective metric to distinguish correct and wrong answers, which achieves 0.70 in AOC.

4

Performance of AF on Answer Classification

We train the QA model using the BERT-QA implementation from the Huggingface Transformers library on the Primary-C set and adopt the exact match (EM) and F1 score (F1) to evaluate its performance. Table 4.3 reports the QA performance. The QA model achieves 27.94% and 60.89% in EM and F1, respectively, on the Test-C set. In contrast, its performance on the Test-NC set is 9.15% and 37.22% in EM and F1, which is much worse than its performance on Test-C and demonstrates the difficulty of automatically answering these questions. Although its performance is poor due to the lack of enough training data, we observe that the AF score is an effective metric for correct answer classification as shown in Figure 4.4 and Figure 4.5 and achieves 0.71 in AOC, which is slightly better than AC. The combination of AC and AF (AC+AF) improves answer classification performance by up to 4% by a simple sum.

Performance of Answer Quality Evaluation on Answer Aggregation

In Table 4.2, the rows **AC**, **AF** and **AC+AF** show the experimental results of performing answer aggregation by ranking answers according to AC, AF or by combining them(AC+AF). AC and AF have comparable performance; both achieve over 57% on P@1 and around 10% improvement over baselines, which shows the effectiveness of the proposed signals. By combining the NLI model prediction and the QA model heuristic signal, we can further improve the P@1 performance by around 3% on both Primary-NC and Test-NC sets, which shows the complementary strengths of the two signals.

4.5.2 Effectiveness of Answer Text Sequence Aggregation

As shown in Table 4.2, ***SMV*** and ***SMS*** can achieve similar performance to AC and AF by using the pre-trained BERT-base model as encoder without any fine-tuning. This suggests the effectiveness of modeling answer aggregation for the extractive QA task as a sequence answer aggregation problem. These methods outperform the context-free sequence aggregation baselines by about 10%, which proves the importance of contextual

Table 4.2: Experimental results of baselines and the proposed approach of answer aggregation on *Primary-NC* and *Test-NC* set using the BERT-base-uncased model.

	Method	Primary-NC				Test-NC			
		P@1	R@1	MAP	NDCG	P@1	R@1	MAP	NDCG
Baselines	**RS**	0.4728	0.3574	0.6550	0.7385	0.4782	0.3610	0.6600	0.7429
	CF-SMV	0.4660	0.3818	0.6536	0.7334	0.4765	0.3940	0.6629	0.7409
	CF-SMS	0.4683	0.3800	0.6545	0.7339	0.4831	0.3952	0.6645	0.7419
Answer Quality	**AC**	0.5703	0.4364	0.7280	0.7902	0.5820	0.4451	0.7369	0.7973
	AF	0.5796	0.4423	0.7310	0.7917	0.5878	0.4512	0.7376	0.7971
	AC+AF	0.6022	0.4595	0.7471	0.8034	0.6128	0.4677	0.7546	0.8096
Sequence Aggregation	**SMV**	0.5685	0.4124	0.7194	0.7822	0.5816	0.4234	0.7278	0.7894
	SMS	0.5701	0.4087	0.7190	0.7816	0.5851	0.4225	0.7282	0.7892
Joint Method	**AC-SMV**	0.6036	0.4467	0.7400	0.7985	0.6124	0.4528	0.7472	0.8047
	AF-SMV	0.6009	0.4544	0.7434	0.7997	0.6106	0.4634	0.7507	0.8057
	AC-SMS	0.6008	0.4450	0.7393	0.7978	0.6194	0.4598	0.7526	0.8084
	AF-SMS	0.6011	0.4538	0.7449	0.8007	0.6190	0.4687	0.7563	0.8099
	ACAF-SMV	**0.6213**	0.4646	**0.7533**	**0.8079**	0.6274	0.4698	0.7606	0.8140
	ACAF-SMS	0.6165	**0.4647**	0.7530	0.8076	**0.6304**	**0.4762**	**0.7635**	**0.8159**

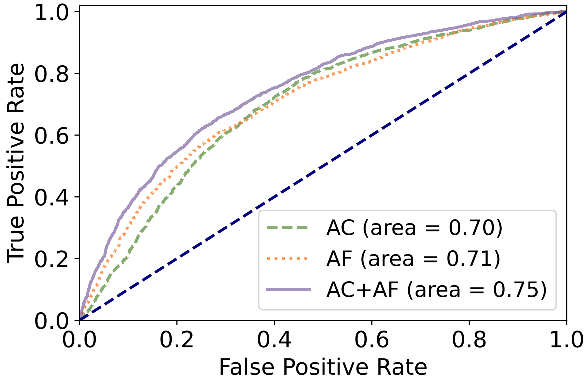


Figure 4.5: ROC Curve and area under the curve (AOC) of different answer classification methods, including answer correctness (AC), answer confidence (AF), and their combination.

Table 4.3: The performance of the QA model trained on the Primary-C set.

Method	Test-C		Test-NC	
	Exact	F1	Exact	F1
Bert-base	27.94	60.89	9.15	37.22
Bert-large	31.21	62.21	12.23	37.33
Roberta-base	32.24	66.65	13.11	43.94

embedding. Since both SMV and SMS are based on the latent semantic similarity among answer candidates, the effectiveness of these methods implies the crowdsourced answers bear some common knowledge or contextual information that can be further explored.

We then conduct experiments by combining AC, AF with SMS and SMV separately (**AC-SMV**, **AF-SMV**, **AC-SMS** and **AF-SMS**). Results in Table 4.2 show that the proposed joint methods achieve around 3% absolute performance improvement on P@1, around 5% on R@1 than using SMS and SMV only and similar to AC+AF (only slightly worse). By combining AC+AF with SMS or SMV (ACAF-SMS / ACAF-SMV), the system performance is further improved by around 2% on P@1 and around 1% on other metrics. These findings first suggest the effectiveness of the joint aggregation method. They also demonstrate that the system can achieve better performance by combining unsupervised contextual answer representation and weak learned signals.

4.5.3 Influence of Encoders

Table 4.4 show the performance of the joint methods ACAF-SMV and ACAF-SMS on Test-NC set using different types of pre-trained encoders `BERT-base`, `BERT-large`, `Roberta-base` and `BART-base`. The results first show the performance of both methods is robust alongside different encoders with different model sizes, types, and pre-training methods, demonstrating the effectiveness and stability of the proposed methods. Second, ACAF-SMS outperforms ACAF-SMV with all kinds of encoders on the Test-NC set.

Table 4.4: Results of answer aggregation using different encoders.

Model	P@1	R@1	MAP	NDCG
	ACAF-SMV			
BERT-base	0.6274	0.4698	**0.7606**	**0.8140**
BERT-large	0.6238	0.4670	0.7575	0.8117
Roberta-base	0.6247	0.4653	0.7570	0.8111
BART-base	**0.6291**	**0.4724**	0.7545	0.8098
	ACAF-SMS			
BERT-base	**0.6304**	0.4762	0.7635	0.8159
BERT-large	0.6247	0.4726	0.7607	0.8140
Roberta-base	0.6300	**0.4786**	**0.7638**	**0.8162**
BART-base	0.6304	0.4750	0.7633	0.8155

4

4.5.4 Evaluation with More Metrics

Besides the rank-aware metrics, we also compare the method performance of the top-1 answer using two evaluation metrics: Exact Match (EM) and the macro-averaged F1 score. The EM metric measures the percentage of answer predictions that match the ground truth answer exactly. The F1 score calculates the F1 of the overlap of the bag-of-words representation of the ground truth and top-1 answers. We use the implementation of Exact Match and F1 from MRQA [174][1]. We report the EM and F1 results in Table 4.5. The results further demonstrate the effectiveness of the joint methods **ACAF-SMV** and **ACAF-SMS**, aligning with the conclusion from Table 4.2.

4.5.5 Impact of Answer Selection on QA Performance

To explore the impact of answer selection on QA performance on the NewsQA dataset. We first train BERT-base-QA models on data where answers are the top-1 answer selected by the proposed methods **ACAF-SMS** and **ACAF-SMV** against answers selected by humans. Table 4.6 demonstrates the EM and F1 scores of these experiments. We observe that the F1-scores of the proposed methods ACAF-SMS (59.68) and ACAF-SMV (60.56) are very close to the performance of the human-labeled data (61.12). We further investigate the effectiveness of our method by using them as additional voters for selecting the best answers in combination with human voting. Results show that the QA performance can be improved to 61.63 (ACAF-SMS as the voter) and 62.27 (ACAF-SMV as the voter), surpassing the human-selection-only setting. These results show that the automated answer aggregation method can be used for aggregating the disagreements among annotators regarding the impacts of the results on the QA models.

4.5.6 Answer Aggregation Results on Other Datasets

In addition to the NewsQA dataset, we perform experiments on two extractive QA datasets, including the SQuAD and Natural Questions datasets. The SQuAD and Natural Questions

[1] `https://github.com/mrqa/MRQA-Shared-Task-2019`

Table 4.5: Performance of answer agreement on *Primary-NC* and *Test-NC* using the BERT-base-uncased model in terms of Exact Match (EM) and F1.

Method	Primary-NC		Test-NC	
	EM	F1	EM	F1
RS	0.4666	0.5246	0.4656	0.5292
CF-SMV	0.4640	0.5690	0.4729	0.5750
CF-SMS	0.4669	0.5696	0.4773	0.5749
AC	0.5638	0.6140	0.5689	0.6182
AF	0.5751	0.6300	0.5829	0.6337
AC+AF	0.5933	0.6426	0.5970	0.6454
SMV	0.5584	0.6179	0.5693	0.6287
SMS	0.5626	0.6202	0.5733	0.6309
AC-SMV	0.5980	0.6478	0.6027	0.6525
AF-SMV	0.5900	0.6449	0.5944	0.6459
AC-SMS	0.5957	0.6445	0.6089	0.6546
AF-SMS	0.5896	0.6423	0.6036	0.6492
ACAF-SMV	**0.6132**	**0.6626**	0.6146	**0.6652**
ACAF-SMS	0.6085	0.6568	**0.6168**	0.6622

datasets only provide multiple annotations for dev sets. Therefore, we perform experiments on both datasets by treating the training set as the Primary-C set and selecting questions with multiple different annotations and one consensus answer as the Primary-NC set. To train the NLI models needed for answer verification, besides the ground truth answers, we create negative answers by sampling different word spans with the same named entity types, if possible, or word spans with the most similar part-of-speech(POS) structures. Table 4.7 presents the results on the two extra datasets. The joint methods (ACAF-SMS and ACAF-SMV) achieve similar performance (the performance differences are around 1%) and generally outperform SMS and SMV across all metrics in both datasets significantly. These results further highlight the effectiveness of incorporating answer quality evaluation like answer correctness and answer confidence for answer aggregation.

4.5.7 Case Study

As shown in Table 4.8, we conduct a case study to examine the performance of the proposed approach. In this case, AC, AC+AF, and SMS suggest *waste* is the correct answer. However, its answer confidence is very low(0.0025). AF points *great pacific garbage patch that stretches* is the best answer. Only **ACAF-SMS** ranks the golden answer *of the pacific ocean* as the best answer, even though the AC and AF scores of this answer are not the highest.

Table 4.9 further shows two examples of answer aggregation results. The first example contains seven answer annotations. It is notable that all answer annotations are generally located in the related parts of the context. However, some answer annotations are too vague or point to the wrong spans. Although the answer correctness evaluation model and

Table 4.6: The performance of the QA model trained on datasets created with different methods, including the answer aggregation with humans (**GroundTruth**), the **ACAF-SMS** method, the **ACAF-SMV** method, and answer selected by both human annotators and the automatic answer aggregation method (**ACAF-SMV**$_{\text{voter}}$ and **ACAF-SMS**$_{\text{voter}}$).

Method	Exact	F1
GroundTruth	28.00	61.12
ACAF-SMS	25.94	59.68
ACAF-SMV	26.37	60.56
ACAF-SMS$_{\textbf{voter}}$	27.44	61.63
ACAF-SMV$_{\textbf{voter}}$	**28.55**	**62.27**

Table 4.7: Performance of answer aggregation on SQuAD and Natural Questions.

Method	P@1	R@1	MAP	NDCG
	SQuAD			
SMS	0.6251	0.4829	0.8064	0.8573
SMV	0.8150	0.4787	0.8074	0.8580
ACAF-SMS	0.8597	**0.5245**	0.9265	0.9460
ACAF-SMV	**0.8602**	0.5244	**0.9266**	**0.9460**
	Natural Questions			
SMS	0.4725	0.4183	0.7159	0.7894
SMV	0.4636	0.4094	0.7118	0.7864
ACAF-SMS	**0.7563**	**0.5233**	**0.8654**	**0.9008**
ACAF-SMV	0.7474	0.5141	0.8587	0.8959

the answer confidence model cannot correctly distinguish the correct answer annotation, by jointly considering the contextualized sequence similarity, the **ACAF-SMS** methods can infer the correct answer annotation. The second example presents a negative case where the answer correctness evaluation model and the answer confidence method make the correct prediction. However, the **ACAF-SMS** method makes the wrong prediction because the answer annotation *Africa's bread basket* has maximum sequence similarity over all answer annotations.

4

Table 4.8: An example from NewsQA dataset. There are 7 different answer annotations for the question. Some of the answers are overlapped. For each answer we report its ranking scores with AC AF SMS ACAF-SMS.

context	The American photographed the remains of albatross chicks that had died from consuming plastic waste found in the surrounding oceans. According to the artist, not a single piece of plastic in any of the photographs was moved, placed or altered in any way. The nesting babies had been fed the plastic by their parents, who collected what looked to them like food to bring back to their young. From cigarette lighters to bottle caps, the plastic is found in what is now known as the great Pacific garbage patch that stretches across thousands of miles of the Pacific Ocean.				
Question	Plastic was found across thousands of miles of what				
Answer Candidates	great pacific garbage patch that stretches	0.0081	0.7406	0.0053	0.4904
	of	0.0837	0.7406	0.0453	0.4737
	of the pacific ocean.	0.7745	0.0898	0.3658	0.5306
	waste	0.9175	0.0025	0.4142	0.4457
	in the	0.0129	0.0017	0.0085	0.0091

Table 4.9: A positive example (top) and a negative example (bottom) from NewsQA dataset.

Context	Question	Answer	AC	AF	ACAF-SMS
Editor's note: Bryan Batt, who plays the closeted art director Salvatore Romano in the Emmy award-winning cable TV series "Mad Men," has acted in nine Broadway and nine Off-Broadway productions, such as "Sunset Boulevard," "Beauty and the Beast," "Jeffrey" and "Starlight Express." Batt, who is 45, has been acting for 23 years. He spoke to CNN.com about being an openly gay actor. "We have to work toward acceptance on all levels," says actor Bryan Batt, who is openly gay.	who is bryan batt?	plays the closeted art director Salvatore Romano	0.0681	0.0023	0.0533
		plays	0.0681	0.0023	0.0023
		"Mad Men,"	0.9317	0.0018	0.5939
		closeted art director Salvatore Romano in	0.0033	0.0004	0.9745
		plays the closeted art director Salvatore Romano in the Emmy award-winning	0.0047	0.0023	0.0053
		cable TV series "Mad Men," has acted in nine Broadway and nine Off-Broadway productions, such as "Sunset Boulevard," "Beauty and the Beast," "Jeffrey" and "Starlight Express."	0.0031	0.0018	0.0036
		an openly gay actor.	0.0015	0.9934	0.5455
Malnutrition has left this baby born in Zimbabwe fighting for her life. She is the face of an unfolding crisis in a country once known as Africa's bread basket.[...] But the World Health Organization (WHO) says the desperate situation has triggered a widening cholera outbreak that has killed 775 people and infected more than 15,000.	What is the outbreak part of?	unfolding crisis	0.9737	0.0009	0.4904
		cholera	0.9848	0.1766	0.4737
		Africa's bread basket.	0.9894	0.0283	0.5306
		the desperate situation	0.9938	0.1811	0.4456

4

4.6 Limitations

While many automatic answer aggregation methods take crowd worker's reliability into consideration [155, 175], to keep the proposed approach simple and concise, we focus on the influence of answer quality and ignore the worker reliability. Moreover, we only use NewsQA to evaluate the proposed method. Although it is possible to consider more real or simulated datasets, as shown by the experiments on SQuAD and Natural Questions in Section 4.5.6, NewsQA is the only large extractive QA dataset that provides all actual annotations to the best of our knowledge. Besides, this chapter assumes there is only one correct answer for each question, while it is possible that there are multiple correct answers in some applications.

4.7 Conclusion

In this chapter, we propose a novel answer annotation aggregation method for EQA crowdsourcing. We show that without any fine-tuning, our methods can achieve comparable performance with the trained QA and NLI model using *limited training data*. We introduce a novel algorithm for combining the NLI model, QA model, and contextual text embedding for answer text sequence aggregation. The experiments on a real large-scale crowdsourced EQA dataset show the effectiveness and stability of the proposed method. The proposed methods outperform the baseline single metric method by around 16% absolute improvement on P@1 and 10% improvement on other ranking metrics. For future work, we will further explore methods incorporating crowd worker reliability and question answerability for better answer aggregation. We will also explore the applicability of our approaches to other tasks that deal with collecting extractive texts [176, 177].

5

Unsupervised Domain Adaptation for Question Generation with Domain Data Selection and Self-training

This chapter focuses on domain adaptation for Automatic Question Generation (AQG) systems. The effectiveness of the trained AQG models can degrade significantly when applied to a different domain due to domain shift. In this chapter, we explore an unsupervised domain adaptation *approach to mitigate the lack of training data and domain shift issue with domain data selection and self-training. We first present a novel answer-aware strategy for domain data selection to select data with the most similarity to a new domain. The selected data are then used as* pseudo *in-domain data to retrain the AQG model. We then present generation confidence-guided self-training with two generation confidence modeling methods: (i) generated questions' perplexity and (ii) the fluency score. We test our approaches on three large public datasets with different domain similarities using a transformer-based pre-trained AQG model. The results show that our proposed approaches outperform the baselines and show the viability of unsupervised domain adaptation with answer-aware data selection and self-training on the AQG task. The code is available at https://github.com/zpeide/transfer_qg.*

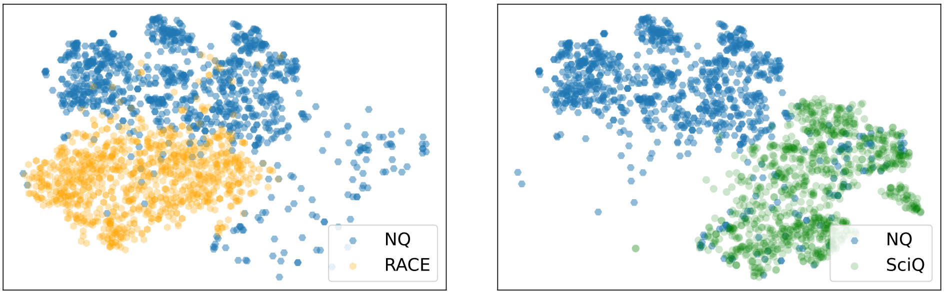
This chapter is based on the following conference paper: *Peide Zhu, Claudia Hauff. 2022. "Unsupervised Domain Adaptation for Question Generation with Domain Data Selection and Self-training", findings@NAACL'22.* [88]

5.1 Introduction

Automatic Question Generation (AQG) aims to generate questions from given text passages. It has been applied to a wide range of applications, such as question answering [24, 178], conversational systems [107], and education [9, 110]. Recently, pre-trained language models (LM) have significantly improved the state-of-the-art performance of various natural language processing tasks [34]. Pre-trained LMs have also substantially advanced the state-of-art performance on AQG [18, 35] by modeling AQG as a sequence-to-sequence task and fine-tuning on task-specific data.

However, with billions of parameters, the performance of these deep neural models heavily relies on the quantity and quality of available training data. As the manual process of creating high-quality questions is expensive in terms of time and money compared with abundant unlabeled data, the available data sources containing well-formed questions are insufficient, especially in the educational domain, where a lot of expertise is required to create questions geared towards human learning. To mitigate the lack of labeled training data, one solution is to pre-train models for AQG on a data-abundant labeled domain (source domain) and transfer the learned knowledge to the unlabeled target domain, which is known as *unsupervised domain adaptation* (UDA) [60]. It is a common challenge in machine learning research to learn knowledge in one domain and apply it in other domains with good generalization performance. One obstacle is the *domain shift* [61] between the source domain and the target domain which are assumed to be independent and identically distributed (i.i.d.), as illustrated in Figure 5.1. This limits the model's generalization and portability. To understand the effect of differences among domains on the performance of downstream AQG tasks, following previous research [179, 180], we perform a preliminary cross-domain study. We first train the AQG model on all domains separately and evaluate them across different domain test sets. As shown in Table 5.1, the model achieves the best performance on the test set from the same domain and degrades dramatically on test sets of other domains, which poses a great challenge to the transferring task. Based on these numbers, we argue that further research into domain adaptation methods for AQG is needed. There is a growing interest in applying unsupervised domain adaptation to tackle the domain shift issue in natural language processing tasks, such as question answering (QA) [181, 182], or neural machine translation (NMT) [179, 183, 184]. However, UDA is under-examined in the context of question generation. Unlike the QA task that can be modeled as a multi-label classification problem, AQG is a sequence generation problem, where it is hard to model the confidence or quality of generations [185]. Therefore, UDA methods for QA, like pseudo-label generation and filtering, cannot be directly extended to the AQG area. Moreover, data augmentation UDA methods for the NMT task, such as domain mixing [186], back-translation [187], or target sentences copying [188] are not directly applicable to AQG.

In this chapter, we propose a two-stage unsupervised domain adaptation approach for AQG to make use of the labeled source domain data and abundant unlabeled data. In the first stage, we focus on unsupervised domain data selection. Although the definition of "domain" in AQG is ambiguous, including the distribution of vocabulary, stylistic preferences, answer types etc, we first confirm that the learned BERT-based context paragraph representation can be used for robust domain data clustering as shown in Figure 5.1, and use Gaussian Mixture Models (GMMs) on the learned representations to find



(a) (b)

Figure 5.1: 2D visualization of average-pool BERT hidden states of data from different domains using t-SNE. **(a)** Datasets NQ and RACE. **(b)** NQ and SciQ.

Table 5.1: Impact of domain shift on AQG. Each row represents the METEOR score of the UniLM [18] model trained on one dataset (the row: NQ, SciQ and RACE) and tested on the test sets (the column).

Dataset	NQ	RACE	SciQ
NQ	**29.64**	13.76	14.32
RACE	16.59	**23.91**	12.37
SciQ	17.36	13.02	**29.47**

clusters, using methods proposed by Aharoni and Goldberg [180]. We perform domain data selection based on the distance of data examples to cluster centroids. To mitigate the gap of answer-type distributions, we further propose an answer-type aware data selection method (AADS) for pseudo-in-domain data selection. The selected pseudo-in-domain data are used to re-train the fine-tuned data to mitigate the domain shift.

In the second stage, we focus on self-training on the unlabeled target domain with the AQG model trained in the first stage. The self-training approach is substantially hindered by noisy and low-quality generated pseudo labels. We first propose a normalization method to avoid re-enforcing poorly generated questions. We also explore using sentence perplexity and fluency scores to model the confidence of sequence generation. We filter pseudo labels with low sequence confidence during self-training to prevent the model from being degraded by wrong or low-quality predictions.

We conduct experiments across three domains, including the Natural Question dataset as the source domain, RACE as one target domain of education, and SciQ as the target domain of science. Our results show our proposed approach is effective even when the target domain is substantially different from the source domain and outperforms several baselines including Latent Dirichlet Allocation (LDA) [189], BERT discriminator-based data selection [190], and unsupervised Gaussian mixture model(GMM) clustering on pre-trained language model features [180].

5

5.2 Background

In Chapter 2, we have introduced automatic question generation. In this section, we present a brief review of UDA and then discuss how our work differs from recent related research.

In many machine learning algorithms, samples in the training set and the test set are assumed to be independent and identically distributed (i.i.d.). When the underlying distributions do not match, the algorithms face the *domain shift* problem [191, 192], i.e., the *source* domain and the *target* domain data are not sampled from the same distribution. This issue happens in real-world scenarios, where labeled training data are scarce, while unlabeled data may be abundant since annotations are time-consuming and costly to acquire. It then translates into performance degradation. Unsupervised domain adaptation provides an elegant and scalable solution for mitigating this issue by learning only from unlabeled target data. In this chapter, we focus on the data-centric methods: data selection and self-training with pseudo-labeling [192].

There are abundant data samples in the source domain, but the importance of samples varies for adaptation. The data selection method [193] aims to select the data samples that are most related to the target domain. It is attracting more attention, thanks to the abundance of data and the large pre-trained models [194]. It has been studied for several NLP tasks, such as machine translation [180, 183], sequence classification [190, 195], and parsing [196]. Various domain similarity metrics have been investigated for data selection, such as Jensen-Shannon divergence on term distribution [196], perplexity [197], and cosine similarity [180]. Moreover, Aharoni and Goldberg [180] showed that sentence representation learned by pre-trained language models, e.g., BERT [34] and Roberta [198] is capable of clustering textual data to domains in an unsupervised way with high precision. In our work, we follow this research and perform domain clustering and selection with BERT.

Self-training is a bootstrapping method that has been used for domain adaptation in multiple NLP tasks [199–202]. The main idea of self-training [203] is to predict labels for unlabeled samples with a trained classifier as their 'pseudo' ground truth and use the synthetic data for further training.

Although the AQG approaches have made great strides in improving AQG effectiveness, they are trained and tested with data from the same dataset. When there is domain shift between training and test data, the model performance deteriorates considerably. Previous research, such as Liao and Koh [204], investigated transfer learning for AQG using supervised and semi-supervised adaptation methods but ignored the unsupervised setting. While some research [181, 205, 206] investigate the unsupervised iterative generation of synthetic QA pairs for question answering, they are not explicitly designed for AQG and do not demonstrate AQG performance. The most related recent work to ours is by Kulshreshtha *et al.* [207], who propose a new training protocol for UDA AQG. However, it requires unlabeled questions in the target domain, which is not always available, and we focus on investigating a more effective self-training method. We compare this chapter in Section 5.7.5. In this chapter, we close the gap by performing the answer-type aware domain data selection and self-training for mitigating the shift in source and target domain distributions.

5.3 Formalization

We now formulate the problem and present our notation. The data in the source domain with ground truth questions are denoted as $\mathscr{D}_s = \{(C^s, Q^s)\}$, while unlabeled data in the target domain is $\mathscr{D}_t = \{(C^t\}$; here, C is denoting the context (the passages, and answer spans used for generating questions). The question generation task is then to generate a sequence $\hat{Q}$ that maximizes the conditional probability of the prediction $\Pr(Q|C, \theta)$:

$$\begin{aligned}\hat{Q} &= \arg\max_{Q} \Pr(Q|C,\theta) \\ &= \arg\min_{Q} \sum_{t=1}^{T} -\log \Pr(Q_t|C,\theta,Q_{<t})\end{aligned} \tag{5.1}$$

where θ represents the parameters of the AQG model, which is initially learned from training data in the source domain. In our work, we aim to learn to adapt the θ from a source domain $\mathscr{D}_S$ to the target domain $\mathscr{D}_T$ and achieve optimal performance.

5

5.4 Domains

5.4.1 Source Domain

We use the question answering corpus Natural Questions (`NQ`) [56] as our source domain. It consists of aggregated questions issued to the Google search engine and answers annotated by crowd-workers from the most related Wikipedia pages. It consists of a large amount of unique passages and covers a range of topics, which makes it a good source domain for transferring. As there are many examples in `NQ` with tables as context, to use this dataset forAQG, we select a subset that contains 89,453 samples in the training set and 3,726 samples in the test set from the original `NQ` dataset.

5.4.2 Target Domains

Education

The first target domain we choose is education, for which we use the `RACE` [51] dataset. `RACE` is a large dataset consisting of questions, answers, and associated passages used in English exams designed for middle-school and high-school Chinese students. Questions in `RACE` are designed by instructors (i.e., domain experts) for evaluating students' reading comprehension ability. There are three types of questions: cloze, general and specific. Following the practice of EQG-RACE [208], we only keep the specific questions. For unsupervised AQG, we use 18.6K data examples in the training set. The original dev and test sets are used for evaluation.

Science

Our second target domain is science, where we make use of the `SciQ` [156] dataset. `SciQ` consists of 13.7K crowdsourced multiple-choice science exam questions, including 11.7K questions in the training set and 1K for the dev and test set each. Each `SciQ` question has an associated passage, the correct answer, and the distractors. The `SciQ` passages are chosen from science study textbooks of different topics, including biology, chemistry, earth science, and physics. For unsupervisedAQG, we utilize the support passages in the training

Table 5.2: Overview of the source domain dataset NQ, and the selected datasets for target domains SciQ and RACE.

Features	NQ	SciQ	RACE
Question	Search Logs	Crowdsourced	Experts
Context	Wikipedia	Textbook	Examinations
Train set	89,453	11,679	18,614
Test set	3,726	1,000	1036
#W/doc	106.27	78.05	318
#Sent./doc	4.43	4.84	17
#W/Sent.	26.81	16.13	17.96
#W/Q	10.20	14.31	10.8

set without questions as unlabeled data; we use the original dev and test sets for AQG evaluation.

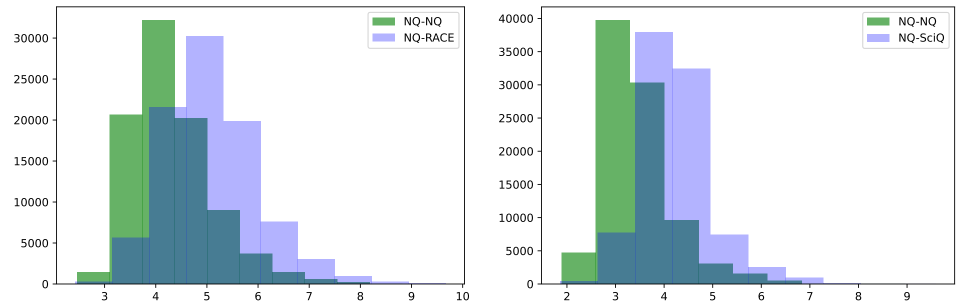
Table 5.2 lists the basic statistics of our three datasets. On those datasets, we can make a thorough evaluation of the AQG model's transfer performance and the effectiveness of the proposed approach.

5.5 Domain Data Selection

Not all data are equally important or useful for domain adaptation. Irrelevant data samples can add noise and affect the learned model's performance and robustness towards cross-domain application considerably [209]. A solution to reduce the impact of irrelevant data is domain data selection, i.e., to retrieve the most appropriate data from the source domain data given the target domain data. Most proposed domain data selection approaches consider ranking training examples from $\mathscr{D}_S$ according to a domain similarity measure and select the top-n examples that are closest to $\mathscr{D}_T$.

We encode the context passage at the paragraph level with BERT, and perform average pooling of the last layer hidden state of each token to create its vector representation. To show that this is a robust representation for mapping sentences to domains in an unsupervised, data-driven approach, we first visualize them with t-SNE, as shown in Figure 5.1. We can observe the encoding vector representation with BERT indeed can cluster data examples to domains. Following the practice of Aharoni and Goldberg [180], we then perform unsupervised clustering by fitting Gaussian Mixture Models (GMMs) to the vector context representations with k predefined clusters. We assign each cluster the domain class by measuring its purity (proportion of examples belonging to each domain). We use the Euclidean distance [210] of each example to cluster center as the measure of domain distance. Figure 5.2 shows the distribution of NQ dataset examples' distance to NQ's, RACE's and SciQ's domain center, respectively. We sort source data examples based on their distance to the target domain center and select data examples with the most domain similarity as the pseudo-in-domain data.

Table 5.3 shows the unsupervised domain clustering results. We compare the proposed methods with Latent Dirichlet Allocation-based (LDA) clustering [189]. We also compare different ways of creating paragraph vector representations, including using BERT [CLS]



(a) (b)

Figure 5.2: Distribution of the distance between each data example to domain cluster center. **(a)** NQ and RACE. **(b)** NQ and SciQ.

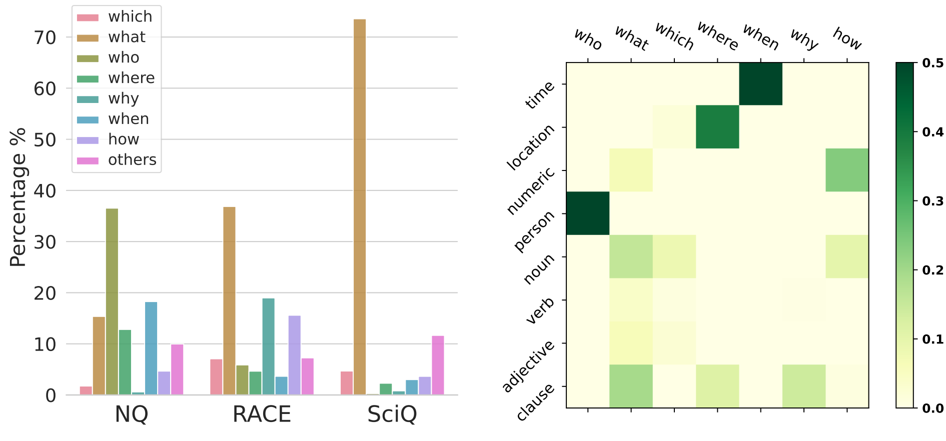
Table 5.3: The performance of different unsupervised domain clustering methods on the RACE and SciQ datasets.

	RACE			SciQ		
Method	Acc	F1	Rc	Acc	F1	Rc
LDA	0.79	0.76	0.72	0.69	0.61	0.55
KM$_{CLS}$	0.37	0.35	0.98	0.33	0.25	0.97
KM$_{All}$	0.94	0.85	**0.99**	0.88	0.63	0.89
KM$_{Last}$	0.97	0.91	0.97	0.91	0.72	0.99
GMM$_{CLS}$	0.42	0.36	0.97	0.37	0.26	0.94
GMM$_{All}$	0.96	0.90	0.95	0.88	0.64	0.89
GMM$_{Last}$	**0.98**	**0.95**	0.96	**0.91**	**0.72**	**0.99**

token encoding ($_{CLS}$), average pooling of all BERT layer hidden states ($_{All}$), and average pooling of the last hidden states ($_{Last}$). Besides GMM clustering methods, we also compare the GMM method with K-Means (KM). To accelerate the clustering, we perform PCA over the paragraph representation first. Our results show the GMM method with the pooling average of the last BERT hidden states to outperform the other methods.

5.5.1 Answer-Type Aware Data Selection

The question type distributions vary a lot for different application domains, as shown in Figure 5.3a. For example, in NQ, the ‘who’ questions account for over 35% of all questions, but in SciQ, 73.6% of questions have the ‘what’ type. Traditional data selection methods are based only on the similarity of context passages, which may suffer from unbalanced target label sampling. As there are no questions available in the target domain, it is a challenge to perform data selection according to the distribution of target question types. We first investigate the correlation between the answer types and question types. The question types are identified by the interrogative ‘w’-word, such as ‘who’, ‘what’, etc. We



(a) (b)

Figure 5.3: **(a)** The distributions of question types of the Natural Questions (NQ) dataset, the RACE dataset, and the SciQ dataset. **(b)** The correlation between answer types (including named entity types like time, location, numeric, and person and POS taggers like noun, verb, adjective, and clause and question types.

identify the answer types such as 'time', 'location', etc. using the spacy[1] NER and part-of-speech (POS) tagger. The correlation matrix (expressed in Pearson correlation coefficient) is shown in Figure 5.3b. We find question types and answer types are strongly correlated to each other. For example, the correlation coefficient between 'time' and 'when' is 0.67, and between 'person' and 'who' is 0.63. Thus, we propose a heuristic answer-type aware data selection strategy for domain data selection from the source domain with a similar answer type distribution in order to mitigate the label divergence. Specifically, we first group the data by answer types, and then conduct data selection on each group.

5.6 Self-Training

When training the AQG model with pseudo-labels, it is natural to put more emphasis on the labels that the model is more confident about. An intuitive solution is to weigh each pseudo-token according to its estimated probability in order to avoid re-enforcing poor predictions. Thus, we propose the following normalized training objective for self-training:

$$\hat{Q} = \arg\min_{Q} \sum_{t=1}^{T} -\log \alpha_t \Pr(Q^{'}_t | C, \theta, Q^{'}_{<t}) \tag{5.2}$$

where $Q^{'}$ is the pseudo-label, and α_t is the predicted probability of the t-th word $Q^{'}_t$, and T is the length of the pseudo-label.

We apply the AQG model to generate questions on unlabeled target-domain data, which are then used as 'pseudo' gold labels for further training. The self-training approach is substantially hindered by noisy, low-quality labels. How to deal with noisy pseudo la-

[1]https://spacy.io/

bels is crucial to the final UDA effectiveness. Classical pseudo label generation methods [211–213] filter generated labels by their 'confidence', i.e., the predicted probability of the label in those classification tasks. How to represent the confidence of sequence generation in pseudo-labeling is insufficiently explored. Traditionally, confidence estimation has been defined as a task of assessing the quality of the whole sequence of words in the target sentence. Therefore, we propose a question quality guided pseudo labeling method to address this problem, with two confidence metrics: (i) the sentence perplexity and (ii) the BERT-based fluency score.

Sentence Perplexity

The first metric is the perplexity of the generated questions. The generation with higher confidence should have lower perplexity. Here, perplexity (PPL) is defined as follows:

$$\mathrm{PPL}(Q) = 2^{-\frac{1}{T}\log\prod_{1}^{T}\Pr(Q_t|Q_{<t})} \tag{5.3}$$

BERT-based Fluency Score

For our second metric, we use *fluency* as the question quality metric, which indicates whether the generation follows grammar rules and correct logic. The perplexity of a sentence under a well-trained language model usually serves as a good indicator of its fluency [115]. We use a fine-tuned BERT language model as the evaluator. The fluency metric $R_{fluency}$ for question Q is calculated as follows:

$$R_{fluency}(Q) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log\mathrm{BERT}(Q_t|Q_{<t})\right). \tag{5.4}$$

During the unsupervised self-training, after each epoch, we perform beam search with the trained model, and the generated questions are ranked according to their fluency score. Only questions with confidence metrics better than the threshold ϕ and PPL are selected as pseudo-labels. If one data sample got selected in the last epoch, but its generated question's confidence metric in the current epoch is not higher than before, it is removed. In this way, only questions of high quality that improve over time are chosen for training. Algorithm 2 demonstrates the pseudo-code of the self-training algorithm.

5

Algorithm 2: Self-training for AQG

Input : Target domain data: $\mathscr{D}_t = \{C^t\}$. AQG model $\mathscr{M}_{QG}$ with parameters θ
repeat
  for $C^t \in \mathscr{D}_t$ **do**
    $[Q_t^{'}, \alpha_t]_1^T = \mathscr{M}_{QG}(C)$
    if *Use Fluency Score* **then**
      $f = \exp(-\frac{1}{T}\sum_{t=1}^{T} \log BERT(Q_t^{'}|Q_{<t}^{'}))$
    else if *Use PPL* **then**
      $f = 2^{-\frac{1}{T}\log\prod_1^T \Pr(Q_t|Q_{<t})}$
    if $f > \phi$ **then**
      $\mathscr{L} = \mathscr{L} + \sum_{t=1}^{T} -\log \alpha_t \Pr(Q_t^{'}|C, \theta, Q_{<t}^{'})$
    end
  end
  $\theta \longleftarrow \mathrm{Adam}(\nabla_\theta \mathscr{L})$.
until *Convergence or Reach Maximum Epochs*;

5.7 Experiments

In this section, we describe the model and the training regime in more detail.

5.7.1 Experimental Settings

QG Model

We use the state-of-the-art pre-trained transformer-based sequence-to-sequence natural language understanding and generating model **UniLM** [18] for question generation. Specifically, we choose the uncased pre-trained `unilm1.2-base-uncased` model for fine-tuning. It has 12 transformer layers and is jointly pre-trained on large amounts of text, optimized for bidirectional, unidirectional, and sequence-to-sequence language model objectives. We use the `s2s-ft` package[2] for fine-tuning. To fine-tune our model, the input context passage, the answer, and the generated question are combined together into a sequence: "`[CLS]` *context passage*`[EOS]` *answer span* `[EOS]` *question* `[EOS]`". Both the input passage and answer are regarded as the first text segment, while the generated question is the second segment in the unified LM.

Training Details

The model is trained on a server consisting of 4 GeForce GTX 1080 GPUs with a batch size of 32, a mask probability of 0.8, and a label smoothing rate of 0.1. The max_source_seq_length is set to 464; the max_target_seq_length is 48. We first fine-tune UniLM with the `NQ` dataset for ten epochs. We use the Adam optimizer with $\epsilon = 1e-8$, learning rate is $1e-4$ with 500 warmup steps.

[2] `https://github.com/microsoft/unilm/tree/master/s2s-ft`

Table 5.4: Results of unsupervised domain adaptation for AQG with answer-type aware (**AA-**) domain data selection(**DDS**) and self-training(ST) on RACE and SciQ test set. We compare three baseline methods: LDA [189], BERT-DDS [190] , GMM [180]. ⋆ denotes the best results for DDS, ♡ denotes best results for ST, and ♣ denotes best results for DDS+ST.

		RACE				SciQ			
	Method	**B-1**	**B-4**	**MT**	**RG**	**B-1**	**B-4**	**MT**	**RG**
	None	21.99	4.11	13.68	21.31	25.94	8.67	15.53	26.59
DDS	**random**	21.91	4.02	13.74	21.26	26.15	8.97	15.56	26.62
	LDA	21.97	4.29	13.72	21.47	26.57	8.88	15.67	27.07
	BERT-DDS	22.06	3.99	13.61	21.30	26.43	9.08	15.70	26.70
	KMeans	22.21	4.45	13.75	21.65	26.45	**9.23**⋆	15.72	**27.15**⋆
	GMM	22.38	4.58	14.05	21.70	26.51	9.08	**15.79**⋆	27.05
	AA-KMeans	22.28	4.40	13.92	21.71	26.26	8.85	15.66	26.82
	AA-GMM	**22.79**⋆	**4.79**⋆	**14.23**⋆	**22.15**⋆	**26.61**⋆	9.09	15.73	26.90
ST	**w/o-Norm**	23.34	4.82	14.45	22.89	27.89	10.37	16.51	28.26
	w/o-Filter	23.83	5.13	14.65	23.06	28.29	10.85	16.95	28.86
	Fluency	24.20	5.11	14.74	23.66	28.22	10.76	16.92	28.92
	PPL	**24.38**♡	**5.22**♡	**14.85**♡	23.43	28.30	11.04	17.12	29.03
	Fluency&PPL	24.32	5.14	14.73	**23.52**♡	**28.30**♡	**11.04**♡	**17.12**♡	**29.03**♡
DDS +ST	**w/o-Filter**	23.43	4.93	14.43	22.78	28.21	11.00	16.90	28.93
	Fluency	24.20	4.85	14.67	23.13	28.82	11.05	16.86	28.94
	PPL	24.43	**5.40**♣	15.08	23.49	29.12	11.04	16.92	29.38
	Fluency&PPL	24.71	5.20	14.96	23.78	**29.40**♣	11.23	17.13	**29.52**♣
	AA-Fluency	24.14	5.17	14.79	23.07	28.10	10.82	16.69	28.54
	AA-PPL	24.50	5.14	**15.09**♣	23.60	28.84	11.65	**17.22**♣	29.30
	AA-Fluency&PPL	**24.71**♣	5.16	14.87	**23.80**♣	28.68	**11.70**♣	17.17	29.36

5

Table 5.5: In-domain test results of the AQG model (fine-tuned and tested on the same dataset).

Dataset	B-1	B-4	MT	RG
NQ	60.05	30.31	29.64	59.26
SciQ	46.99	33.22	29.47	42.73
RACE	37.86	17.90	23.91	37.56

Unsupervised Domain Data Clustering
We use 4,500 examples randomly selected from NQ, SciQ, and RACE for unsupervised data clustering. We set the number of clusters as 2 since we intend to investigate the separability between the source domain and the target domain.

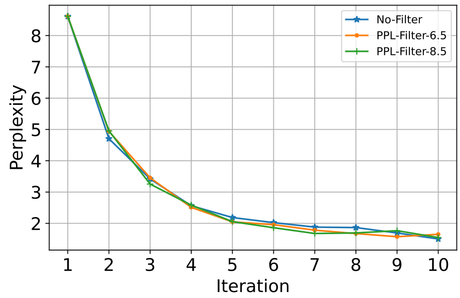
Evaluation Metrics
We compare the model performance along three automatic evaluation metrics: **BLEU** [66], which is computed with the geometric average of the modified n-gram precision and the brevity penalty; **Meteor** [68], which compares the generation with the gold question in terms of exact, stem, synonym, and paraphrase matches; and **Rouge-L** [67], which measures the shared longest common sub-sequence. We calculate these metrics with the package released by Du *et al.* [13]. We also conduct a human evaluation.As a sanity check and to evaluate the AQG model's ability to generate questions based on these datasets, we first conduct in-domain tests on these three datasets separately, i.e., we fine-tune and test the model on the training/test set from the same dataset. As shown in Table 5.5, we achieve results comparable with state-of-art for the NQ, RACE and SciQ datasets.

5

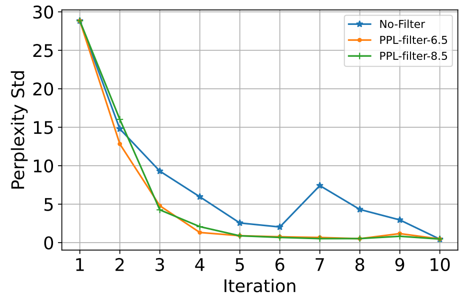
5.7.2 Experiments on Data Selection

In this experiment, we compare the proposed answer-type aware data selection with several baselines. We train the AQG model with the selected data and evaluate the data selection method by comparing its performance. The first baseline is random data selection (**random**). With this baseline, we randomly sample 1,000 samples from NQ. The second baseline is LDA-based clustering [189]. We use the gensim [214] LDA implementation for this baseline. The third method (**BERT-DDS**) is proposed by Ma *et al.* [190], where a BERT-based domain discriminator is used for data selection. The discriminator is first trained with randomly sampled data from the datasets. The baseline model achieved 99.85% for RACE and 92.35% accuracy for the SciQ dataset. The last baseline method we compare is adopted from the unsupervised domain clustering method (**GMM**) proposed by Aharoni and Goldberg [180], as described in Section 5.5. We use the BERT-base model implementation of huggingface transformers [157] to get the context passage encoding. In addition to GMM, we also compare the K-Means method [215]. The results are presented in Table 5.4.

Impact of Domain Data Selection
Re-training with randomly selected data does not improve our model's generalization performance. All other data selection methods outperform random data selection except BERT-DDS. One reason is that BERT-DDS training needs sampling data from different domains; its performance relies on the sampled data and also label examples that are sim-



(a)



(b)

Figure 5.4: Change of **(a)** average perplexity, and **(b)** standard deviation of generations along iterations.

ilar to the target domain as the source domain. *Data selection with unsupervised domain clustering with BERT context encoding outperforms other methods, which confirms its effectiveness.*

On the RACE dataset, answer-type aware data selection with K-Means (**AA-KMeans**) and GMM (**AA-GMM**) outperform the same selection method without answer-type awareness. We note this result does not always hold for the SciQ dataset. One possible reason is due to the extremely unbalanced answer type distribution in SciQ: we have to select examples with generally low domain similarities w.r.t. the source domain to create identical answer-type distributions.

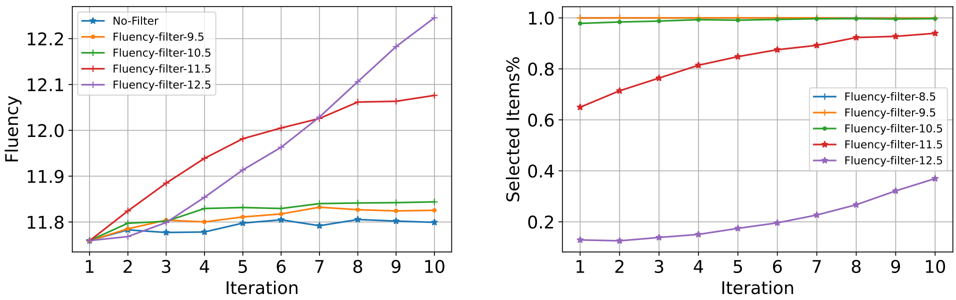
5.7.3 Experiments on Self-Training

We conduct self-training with the target-domain unlabeled data on the AQG model fine-tuned on the NQ dataset. We first verify the effectiveness of the proposed normalized training objective. As the results show in Table 5.4, self-training with normalization (**w/o-Filter**) outperforms self-training without any confidence filtering and normalization (**w/o-Norm**), which indicates its effectiveness.

Impact of Generation Confidence Guided Self-training

We explore two generation confidence metrics for self-training, the sentence perplexity and the question fluency score. To train the BERT LM for generating fluency scores for question quality evaluation, we combine all questions from NQ and the Quora Question Pairs dataset[3], creating a dataset consisting of 834,834 questions. The final model achieves a perplexity of 9.27 on the evaluation set. As the results in the **ST** part of Table 5.4 show, both proposed generation confidence metrics improve the performance considerably up to 6%. This can be explained by removing low-quality and noisy data, which hinders model training. As Figure 5.4 shows, with perplexity filtering—although the changing curves of mean perplexity of the generated pseudo-labels in each iteration are similar—the standard deviation drops faster and more steady. As Figure 5.5 shows, the average fluency score improves along iterations even without fluency filtering, but with fluency filtering,

[3]https://www.kaggle.com/c/quora-question-pairs



(a) (b)

Figure 5.5: Change of **(a)** average fluency score, and **(b)** the percentage of generated questions whose fluency score is higher than ϕ along iterations.

Table 5.6: Influence of the fluency threshold (ϕ).

	ϕ	B-1	B-4	MT	RG
RACE	8.5	23.83	**5.12**	14.65	23.06
	9.5	24.20	5.11	**14.74**	**23.66**
	10.5	**24.23**	5.06	14.68	23.29
	11.5	23.93	5.10	14.46	23.09
	12.5	23.78	4.55	14.41	23.05

the average fluency score improves more steadily and increases towards the threshold value ϕ. The proportion of questions with higher fluency scores than ϕ increases along iterations. As reflected in Figure 5.5b and Table 5.6, if the threshold value is too low, fewer noisy pseudo examples can be filtered out. If the threshold is too high, there would be less supervision for the AQG model. Both of these settings would lead to performance degradation.

Impact of Joining Domain Data Selection and Self-Training

We also conduct domain adaptation by joining domain data selection and self-training (**DDS+ST**). As shown in Table 5.4, joining DDS and self-training without filtering does not show performance improvement on both datasets, which implies with DDS, pseudo-labels during self-training may be noisier. With the proposed filtering with fluency score or question perplexity, the joint method outperforms DDS and self-training. On the RACE dataset, the answer-type aware joint method generally achieves the best performance across all evaluation metrics.

5.7.4 Human Evaluation

In addition to the automatic evaluation results shown in Table 5.4, we also report on our human evaluation in Table 5.7. We randomly sampled 50 generated questions from the RACE and SciQ test sets, respectively, and asked three domain experts (both male and fe-

Table 5.7: Human evaluation (mean and standard deviation) on `RACE` and `SciQ` datasets. Syntax and Relevance evaluation adopt a 3-point scale. Higher is better; Answerability is boolean type (0-1).

	Syntax	**Relevance**	**Answerability**
Method	`RACE`		
w/o-UDA	2.60 (0.66)	2.00 (0.78)	0.43 (0.49)
ST	2.78 (0.51)	2.12(0.73)	0.46 (0.50)
DDS+ST	**2.81(0.47)**	**2.12 (0.75)**	**0.51 (0.50)**
	`SciQ`		
w/o-UDA	2.83 (0.40)	2.40 (0.64)	0.57 (0.50)
ST	**2.94 (0.26)**	2.49 (0.64)	**0.67 (0.47)**
DDS+ST	2.92 (0.27)	**2.53 (0.63)**	0.67 (0.47)

Table 5.8: Unsupervised domain adaptation results on MLQuestions dataset.

Dataset	**B-1**	**B-4**	**MT**	**RG**
w/o-UDA	30.06	7.96	18.62	31.60
DDS	29.89	8.27	18.63	31.64
ST	32.58	9.41	19.41	34.20
DDS+ST	**34.76**	**10.57**	**20.41**	**37.02**
Net Gain	**4.7↑**	**2.61↑**	**1.79↑**	**5.42↑**

male, ages ranging from 25 to 35) to rate the generated questions by the AQG model without UDA (w/o-UDA), with self-training (ST), and self-training and domain data selection(DDS+ST). The experts are also presented with the context paragraphs and the answers, as shown in Figure 5.6. The generated questions are shown in Table 5.10 and Table 5.11. We rate questions along three dimensions: (i) syntax, (ii) relevance, and (iii) answerability. We evaluate the syntax correctness on a 3-point scale: score 1 for significant syntax issues, score 2 for minor issues, and score 3 for the question is syntactically correct. We evaluate the relevance, i.e., whether the question is relevant to the context and the answer on a 3-point scale: score 1 for irrelevance, score 2 for partial relevance, and score three for meaning entirely relevant. In contrast, we regard the answerability as a boolean-type value, indicating whether the question can be answered given the context and answer. As the results show, all AQG with UDA methods outperform the AQG model without domain adaptation. On the `RACE` dataset, the proposed unsupervised domain adaptation for AQG with data selection and self-training (DDS+ST) achieves the best performance along with all metrics; although the performance of UDA with self-training only outperforms DDS+ST slightly in terms of syntax and answerability, DDS+ST outperforms self-training.

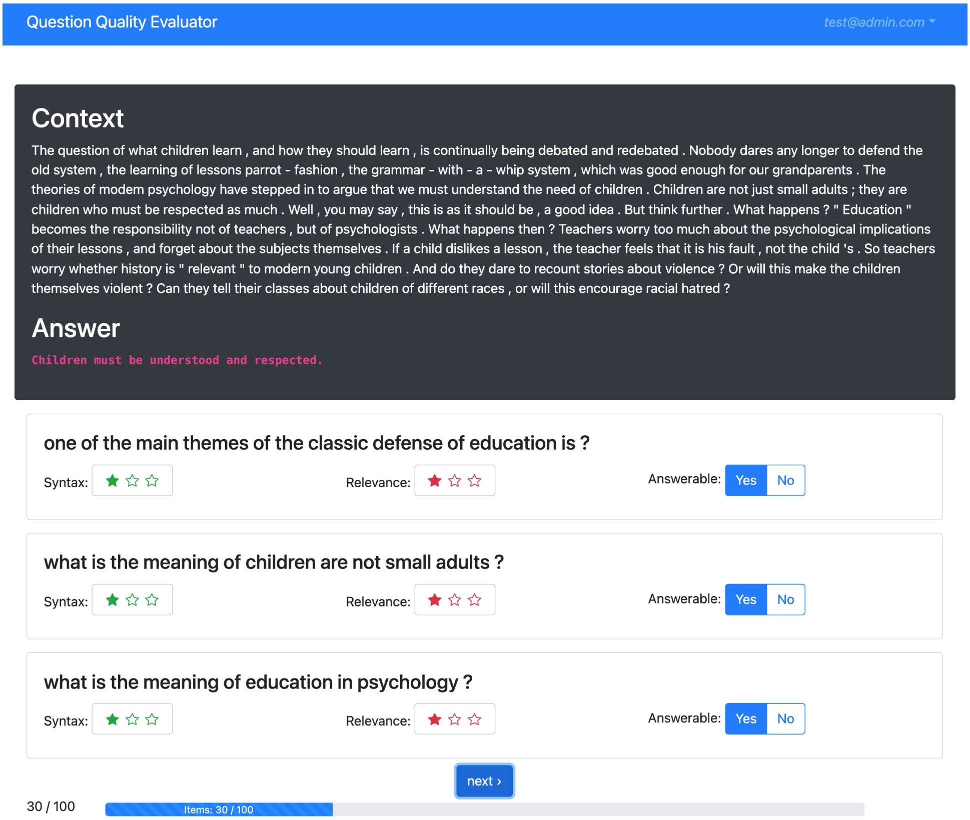


Figure 5.6: The interface for human annotation. We display the context and the ground truth answer to the annotators. Questions generated by different methods are displayed in a random order.

5.7.5 Experiments on MLQuestions

In addition to experiments on RACE and SciQ datasets, we also conduct unsupervised domain adaptation experiments on MLQuestions [207].

We first conduct unsupervised domain data selection with GMM_{last} method and present the confusion matrix in Figure 5.7 and select 1,000 data examples from NQ that are closest to MLQuestions clustering center. We set the number of clusters as two because we want to directly investigate the unsupervised separability between NQ and MLQuestions. We use the provided development set and the test set of MLQuestions. Then, we perform domain adaptation for AQG and show results in Table 5.5. Compared with the self-training method explored in [207], the proposed method in this chapter achieves more performance increase, e.g., **DDS+ST** method achieved 5.42 and 4.7 net gain in Rouge-L and BLEU-1 score respectively, compared with 0.58 and 0.80 net gain with self-training in [207]. In this chapter, we focus on the self-training method, so we consider conducting open-domain retrieval-based methods like **Back-Training** in future research.

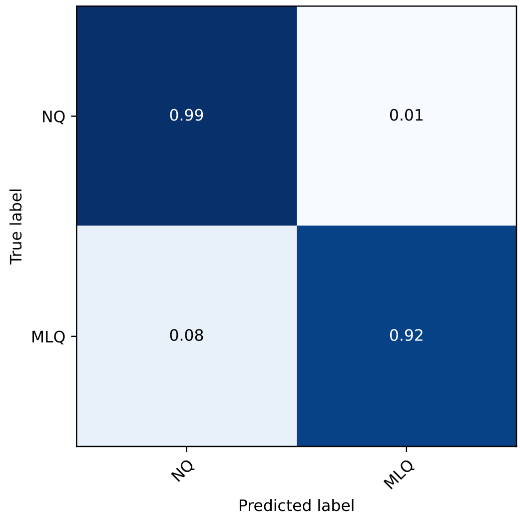


Figure 5.7: Confusion matrix for unsupervised domain data clustering results on MLQuestions and NQ datasets. We use 3,000 data examples from NQ and MLQuestions each.

5

5.8 Case Study

5.8.1 Examples Selected Data

Table 5.9 illustrates several data examples of selected from NQ dataset that are most similar to *education* domain, i.e. the RACE dataset, and to *science* domain, i.e. SciQ dataset using the GMM_{last} BERT based domain data selection method. The RACE dataset is a large dataset of English exams for middle-school and high-school Chinese students. Its vocabulary is middle-school and high-school level. Many passages in it are story-style. As NQ⟶RACE data examples show, the selected data from NQ are close to SciQ in terms of both the vocabulary and text style. Meanwhile, SciQ passages are chosen from science study textbooks of different topics, including biology, chemistry, earth science, and physics. The examples of selected data (NQ⟶RACE) can be categorized into the biology domain, which includes a lot of the biology terms, elucidating biological processes. These examples show the effectiveness of the data selection method.

Table 5.9: Examples of selected data from NQ dataset that are most similar to RACE dataset (NQ⟶RACE) and SciQ dataset (NQ⟶RACE).

NQ ⟶ RACE	NQ ⟶ SciQ
To expand the number of women smokers Hill decided to hire Edward Bernays, who today is known as the father of public relations, to help him recruit women smokers. Bernays decided to attempt to eliminate the social taboo against women smoking in public. ...The targeting of women in tobacco advertising led to higher rates of smoking among women. In 1923 women only purchased 5% of cigarettes sold, in 1929 that percentage increased to 12%, in 1935 to 18.1%, peaking in 1965 at 33.3%, and remaining at this level until 1977.	The lysosomes also act as the waste disposal system of the cell by digesting unwanted materials in the cytoplasm, both from outside the cell and obsolete components inside the cell. Material from outside the cell is taken - up through endocytosis, while material from the inside of the cell is digested through autophagy. Their sizes can be very different. They were discovered and named by Belgian biologist Christian de Duve, who eventually received the Nobel Prize in Physiology or Medicine in 1974.
A man named Bailey intends to take his family from Georgia to Florida for a summer vacation , but his mother , (referred to as "the grandmother" in the story) wants him to drive to East Tennessee , where the grandmother has friends ("connections"). She argues that his children, John Wesley and June Star, have never been to East Tennessee, and she shows him a news article in the Atlanta Journal Constitution ...He and the grandmother agree that things were much better in the past and that the world at present is degenerate; she concurs with Sammy's remark that "a good man is hard to find."	Decomposition is the process by which organic substances are broken down into simpler matter. The process is a part of nutrient cycle and is essential for recycling the finite matter that occupies physical space in the biosphere. Bodies of living organisms begin to decompose shortly after death. Animals, such as worms, also help decompose the organic materials. Organisms that do this are known as decomposers. Although no two organisms decompose in the same way, they all undergo the same sequential stages of decomposition. The science which studies decomposition is generally referred to as taphonomy from the Greek word taphos, meaning tomb.
The next day, just before Lincoln and Sara board a boat to escape to the Dominican Republic, Sucre gives Sara the $100,000 they stole from the General, apologizing for not being able to wire the money to them the night before as planned. Mahone gives Sara the paper Michael asked him to deliver, ..., but don't ever, say. He then says what he wants to say is that he loves them both, very much. He tells them to make sure his child is told every day how much he is loved and how lucky he is to be free. The video, and the entire series	An elater is a cell (or structure attached to a cell) that is hygroscopic, and therefore will change shape in response to changes in moisture in the environment. Elaters come in a variety of forms, but are always associated with plant spores. In many plants that do not have seeds, they function in dispersing the spores to a new location. Mosses do not have elaters, but peristome which also change shape with changes in humidity or moisture to allow for a gradual release of spores

5

Table 5.10: Examples of generated questions on the RACE dataset with different methods.

	RACE	RACE
Context	Jenny was a pretty five-year-old girl. One day when she and her mother were checking out at the grocery store, Jenny saw a plastic pearl necklace priced at $2.50. Her mother bought the necklace for her on condition that she had to do some homework to pay it off. Jenny agreed. She worked very hard every day, and soon Jenny paid off the necklace. Jenny loved it so much that she wore it everywhere except when she was in the shower. Her mother had told her it would turn her neck green! Jenny had a very loving daddy. When Jenny went to bed, he would read Jenny her favorite story. One night when he finished the story, he said, "Jenny, could you give me your necklace?" "Oh! Daddy, not my necklace!" Jenny said." But you can have Rosy, my favorite doll. Remember her?	Lawmakers in the United States have expanded an investigation into the use of location-tracking systems on mobile devices. The action follows recent reports about the storing of information on the Apple iPhone. Some people consider location tracking to be a threat to personal privacy and security. Allan Friedman, the research director, says, "All wireless companies do some location tracking as part of their networks. This information is usually stored by the companies, not the devices, and there are laws to protect it. Law enforcement agencies, ...There's the idea that because it's on my phone and on my computer, rogue applications that I pay for or that I'm tricked into downloading may be able to access this data and somehow misuse it." Apple says it is "not tracking the location of your iPhone". It is simply keeping a database of Wi-Fi hotspots and cell phone towers near the user's location.
Answer	**She had to help her mother do some housework.**	**Because it is thought to threaten users' privacy.**
w/o-UDA	what's the meaning of the name jenny?	why is there a tracking system on my phone?
ST	what is the name of jenny's necklace?	why is there a location tracking system on apple?
DDS+ST	**how did jenny get her necklace in the movie?**	**why do we not use location tracking on iphone?**

5

5

Table 5.11: Examples of generated questions on the `SciQ` dataset with different methods.

	`SciQ`	`SciQ`
Context	Gamma rays are produced when radioactive elements decay. Radioactive elements are elements with unstable nuclei. To become more stable, the nuclei undergo radioactive decay. In this process, the nuclei give off energy and may also emit charged particles of matter. Types of radioactive decay include alpha, beta, and gamma decay. In alpha and beta decay, both particles and energy are emitted. In gamma decay, only energy, in the form of gamma rays, is emitted.	Not all wetlands are alike, as you can see below (Figure below). Wetlands vary in how wet they are and how much of the year they are soaked. Wetlands also vary in the kinds of plants that live in them. This depends mostly on the climate where the wetland is found. Types of wetlands include marshes, swamps, and bogs.
Answer	**radioactive**	**wetland**
w/o-UDA	where do gamma rays come from when they decay?	what do you call a place that is covered with water?
ST	what type of element is the source of gamma rays?	what do you call marshes that are wet?
DDS+ST	**what type of elements give off gamma rays?**	**what are marshes and bogs called?**

5.9 Conclusion

We proposed an unsupervised domain adaptation approach for question generation. Our approach includes an answer-type aware unsupervised domain data selection method and a sequence generation confidence-guided self-training algorithm. We conduct experiments on three domains. We use the Natural Questions dataset as the labeled source domain, `RACE` as the target education domain, and `SciQ` as the target science domain. Our results suggest our approach is effective for this application setting. We find that it significantly improves domain adaptation performance of our AQG model.

6

On the Effects of Automatically Generated Adjunct Questions for Search as Learning

In this chapter, we focus on the impacts of automatically generated questions on learners SAL scenario. Actively engaging learners with learning materials has been considered necessary, especially in the SAL setting. One active reading strategy relies on asking so-called adjunct questions, *i.e., manually curated questions geared towards essential concepts of the target material. However, manual question creation is impractical, given the vast online content. Recent research has explored the effects of AQG on prompting human learning. These studies have primarily focused on user studies in controlled online reading scenarios with limited documents. However, the impacts of adjunct questions on learning in the SAL setting, which involves learning through web searching, are not yet well understood. This chapter addresses this gap by conducting a user study with automatically generated adjunct questions integrated into the reading interface built on top of a search system. We conduct a between-subjects user study (N = 144) to investigate the incorporation of automatically generated adjunct questions on participants' learning by employing three different generation strategies as well as a control condition: (i) synthesis questions; (ii) factoid questions targeting random text spans; and (iii) factoid questions targeting terms and phrases relevant to the information need at hand.*

This chapter is based on the following conference paper: *Peide Zhu, Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. 2024. "On the Effects of Automatically Generated Adjunct Questions for Search as Learning", to appear@CHIIR'24.*

6.1 Introduction

Searching and reading online materials has become a crucial way of learning. However, it is generally considered inefficient to learn by passively browsing and reading documents [216, 217]. In contrast, actively engaging learners during this process with retrieval practice methods like *adjunct questions*— i.e., asking questions about specific parts of a document to draw attention to the reading materials [217] and retrieving information from one's memory [218, 219]—leads to better learning outcomes. Research on the effects of asking such questions has been extensively studied and generally been found to have a positive effect on learning. However, these studies have been conducted in controlled classroom settings [3, 220, 221] and with *manually* curated questions.

Considering the amount of content available on the web, this is not a feasible approach in the Search as Learning (SAL) setting, where learning behaviour commonly involves searching over open-domain resources, targeting complex concepts instead of fact-finding, and learning by reading and integrating knowledge across documents [93, 222–224].

With the ever-improving generation quality of PLM, some works have analyzed the effectiveness and potential benefits of Automatic Question Generation (AQG) on human learning. For example, Syed *et al.* [80], Van Campenhout *et al.* [84] demonstrated that automatically generated questions performed comparably to human-authored questions. Moreover, some works [80, 83] highlighted the potential importance of incorporating automatically generated adjunct questions. Notably, Syed *et al.* [80] found that in the context of reading comprehension, learners who received automatically generated adjunct questions spent longer reading time and paid more attention to reading material than those without such questions. Additionally, the impact of adjunct questions on learning outcomes varied depending on learners' prior knowledge. Learners with low prior knowledge benefited from adjunct questions significantly in terms of long-term retention, in contrast to other conditions and short-term retention.

Despite these insights, two major limitations remain. First of all, as a crucial way of learning, the searching component is absent from these studies [80, 83, 84] which focused on the controlled online reading scenario with a relatively small number (around 100) of questions. Secondly, these studies evaluate learning outcomes with factoid questions instead of higher-level skills like writing. Learning through web search engines significantly differs from the controlled reading comprehension settings—learners choose their own queries, have access to a much larger body of documents, and self-select documents to view—the effect of including adjunct questions in the SAL setting is not yet well understood.

In our study, we endeavor to address this gap by investigating the effects and factors of actively involving learners with adjunct questions integrated within the reading interface. We aim to assess the impact of this approach on both learners' behavior and learning outcomes in the context of learning-oriented search tasks. We implemented a search system that supports adjunct questions with a corresponding UI widget on top of the open-source SearchX system [225] as shown in Figure 6.1. We conducted a between-subjects user study with $N = 144$ participants, where participants were assigned to one of the following four variants: *(i)* $\mathbf{Q}_{none}$*, the control condition without adjunct questions; (ii)* $\mathbf{Q}_{synthesis}$*, synthesis questions; (iii)* $\mathbf{Q}_{random}$*, factoid questions targeting random text spans; and (iv)* $\mathbf{Q}_{term}$*, factoid questions targeting terms and phrases relevant to the information*

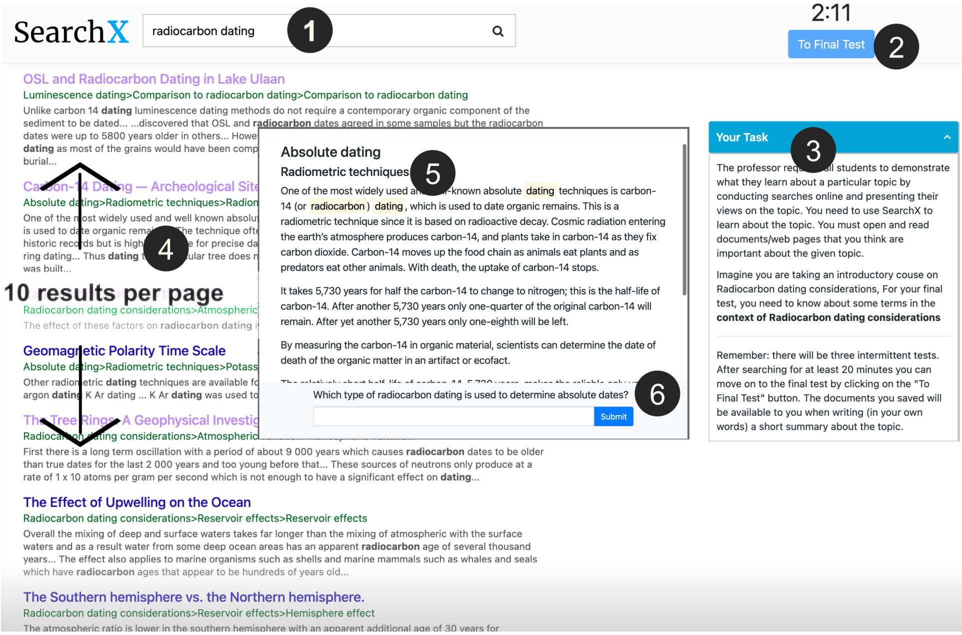


Figure 6.1: Screenshot of the system interface used by participants for searching and learning on the assigned topic (e.g., *radiocarbon dating considerations*). The circled numbers correspond to the narrative of Section 6.3.1.

need at hand. Participants' learning outcomes were measured by two tasks: a recall-based vocabulary learning task [223, 226, 227], and an essay writing task [223, 228, 229] that involved higher cognitive complexity. With this user study, we aim to answer the following research questions *in the SAL setting*:

RQ1 *To what extent do automatically generated adjunct questions impact learners' behaviour and learning outcomes?*

RQ2 *How do the characteristics of adjunct questions, including the question types—factoid questions vs. synthesis questions*[1]*—and the selection of questioning targets and the participants' prior knowledge affect the participants' learning?*

Overall, we find **(i)** compared to the control condition, adjunct questions have a significant influence on learners' behaviour, such as more fine-grained reading evidenced by more reading time and scrolls, as well as fewer queries in the search session; **(ii)** question types (factoid v.s. synthesis) have significant influence on participants' reading behaviour, and with synthesis questions, participants achieve better learning outcomes on the task that requires higher cognitive complexity than those with factoid questions regarding random text spans; **(iii)** the target spans of adjunct questions (random vs. focused) have a significant influence on learning outcomes. $\mathbf{Q}_{\text{term}}$ participants have 76.9% higher vocabulary knowledge gains than $\mathbf{Q}_{\text{random}}$; **(iv)** Participants' prior knowledge levels affect adjunct

[1]Factoid questions require only the extraction of basic facts. Synthesis questions require higher-level cognitive skills like integrating, evaluating, and analyzing different facts.

questions' effects on their learning outcomes and reactions to different AQG strategies. Participants with higher prior-knowledge, in general, achieve better learning gains. These findings provide empirical evidence that to incorporate adjunct questions into a learning-oriented searching system, it is essential to identify the learners' learning target and their prior-knowledge and then generate different types of questions accordingly.

6.2 Background

6.2.1 Search as Learning

Unlike traditional ad-hoc search systems that generally consider a user's information need as atomic (i.e., a single information need is covered by a single user query) [93, 230, 231], a search system designed for SAL must be aware of the nature of users' tasks [222, 224, 232, 233], as these may encompass multiple rounds of interaction with the system, with varying degrees of complexity. Over the past decade, SAL has attracted considerable attention, and many different approaches which touch different parts of the search system to help users learn while searching have been proposed.

Backend adaptations.

Search systems are naturally complex, with multiple components working together to help the user search for relevant documents. While many prior SAL works have focused on front-end adaptations (as we will discuss below), studies investigating how changes made directly to the retrieval pipeline impact learner's behaviour are still rare. For example, Syed and Collins-Thompson [234] designed a retrieval algorithm to improve the ranking of documents with a higher density of vocabulary terms related to the topic the learner is interested in. Collins-Thompson *et al.* [235] demonstrated how tweaking the ranking system according to the learner's reading level can also be beneficial. Finally, Athukorala *et al.* [236] showed that a reinforcement-learning-based ranking algorithm can improve the learner's experience by balancing the diversity or depth of the search results according to the learner's intention.

Frontend adaptations.

Most prior SAL works have focused on aiding users in writing queries and organizing thoughts and content. Learning-oriented adaptations to the Search Engine Results Page (SERP), such as displaying an outline of the topic the learner is interested in [223], providing entity cards [237], or including conversational interfaces [238] have been shown to help users with their knowledge acquisition process—at least to some extent. Approaches that help formulate queries have also been studied since learners' querying behaviour plays a vital role in their learning process [239, 240]. For instance, inspired by [241], Câmara *et al.* [223] displayed a progress bar that estimates how much topic exploration has been done, and this considerably influenced learners' querying behaviour. Another type of change made to the UI is related to how learners organize their materials and thoughts, as explored in [227] where learners were prompted to highlight parts of the text they may find relevant and take notes directly on the SERP. Liu *et al.* [242] asked learners to build mind-maps. They considered search as a method of keeping track and organizing complex information, leading to a measurable change in user behaviour and knowledge gains.

Active learning.
Some of the strategies mentioned above are examples of educational active learning techniques. Instead of passively reading, the learner *actively* engages with the learning material. These strategies have consistently been shown to considerably improve learner's knowledge retention [243–245]. A popular method of implementing active learning is asking learners questions about the material they come across during the search process. These questions are designed to guide learners' attention to specific portions of the material (ideally those covering the key ideas) and, therefore, help learners to understand and remember the material better [216, 246]. The effects of using questions to foster learning (i.e., adjunct question effects) are well known in the classroom setting [3, 220, 221, 247]. Importantly, these questions are typically created manually by topical and educational experts. This is an expensive and slow process and not feasible to do at scale, considering the quantity and diversity of online learning materials. In contrast, automatic question generation is scalable.

6.2.2 Automatic Question Generation

As a critical Natural Language Processing (NLP) task, AQG has been heavily researched over the past decades. Various template-based [11, 21, 24] and neural network-based [13, 18, 248] methods have been proposed. Like other NLP tasks, with the advance of PLM, AQG approaches have jumped considerably in quality as measured by automatic metrics and human evaluations [34, 36, 249–252]. Some works have investigated the application of AQG to education [9, 15, 19, 54, 253–257]. These prior works though, focus mainly on how to apply AQG methods to educational materials and how to generate various types of questions for educational purposes. *The effects of automatically generated questions on human learning still need to be well investigated.*

6

Several works [80–83] have recently begun to study this question. In particular, Syed *et al.* [80] systematically analyzed the effectiveness of AQG on human learning compared to manually curated questions, as well as other impact factors such as learners' prior knowledge, the type of adjunct questions (factoid or synthesis), and the content that questions focused on. Like Syed *et al.* [80], Steuer *et al.* [83] studied automatically generated adjunct questions' effects on non-native speakers' English vocabulary learning. The effects were evaluated by the self-report of prior-knowledge on the topic and the correctness of post-test questions. Van Campenhout *et al.* [84] used automatically generated questions in a university course as formative practice and evaluated the questions' effects by measuring the students' behaviour such as engagement in practice. As mentioned, these works were conducted in a *controlled reading comprehension scenario* by showing participants *one Wikipedia article* or *a fixed list of documents* and corresponding questions (around 100).

Finally, we point out that two types of questions were considered: factoid questions, i.e., questions that ask the knowledge of specific facts from the document which primarily address the *Remembering* level of cognitive complexity in Bloom's taxonomy [8], and the synthesis questions which require higher levels of cognitive complexity like *Analyze* and *Evaluate*. In Syed *et al.* [80], while the factoid questions were automatically generated, the synthesis questions were not. This chapter extends prior work in two directions: (i) we instantiate the concept of adjunct questions in an actual search system, and (ii) we automatically generate different types of questions and investigate their effect on behaviour

& learning.

6.3 Adjunct Questions in SearchX

6.3.1 SearchX Interface

To carry out this study, inspired by [80, 223, 227], we used SearchX [225], a modular, open-source framework that supports IR experiments. SearchX contains a number of modern search engine front-end features and widgets akin to a contemporary web search engine's SERP. Moreover, combined with LogUI [258], it offers fine-grained search logs (hovers, clicks, scrolls, etc.). Figure 6.1 shows the interface we implemented for our experiments. ❶ represents the query box (without query auto-completion). ❷ denotes the timer to help participants count the task time. After the search session lasts at least 20 minutes, the *To Final Test* button becomes available and leads the participant to the post-test when clicked. The task description is shown in ❸, where the assigned topic is bold-faced. ❹ represents the search results page. We show 10 results per page and up to 5 pages, which we consider sufficient search depth as participants only sometimes go beyond the second page [233, 259]. The search results are provided by the BM25 ranker of ElasticSearch[2]. Notably, we show a short snippet created by extracting document sentences containing content words of the query in order to provide participants with essential information. Once the participant clicks on a link, a scrollable document viewer ❺ pops up and displays the document. At the bottom of the viewer is the AQG widget ❻ which is invisible for participants in the control condition ($\mathbf{Q}_{\text{none}}$). In the other three conditions, it shows one automatically generated question about the document. A participant can only proceed to another document or the SERP if they provide *some* answer to the question. The answer correctness does not affect participants' payment and is only used for further analysis.

6.3.2 Automatic Adjunct Question Generation

Dataset

Realistic learning by searching involves searching, reading, and gathering knowledge over large-scale open-domain documents. While we could have opted for a web search API as a retrieval backend, this was not feasible as we could not generate questions at scale from any website within a few milliseconds. Instead, we selected a corpus and pre-computed the questions of each type. Specifically, we used the `benchmarkY1train` set from TREC-CAR v1.5 dataset [260]. This dataset contains a set of structured Wikipedia topics with headings designed for retrieving answers for complex information needs and also used for SAL research [223, 227]. Moreover, we used topics and vocabulary terms (phrases representing the topic) created by [223] from the same subset for the following learning tasks. Additionally, we extracted 136 topics in the Wikipedia dumps that contained the vocabulary terms to ensure participants were posed with plenty of documents containing the target vocabulary terms. In total, we used 253 Wikipedia topics. As each topic corresponds to one long Wikipedia article that requires considerable reading time, as shown in [80], we split articles into 1,627 documents based on their heading structures to engage participants with more searching and reading behaviour.

[2]https://www.elastic.co/

Table 6.1: An example of automatically generated questions from a given document. Shown here are two factoid questions (Q_{random}, Q_{term}) and a synthesis question ($Q_{synthesis}$). Highlighted in cyan and green are answers for creating the corresponding factoid questions. Extracted word spans that are filtered out are highlighted in violet.

Example 2	Irritable bowel syndrome	
Document	Approximately 10 percent of IBS cases are triggered by an acute gastroenteritis infection. Genetic defects relating to the innate immune system and epithelial barrier as well as high stress and anxiety levels appear from evidence to increase the risk of developing post-infectious IBS. Post-infectious IBS usually manifests itself as the diarrhea predominant subtype. Evidence has demonstrated that the release of high levels of proinflammatory cytokines during acute enteric infection causes increased gut permeability leading to translocation of the commensal bacteria across the epithelial barrier resulting in significant damage to local tissues which is likely to result in chronic gut abnormalities in sensitive individuals. However, increased gut permeability is strongly associated with IBS regardless of whether IBS was initiated by an infection or not.	
Q_{random}	What percentage of ibs cases are triggered by an acute gastroenteritis infection?	10 percent
Q_{term}	What part of the gut is affected by irritable bowel syndrome?	epithelial barrier
$Q_{synthesis}$	Why do some people develop IBS more often than others?	

We studied two categories of questions and employed separate question generators for each: (i) factoid (or low-level) questions and (ii) synthesis (or high-level) questions. As illustrated in Table 6.1, factoid questions seek text spans that pertain to specific facts, such as concepts and numbers, which can be directly retrieved from the text. In contrast, synthesis questions necessitate comprehensive efforts, such as integrating and analyzing document information, surpassing the mere extraction of text spans.

Factoid Question Generation

We used the PAQ [250] framework for generating factoid questions. First, we utilized two extraction methods provided by PAQ to identify text spans within a document that are worth questioning. One method involved extracting all named entities as potential answers, as named entities such as names, numbers, and locations often convey significant information. The other method involved a trained neural model as the answer span extractor called `Span2DAnswerExtractor`[3]. In addition, we also included all vocabulary terms as question-worthy text spans as experts chose them as the most representative terms for each topic. We opted for the `qgen_multi_base`[4], a BART [36]-based model fine-tuned on various QA datasets as the factoid question generator. It took the document and extracted text spans as inputs, resulting in 65,237 questions. The generated questions underwent

[3] https://github.com/facebookresearch/PAQ#answer-extraction
[4] https://github.com/facebookresearch/PAQ

6

a filtering process concerning question length and consistency. First, questions shorter than 6 words were disregarded, resulting in the removal of 392 questions. Subsequently, the remaining questions were filtered using PAQ's QA-Pair filtering tool, which assessed the consistency between the answer and the generated question. This step led to the further filtering of 37,016 questions. If multiple valid QA pairs existed for a single document, we selected the pair with the highest answer score for that document. We then separated all factoid questions into two groups: questions regarding the vocabulary terms ($\mathbf{Q}_{\text{term}}$, 750) and other text spans ($\mathbf{Q}_{\text{random}}$, 1,627). Although answers for $\mathbf{Q}_{\text{random}}$ tend to be informative and important, they were extracted regardless of the participants' learning goals. Therefore they are independent or *random* from the learning purpose.

Synthesis Question Generation

Synthesis questions typically require more than text spans from the documents to provide comprehensive answers. PAQ is primarily trained to cater to factoid questions, so it may not be well-suited for generating synthesis questions. To address this limitation, we opted to fine-tune the BART model [36] using the ELI5 dataset [53] which comprises complex, diverse questions that require long-form multi-sentence answers, e.g., *Why are flutes classified as woodwinds when most of them are made out of metal?*, aligning with the requirements of synthesis question generation. We generated one synthesis question for each document paragraph, resulting in 5,393 synthesis questions. Among the questions of the same document, we selected the longest one as the synthesis question for the study. Table 6.1 shows examples of our generated factoid and synthesis questions. As shown in these examples, facts to answer the $\mathbf{Q}_{\text{random}}$ and $\mathbf{Q}_{\text{term}}$ questions can be directly found in the document as text spans. In contrast, the generated synthesis questions require comparing and analyzing document contents.

6

6.3.3 Question Quality Evaluation

We took a random sample of 30 generated questions from $\mathbf{Q}_{\text{random}}$, $\mathbf{Q}_{\text{term}}$, and $\mathbf{Q}_{\text{synthesis}}$, respectively, in order to evaluate the quality of the generated questions. In addition, we also chose 30 human-curated questions from the 4 SQuAD [50] articles used in [80] for comparison. We conducted human evaluation by recruiting five native English speakers with at least undergraduate degrees as annotators. The questions were rated on a 5-point scale concerning their relevance to their context, the answerability, i.e., whether they can be answered with information from the document, and the possibility that a human wrote the question. The final rating of each question is determined by averaging all annotators' ratings. Table 6.2 reports the average score along all these measures. We conducted a one-way ANOVA test on the measures with respect to the question type factor. First, the average length of questions ranged from 11.3 to 13.2, and there was no significant difference. Second, we can observe that although automatically generated questions were considered less human-written, they were still considered as likely written by humans (> 3.4 on a 5-point scale, compared to 4.28 for human-curated SQuAD questions). Furthermore, $\mathbf{Q}_{\text{synthesis}}$ questions were significantly lower than the SQuAD questions in terms of relevance ($p < 10^{-4}$) and answerability ($p < 10^{-4}$). One possible reason is synthesis questions tend to require more cognitive complexity and background knowledge than SQuAD questions which are simple factoid questions. Notably, the answerability of $\mathbf{Q}_{\text{random}}$ questions

Table 6.2: Comparison of SQuAD and automatically generated questions in terms of average question length and human evaluation for Relevance, Answerability, and Human-Written (H-W) on a 5-point scale. † denotes the one-way ANOVA significance, while $\mathscr{U}$(SQuAD), $\mathscr{S}(\mathbf{Q}_{\text{synthesis}})$, $\mathscr{R}(\mathbf{Q}_{\text{random}})$, $\mathscr{T}(\mathbf{Q}_{\text{term}})$ indicate post-hoc significance (TukeyHSD pairwise test, p<0.05) over four groups of questions.

Method	**Length**	**Relevance**†	**Answerability**†	**H-W**†
SQuAD	11.3	$4.75^{\mathscr{T}\mathscr{S}}$	$4.68^{\mathscr{T}\mathscr{S}}$	$4.28^{\mathscr{T}\mathscr{S}}$
$\mathbf{Q}_{\text{random}}$	13.2	4.34	$4.10^{\mathscr{S}}$	3.79
$\mathbf{Q}_{\text{term}}$	12.8	$4.03^{\mathscr{U}}$	$3.55^{\mathscr{U}}$	$3.55^{\mathscr{U}}$
$\mathbf{Q}_{\text{synthesis}}$	12.0	$3.87^{\mathscr{U}}$	$3.47^{\mathscr{U}\mathscr{R}}$	$3.44^{\mathscr{U}}$

was significantly higher than that of synthesis questions ($p = 0.034$). This is aligned with our design since the synthesis questions are supposed to be more challenging to answer, which leads to lower answerability ratings from annotators.

6.4 User Study Design

6.4.1 Topics

In line with prior research [223, 227, 228], we assessed participants' learning outcomes with two learning-focused tasks: a recall-based vocabulary learning task and an essay writing task. The vocabulary-learning task assessed knowledge levels on vocabulary terms at cognitive levels like remembering and understanding based on revised BLOOM's taxonomy [8]. On the other hand, the essay writing task required participants to compose a summary of at least 100 words based on their acquired knowledge during the search session. This task aimed to assess higher cognitive levels, such as evaluating and analyzing. We chose seven topics along with the vocabulary terms most representative of each topic created by Câmara *et al.* [223]. These topics have suitable complexity, so they are not too easy that most participants already have plenty of knowledge or are too hard to learn in twenty minutes. Table 6.3 presents the topics and vocabulary terms.

6.4.2 Experimental Conditions

As mentioned earlier, in our user study, we randomly assigned each participant to one of the following four conditions:

$\mathbf{Q}_{\text{random}}$ For participants in this condition, to each document the participant opened, we presented one automatically generated factoid question regarding a text span like one named entity randomly sampled from each opened document.

$\mathbf{Q}_{\text{term}}$ In this condition, if there were vocabulary terms of the assigned topic in the document, we presented the participant with an automatically generated factoid question regarding one of the vocabulary terms. Otherwise, a random factoid question would be presented instead.

$\mathbf{Q}_{\text{synthesis}}$ In this condition, we presented a participant with a high-level synthesis question about the opened document.

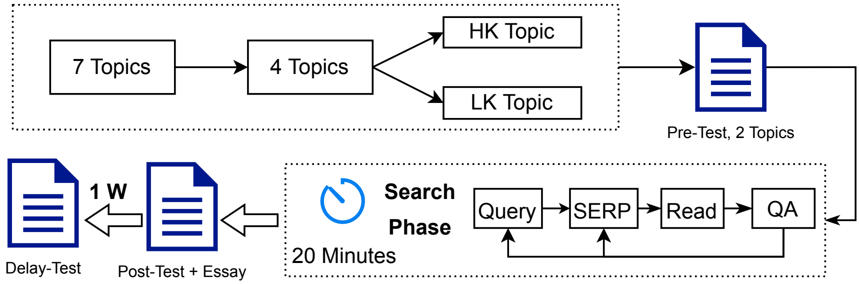


Figure 6.2: Illustration of the user study workflow. This flow describes the experimental conditions of $\mathbf{Q}_{\text{random}}$, $\mathbf{Q}_{\text{synthesis}}$, and $\mathbf{Q}_{\text{term}}$. The $\mathbf{Q}_{\text{none}}$ condition does not take the QA step.

$\mathbf{Q}_{\text{none}}$ In the control condition we did not show participants the AQG widget (6 in Figure 6.1) in the document viewer.

6.4.3 Study Workflow

We now briefly introduce the experimental procedure that consists of seven phases.

1. **Task Introduction.** Participants read a general introduction to the entire study workflow.

2. **Survey.** Participants were asked to complete a demographics survey containing questions regarding their education level, language skills, and their use of web search engines and online documents for learning.

3. **Topic Selection.** We selected seven topics for our user study. To prevent participants from the familiarity bias [261], we designed a two-step knowledge selection procedure. First, we randomly chose four of seven topics and asked participants to choose one topic they knew best and one they knew least. Both of the topics were used for the vocabulary knowledge pre-test.

4. **Pre-Test.** Participants were asked to complete two vocabulary knowledge tests. Each test consisted of 10 vocabulary questions on the topics selected in the above topic selection phase. We chose one topic randomly with equal probability as the assigned topic for participants to learn more about in the following phases.

5. **Search Phase** Participants were randomly assigned to one of our four experimental conditions. Participants needed to spend at least 20 minutes searching and reading documents to learn about the assigned topic in line with prior research [80, 233].

6. **Post-test.** After 20 minutes in the search phase, participants could continue to the post-test which consisted of a vocabulary test on the assigned topic (same 10 questions as in the pre-test in shuffled order) and an essay writing assignment (100+ words).

7. **Delay-test** One week after the post-test, participants were invited to take a delay-test which consisted of the vocabulary test as the post-test in different question order.

6.4.4 Participants

We conducted our user study on the *Prolific Academic*[5] platform. The required number of participants was determined by a statistical power analysis conducted with a significance level of $\alpha = 0.05$, a power of $1 - \beta = 0.80$, an expected effect size of 0.25 and a group size of 4 using the software *GPower* [262]. This gave a minimum required number of $n = 136$ participants. To ensure the response quality, we only recruited native English-speaking participants within the age range of 18 to 51 with a minimum of 95% approval rate, at least 100 successful task submissions, and at least a high-school level of education. The entire study lasted for around 35 minutes. We paid each participant GBP £5 for the study. Overall, 178 participants completed the post-test; we rejected 18 of them because of a lack of attention (over 5 minutes of no activity in the browser tab) in the search phase, which led to 160 valid participants. We further paid £1 bonus for participants who took the delay-test after one week, and 144 valid participants returned and completed the delay-test. Among the 144 participants (77 male, 67 female), the median age is 34.5 (min. 20, max. 51). Forty reported a high school degree as the highest education degree, 17 reported a community college degree, 58 reported an undergraduate degree, 25 reported a graduated degree, and 4 reported doctorate degrees. Table 6.3 reports the distributions of participants over topics and experimental conditions. The 144 participants were evenly distributed among the topics, each with a participant count ranging from 19 to 23. Table 6.3 also shows the average number of queries over each topic, which ranges from 3.26 to 4.65, indicating that our participants actively engaged in the search phase.

6.4.5 Metrics

Learning Gains

In the pre-, post-, and delay–tests, we asked our participants to self-assess their knowledge levels on a set of vocabulary terms. In line with [223, 227, 229, 233], we evaluated the study participants' knowledge of a term with the *Vocabulary Knowledge Scale (VKS)* [226] across four levels:

1 *I don't remember having seen this term/phrase before.*

2 *I have seen this term/phrase before, but I don't think I know what it means.*

3 *I have seen this term/phrase before, and I think it means ...*

4 *I know this term/phrase. It means ...*

Besides, we further asked participants to write down the meaning of the vocabulary term in their own words for vocabulary in knowledge level *(3)* and *(4)*, which we can use to judge the quality and reliability of the self-assessment. To reduce the question priming effects, participants did not know that vocabulary terms asked in the pre-test would be asked again in the post-test. Following earlier works [223, 227, 233], we first rescored the knowledge level self-assessments to 0 – 2, specifically, knowledge level *(1)* and *(2)* was rescored as 0, knowledge level *(3)* was rescored as 1, and knowledge level *(4)* as 2. Then we evaluated the learning gain with *Realized Potential Learning (RPL)* [233, 263, 264]. RPL

[5]https://app.prolific.co

Table 6.3: Overview of topics and corresponding vocabulary terms chosen for learning tasks, as well as number of participants and other associated statistics (± represents the standard deviation) over topics. Two-way ANOVA tests revealed no significant differences in the average number of queries ($F(6, 132) = 0.839$, $p = 0.542$).

	Ethics	**Genetically modified organism**	**Noise-induced hearing loss**	**Radiocarbon dating considerations**	**Business cycle**	**Irritable bowel syndrome**	**Theory of mind**
Vocabulary Terms	anarchist ethics, descriptive ethics, normative ethics, relational ethics, virtue ethics, ethical resistance, consequentialism, epicurean ethics, ethics feasible, ethics spheres	transgenic, genomes, selective breeding, microinjection enzyme, chromosome, plasmid, myxoma, kanamycin, severe combined immunodeficiency, Leber's congenital amaurosis	acoustic trauma, discomfort threshold, cochlear damage, audiogram, overstimulation of hair cells, noise conditioning, excitotoxicity, OSHA, sensorineural hearing loss, tinnitus, Threshold shift	carbon exchange reservoir, isotopic fractionation, polarity excursion, carbonate, geomagnetic reversals, mass spectrometry, upwelling, radiocarbon, neutrons, photosynthesis pathways	economic cycles, distribution cycles, swing cycle, wage cycle, marxist model, endogenous causes, friedman, capital profitability, model recession, austrian school	bifidobacteria infantis, mesalazin, bile acid malabsorption, selective serotonin reuptake inhibitors, Gut-brain axis, antidepressants, laxatives, probiotics, celiac disease, epithelial barrier	asperger syndrome, theory of mindreading, attentional reorienting, mind development, mind autism, hyperactivity, perspective experiment, intentionality, perception, belief
Number of participants	21	20	21	19	19	21	23
Q_{none}	3	4	6	5	5	4	5
$Q_{synthesis}$	7	6	5	3	5	4	7
Q_{random}	6	6	4	5	4	5	6
Q_{term}	5	4	6	6	5	8	5
Average number of queries	4.43(±4.20)	4.65(±3.48)	3.48(±2.34)	4.00(±3.64)	4.42(±3.44)	3.71(±2.63)	3.26(±3.52)
Median number of queries	3.00	3.50	3.00	2.00	3.00	3.00	2.00

is the absolute knowledge gain (ALG) normalized by the maximum possible learning gain (MLG). The absolute knowledge gain represents the sum of knowledge level changes from the pre-test to subsequent tests. The maximum possible learning gain means the total possible knowledge level changes from the pre-test to the subsequent tests.

$$ALG = \frac{1}{n}\sum_{i=1}^{n} \max(0, vks^{x}(v_i) - vks^{pre}(v_i)) \tag{6.1}$$

$$MLG = \frac{1}{n}\sum_{i=1}^{n} 2 - vks^{pre}(v_i) \tag{6.2}$$

$$RPL = \frac{ALG}{MLG} \tag{6.3}$$

$$\tag{6.4}$$

where $vks^{pre}(v_i)$ is the rescored knowledge level of vocabulary v_i in pre-test; $vks^{x}(v_i), x \in \{\text{post}, \text{delay}\}$ is the rescored knowledge level of vocabulary v_i in post- or delay-test. $vks(v_i) \in \{0, 1, 2\}$ and n is the number of vocabulary items under the tested topic.

6

Self-assessment Quality

In order to determine the quality of vocabulary knowledge self-assessments, we sampled $approx$10% of term definitions of knowledge levels (3) and (4) from both the pre- and post-tests (specifically, 40 from the pre-test and 60 from the post-test) written by participants. We tasked two experts to label these definitions as either *correct*, *partially correct*, or *incorrect*, keeping in mind that the definitions were written by topical novices. Based on the expert labels, among definitions of knowledge level (3), 20% were correct, 68% were partially correct, and the rest 12% were incorrect. Among the definitions of knowledge level (4), 70% were correct, 24% were partially correct, and 6% were incorrect. Based on the low incorrect rate, we consider the self-assessment reliable.

Automatic Assessment

Another way to scale up the assessment of our participants' definitions is to rely on large-scale language models (LLMs). State-of-the-art LLMs such as GPT-3.5 or GPT-4 [265] have reportedly achieved human-level performance on various complex natural language tasks. To evaluate the influence of uncertainty in self-assessment, we evaluated all definitions with the following prompt template (with `{...}` indicating placeholders):

> You are an expert in the topic `{topic}`, and your goal is to evaluate if a beginner's definition of a term related to this topic is correct. The definitions are written by novices who just started reading the materials. The definitions are supposed to be short, somewhat unprofessional, and with no more than 10 words.
> First, write your own short and correct definition for each term. Then, reason whether or not the user's answer is correct. After this, evaluate the user's definition with the following scale:
>
> 1) Incorrect: the definition is entirely incorrect. (e.g. "Positively charged particles that form part of the atomic structure in the nucleus." is an incorrect description of "neutrons").
> 2) Partially correct: the definition captures part of the base concepts correctly, but part of the definition is incorrect. (e.g., "A part of an atom." is a Partially Correct answer to the term "neutrons").
> 3) Correct: the definition captures some basic concepts and contains no incorrect description. (e.g., "Carbon 14 decays and dating can be done using this fact." is a correct definition for the term "radiocarbon").
>
> Answer it in the following format:
> Definition:
> Reason:
> Evaluation: 1, 2, or 3
> —
> Term: `{term}`
> Definition: `{definition}`

6

Based on GPT-3.5's output, we categorized *partially correct* term definitions as knowledge level (3) and *correct* term definitions as knowledge level (4). Incorrect term definitions were designated as level (2). We conducted our data analyses with the self-assessment and knowledge levels as determined by GPT-3.5. The trends and statistical outcomes do not differ between self-assessment and GPT-3.5 based assessment[6]. **Thus, due to space constraints, in the remainder of this chapter, we report the learning gain evaluation based on the self-assessed vocabulary knowledge levels only.**

Essay Quality

In addition to RPL, we evaluated knowledge expressed in participants' essays with two additional measures as learning indicators: F-Fact and T-Depth, following [227, 266]. Concretely, F-Fact represents the number of individual facts in an essay, and T-Depth represents the extent to which each subtopic is covered. We manually annotated the written essays for both measures. For F-Fact, the annotators were required to identify topic-related facts and count the number of facts in an essay. For T-Depth, annotators scored the essay on a scale of 0 to 3, where 0 represented not covered and 3 indicated the essay covered the topic with great focus. Five annotators divided 160 essays among themselves. Twenty

[6]As an example, for the learning gain metric RPL, the scores were no larger than 0.01.

essays were annotated by all annotators, achieving a Pearson correlation of 0.73 for *F-Fact* and 0.75 for *T-Depth*, indicating high inter-annotator agreement for the metrics.

Behaviour Metrics

Following prior research [80, 229, 267], we extracted seven types of search and reading behaviour from our collected search logs: (i) the **number of queries** a participant formulates; (ii) the **number of unique documents** a participant viewed; (iii) the **number of snippets** a participant viewed; (iv) the **average time of between queries**; (v) the **average time between documents**; (vi) the **average document dwell time**; (vii) the **number of mouse scrolls** over the opened documents.

Answer Quality

To examine whether participants indeed engaged with the adjunct questions, we evaluated participant-written answer quality of the factoid questions with EM (Exact Match) score, which measures the percentage of answers that match exactly with ground-truth answers, and (macro-averaged) F1 score [50, 174], which treats all answers as bags of tokens and calculate the average overlap between the participants' answers and the ground truth answer. We found the F1 scores of answers to $\mathbf{Q}_{\text{random}}$ and $\mathbf{Q}_{\text{term}}$ were 0.589 and 0.408, and the EM scores of answers to $\mathbf{Q}_{\text{random}}$ and $\mathbf{Q}_{\text{term}}$ were 0.523 and 0.322, respectively. These results confirmed that participants indeed engaged with the questions.

6.5 Results

In this section, we discuss the results of our user study. As a sanity check, we first analyze participants' overall learning gains. Figure 6.3a reports the distribution of knowledge levels reported in pre-, post-, and delay-tests. Participants marked fewer vocabulary terms as knowledge levels 1 or 2 and more as knowledge levels 3 and 4 in post- and delay-tests than the pre-test, which shows that participants learned both short-term (post-test) and long-term (delay-test) vocabulary knowledge over the assigned topics in the task. Furthermore, Figure 6.3b shows detailed knowledge state transitions on each condition from pre-test to post-test. Although the assessment on most vocabulary terms (> 50%) remained unchanged and transitions among lower knowledge levels accounted for most learning gains, participants did achieve learning gains in all conditions. These results, together with the evaluation of participants' self-assessment quality (Section 6.4.5) and the quality of answers to the adjunct questions (Section 6.4.5), validate our system and experimental design. On average, participants were indeed actively engaged and learning throughout the study.

We now present study results in line with the research questions. Table 6.4 presents the main results. We conducted two-way ANOVA tests on these measures, considering the assigned topics and the conditions as factors, and examined the main effects with $\alpha = 0.05$. We then used TukeyHSD pairwise tests for post-hoc analysis.

6.5.1 Adjunct Question Effects in SAL

Effects of Adjunct Questions on Participants' Search Behaviour

In our user study, participants were required to learn one topic by searching and reading for at least 20 minutes. The average document dwell time (Row **XIV** in Table 6.4) of par-

Table 6.4: Mean (± standard deviations) of evaluation metrics across all participants in each condition. † denotes the two-way ANOVA significance, while $\mathcal{N}$,$\mathcal{S}$,$\mathcal{R}$,$\mathcal{T}$ indicate post-hoc significance (TukeyHSD pairwise test, p<0.05) over the four conditions $\mathbf{Q}_{\text{none}}$, $\mathbf{Q}_{\text{synthesis}}$, $\mathbf{Q}_{\text{random}}$, and $\mathbf{Q}_{\text{term}}$, respectively.

	Measure	$\mathbf{Q}_{\text{none}}$	$\mathbf{Q}_{\text{synthesis}}$	$\mathbf{Q}_{\text{random}}$	$\mathbf{Q}_{\text{term}}$
I.	**Number of participants**	32	37	36	39
II.	**Search phase**	19m47s(±1m46s)	21m5s(±3m16s)	20m6s(±2m30s)	20m18s(±2m56s)
III.	**Post-Test RPL**†	0.23(±0.18)$^{\mathcal{R}}$	0.14(±0.14)	0.13(±0.12)$^{\mathcal{N}\mathcal{T}}$	0.23(±0.15)$^{\mathcal{R}}$
IV.	**Delay-Test RPL**	0.17(±0.20)	0.11(±0.12)	0.10(±0.12)	0.16(±0.19)
V.	**Flesch score**	52.00(±12.33)	49.67(±17.06)	52.56(±13.23)	56.49(±13.42)
VI.	**T-Depth**	0.87(±0.35)	0.91(±0.48)	0.78(±0.47)	0.81(±0.41)
VII.	**F-Fact**	13.88(±7.82)	12.68(±6.38)	10.16(±7.28)	12.68(±8.57)
VIII.	**Fraction of topical terms used by essays**	0.05(±0.04)	0.04(±0.03)	0.03(±0.03)	0.04(±0.02)
IX.	**Number of queries**†	6.09(±3.90)$^{\mathcal{S}\mathcal{R}\mathcal{T}}$	3.49(±2.80)$^{\mathcal{N}}$	3.11(±2.78)$^{\mathcal{N}}$	3.49(±3.16)$^{\mathcal{N}}$
X.	**Number of unique documents viewed**†	13.44(±6.65)$^{\mathcal{S}\mathcal{R}\mathcal{T}}$	7.49(±3.01)$^{\mathcal{N}}$	9.47(±5.14)$^{\mathcal{N}}$	9.10(±3.89)$^{\mathcal{N}}$
XI.	**Number of snippets**†	45.09(±21.16)$^{\mathcal{S}\mathcal{R}\mathcal{T}}$	31.57(±17.55)$^{\mathcal{N}}$	32.47(±16.61)$^{\mathcal{N}}$	28.38(±15.70)$^{\mathcal{N}}$
XII.	**Average time between queries (secs.)**†	379.97(±329.66)$^{\mathcal{S}\mathcal{R}\mathcal{T}}$	685.47(±368.11)$^{\mathcal{N}}$	740.67(±375.35)$^{\mathcal{N}}$	615.45(±380.46)$^{\mathcal{N}}$
XIII.	**Average time between documents (secs.)**	18.78(±16.89)	20.20(±13.70)	20.40(±19.73)	19.85(±20.87)
XIV.	**Average document dwell time(s)**†	73.99(±35.46)$^{\mathcal{S}\mathcal{T}}$	182.89(±149.15)$^{\mathcal{N}\mathcal{R}\mathcal{T}}$	128.21(±54.73)$^{\mathcal{S}}$	128.93(±59.32)$^{\mathcal{N}\mathcal{S}}$
XV	**Number of scrolls**†	14.62(±18.76)$^{\mathcal{S}}$	98.18(±94.79)$^{\mathcal{N}\mathcal{R}\mathcal{T}}$	38.99(±43.97)$^{\mathcal{S}}$	36.30(±28.86)$^{\mathcal{S}}$
XVI.	**Average number of non-stopwords in answers**†	–	6.75(±4.50)$^{\mathcal{R}\mathcal{T}}$	0.82(±1.02)$^{\mathcal{S}}$	1.15(±1.38)$^{\mathcal{S}}$
XVII.	**Average reading time before answering (secs.)**	–	116.43(±120.90)	98.32(±46.91)	89.73(±48.79)
XVIII.	**Average time to create answers (secs.)**†	–	36.55(±24.17)$^{\mathcal{R}\mathcal{T}}$	9.82(±8.79)$^{\mathcal{S}}$	13.49(±16.28)$^{\mathcal{S}}$
XIX.	**F1 score**†	–	–	0.58(±0.20)$^{\mathcal{T}}$	0.39(±0.23)$^{\mathcal{R}}$
XX.	**EM score**†	–	–	0.51(±0.24)$^{\mathcal{T}}$	0.30(±0.26)$^{\mathcal{R}}$

6

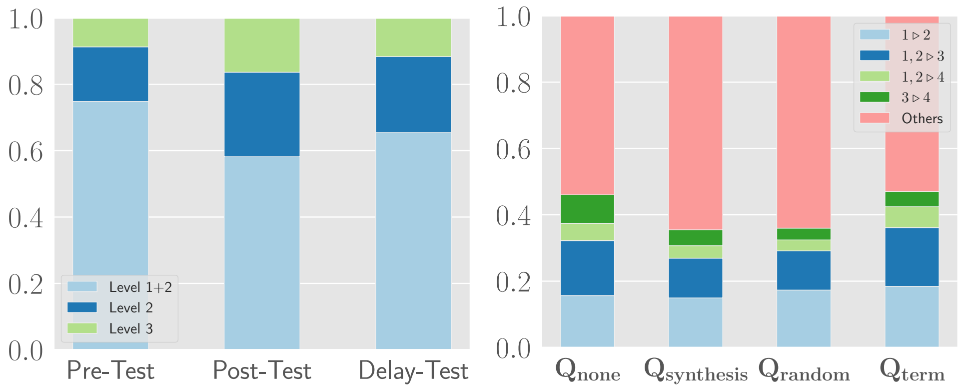
Table 6.5: Mean (± standard deviations) of evaluation metrics across low knowledge participants in each condition. † denotes the two-way ANOVA significance, while $\mathcal{N}$,$\mathcal{S}$,$\mathcal{R}$,$\mathcal{T}$ indicate post-hoc significance (TukeyHSD pairwise test, $p<0.05$) over the four conditions $\mathbf{Q}_{\text{none}}$, $\mathbf{Q}_{\text{synthesis}}$, $\mathbf{Q}_{\text{random}}$, and $\mathbf{Q}_{\text{term}}$, respectively.

	Measure	$\mathbf{Q}_{\text{none}}$	$\mathbf{Q}_{\text{synthesis}}$	$\mathbf{Q}_{\text{random}}$	$\mathbf{Q}_{\text{term}}$
I.	**#Participant**	17	19	18	19
II.	**Reading phase duration (minutes)**	19m22s($\pm 2m4s$)	20m58s($\pm 2m53s$)	19m49s($\pm 2m1s$)	19m58s($\pm 2m46s$)
III.	**Post-Test RPL**	0.19(±0.16)	0.12(±0.14)	0.14(±0.12)	0.19(±0.11)
IV.	**Delay-Test RPL**	0.13(±0.17)	0.06(±0.09)	0.09(±0.11)	0.12(±0.14)
V.	**Flesch Score**	57.12(±9.78)	45.92(±19.23)$^{\mathcal{T}}$	53.05(±13.53)	60.87(±11.42)$^{\mathcal{S}}$
VI.	**T-Depth**	0.84(±0.31)	0.92(±0.30)	0.83(±0.48)	0.83(±0.45)
VII.	**F-Fact**	13.06(±6.06)	13.60(±5.03)	9.93(±7.69)	11.62(±6.08)
VIII.	**Fraction of topical terms used by essays**	0.04(±0.03)	0.05(±0.03)	0.03(±0.02)	0.05(±0.03)
IX.	**Number of queries**	6.00(±3.95)$^{\mathcal{SRT}}$	2.47(±1.61)$^{\mathcal{N}}$	2.61(±1.82)$^{\mathcal{N}}$	3.21(±2.02)$^{\mathcal{N}}$
X.	**Number of unique documents viewed**	13.29(±7.15)$^{\mathcal{ST}}$	6.79(±3.01)$^{\mathcal{N}}$	9.50(±5.66)	8.84(±3.52)$^{\mathcal{N}}$
XI.	**Number of snippets**	47.24(±21.87)$^{\mathcal{SRT}}$	24.47(±13.05)$^{\mathcal{N}}$	31.50(±16.66)$^{\mathcal{N}}$	27.95(±12.77)$^{\mathcal{N}}$
XII.	**Average time between queries (secs.)**	341.84(±256.41)$^{\mathcal{SR}}$	787.13(±350.49)$^{\mathcal{N}}$	760.25(±367.14)$^{\mathcal{N}}$	541.17(±310.61)
XIII.	**Average time between documents (secs.)**	21.55(±19.82)	16.88(±11.15)	23.69(±25.81)	21.83(±23.34)
XIV.	**Average document dwell time (secs.)**	75.29(±35.65)$^{\mathcal{S}}$	221.04(±194.78)$^{\mathcal{NRT}}$	123.12(±49.44)$^{\mathcal{S}}$	125.45(±61.81)$^{\mathcal{S}}$
XV.	**Number of scrolls**	14.61(±21.57)$^{\mathcal{S}}$	115.06(±118.82)$^{\mathcal{NRT}}$	25.85(±20.15)$^{\mathcal{S}}$	31.14(±21.41)$^{\mathcal{S}}$
XVI.	**Average reading time before answering (secs.)**	–	146.57(±158.02)	92.17(±44.11)	99.05(±58.02)
XVII.	**Average number of non-stopwords in answers**	–	7.84(±5.81)	0.58(±0.68)	1.13(±1.43)
XVIII.	**Average time to create answers (secs.)**	–	39.15(±28.58)$^{\mathcal{RT}}$	9.23(±8.39)$^{\mathcal{S}}$	9.24(±8.34)$^{\mathcal{S}}$
XIX.	**F1 score**	–	–	0.62(±0.19)	0.37(±0.25)
XX.	**EM score**	–	–	0.56(±0.20)	0.27(±0.28)

6

Table 6.6: Mean (± standard deviations) of evaluation metrics across high knowledge participants in each condition. † denotes the two-way ANOVA significance, while $\mathcal{N}$,$\mathcal{S}$,$\mathcal{R}$,$\mathcal{T}$ indicate post-hoc significance (TukeyHSD pairwise test, p<0.05) over the four conditions $\mathbf{Q}_{\text{none}}$, $\mathbf{Q}_{\text{synthesis}}$, $\mathbf{Q}_{\text{random}}$, and $\mathbf{Q}_{\text{term}}$, respectively.

	Measure	$\mathbf{Q}_{\text{none}}$	$\mathbf{Q}_{\text{synthesis}}$	$\mathbf{Q}_{\text{random}}$	$\mathbf{Q}_{\text{term}}$
I.	**Number of participant**	15	18	18	20
II.	**Search Phase**	20m16s(±1m17s)	21m14s(±3m43s)	20m22s(±2m57s)	20m37s(±3m7s)
III.	**Post-Test RPL**	0.28(±0.19)$^{\mathcal{R}}$	0.16(±0.14)	0.12(±0.13)$^{\mathcal{N}\mathcal{T}}$	0.26(±0.19)$^{\mathcal{R}}$
IV.	**Delay-Test RPL**	0.22(±0.23)	0.15(±0.14)	0.11(±0.14)	0.20(±0.23)
V.	**Flesch Score**	46.19(±12.63)	53.63(±13.88)	52.08(±13.30)	52.33(±14.11)
VI.	**T-Depth**	0.90(±0.39)	0.90(±0.62)	0.73(±0.46)	0.80(±0.38)
VII.	**F-Fact**	14.80(±9.57)	11.70(±7.57)	10.39(±7.06)	13.69(±10.48)
VIII.	**Fraction of topical terms used by essays**	0.05(±0.04)	0.03(±0.03)	0.03(±0.03)	0.03(±0.02)
IX.	**Number of queries**	6.20(±3.97)	4.56(±3.38)	3.61(±3.47)	3.75(±4.00)
X.	**Number of unique documents viewed**	13.60(±6.29)$^{\mathcal{S}\mathcal{T}}$	8.22(±2.92)$^{\mathcal{N}}$	9.44(±4.73)	9.35(±4.30)$^{\mathcal{N}}$
XI.	**Number of snippets**	42.67(±20.80)	39.06(±18.88)	33.44(±16.99)	28.80(±18.39)
XII.	**Average time between queries (secs.):**	423.18(±402.17)	578.17(±364.89)	721.10(±393.01)	686.01(±432.83)
XIII.	**Average time between documents (secs.)**	15.65(±12.74)	23.70(±15.50)	17.10(±10.60)	17.97(±18.64)
XIV.	**Average document dwell time (secs.)**	72.51(±36.42)$^{\mathcal{S}\mathcal{R}\mathcal{T}}$	142.61(±59.97)$^{\mathcal{N}}$	133.29(±60.57)$^{\mathcal{N}}$	132.24(±58.27)$^{\mathcal{N}}$
XV.	**Number of scrolls**	14.63(±15.73)$^{\mathcal{S}}$	80.37(±58.53)$^{\mathcal{N}\mathcal{T}}$	52.12(±56.65)	41.20(±34.34)$^{\mathcal{S}}$
XVI.	**Average reading time before answering (secs.)**	–	84.62(±49.31)	104.48(±50.04)	80.89(±37.46)
XVII.	**Average number of non-stopwords in answers**	–	5.59(±2.08)	1.06(±1.24)	1.17(±1.37)
XVIII.	**Average time to create answers (secs.)**	–	33.80(±18.89)$^{\mathcal{R}\mathcal{T}}$	10.42(±9.38)$^{\mathcal{S}}$	17.52(±20.71)$^{\mathcal{S}}$
XIX.	**F1 score**	–	–	0.54(±0.21)	0.40(±0.21)
XX.	**EM score**	–	–	0.45(±0.27)	0.32(±0.24)



(a) (b)

Figure 6.3: (a) Distributions of vocabulary knowledge levels in the pre-test, post-test, and delay-test. (b) The fraction of vocabulary knowledge changes from pre-test to post-test.

ticipants who received adjunct questions was significantly longer than $\mathbf{Q}_{\text{none}}$ ($p < 0.05$). As a consequence of the longer dwell time, we also observed the number of queries (Row **IX**, $F(3, 116) = 6.70, p = 3 \times 10^{-4}$), the number of unique documents (Row **X**, $p = 2 \times 10^{-6}$), and the number of unique snippets (Row **XI**, $p = 0.001$) that participants viewed to be significantly lower than participants in $\mathbf{Q}_{\text{none}}$. The average time between queries of participants with adjunct questions (ranging from 636 s to 741 s) was significantly longer than $\mathbf{Q}_{\text{none}}$ participants (Row **XII**, $p = 3.7 \times 10^{-4}$). In addition, we also measured participants' in-document mouse activities, i.e., the number of scrolls while reading one document (Row **XV**). We observed that participants had more scrolls when presented with adjunct questions, indicating more concentrated reading behaviour. *These results confirm that adjunct questions significantly impact participants' behaviour, which is consistent with findings from [80].*

Effects of Adjunct Questions on Participants' Learning Gains

Recall that we evaluated participants' learning outcomes with RPL. Figure 6.4a shows that in both the post- and delay-test, the RPL of $\mathbf{Q}_{\text{none}}$ participants was higher than that of participants who received adjunct questions ($\mathbf{Q}_{\text{synthesis}}$, $\mathbf{Q}_{\text{random}}$, and $\mathbf{Q}_{\text{term}}$) (Post-test: $M. = 0.23$ vs. $M = 0.17, p = 0.026$, Delay-test: $M. = 0.17$ vs. $M. = 0.12, p = 0.14$, where $M.$ represents the Mean value). Figure 6.4b shows a detailed comparison broken down to all conditions. The $\mathbf{Q}_{\text{none}}$ and the $\mathbf{Q}_{\text{term}}$ participants had similar short-term retention ($M. = 0.23$ vs. $M. = 0.23$). Both were significantly higher than the $\mathbf{Q}_{\text{random}}$ condition ($M. = 0.23$ vs $M. = 0.13, p = 0.02$) and higher than $\mathbf{Q}_{\text{synthesis}}$ ($M. = 0.24$ vs $M. = 0.14, p = 0.06$). The delay-test RPL (Row **IV**) reflects the long-term learning outcomes. $\mathbf{Q}_{\text{random}}$ exhibited the worst results; $\mathbf{Q}_{\text{term}}$ was close to $\mathbf{Q}_{\text{none}}$. Previous work like [80] showed that participants spent substantially more time reading the same reading materials when presented with adjunct questions. Recall that participants had limited task time, and in the adjunct question conditions, participants read significantly fewer documents, which can partly explain the negative effects of adjunct questions. This is also aligned with an earlier classroom study of

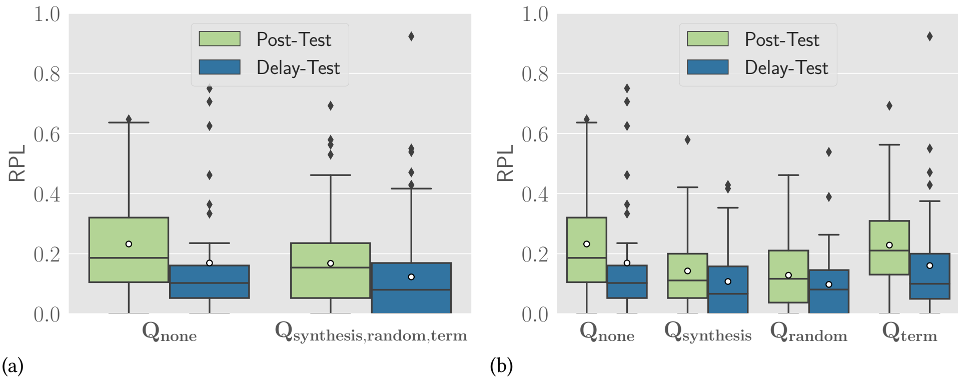


Figure 6.4: Distribution of post-test and delay-test RPL scores (a) w/ vs. w/o adjunct questions, (b) with all conditions.

adjunct questions [3], where the length of the task time is an essential factor in learning outcomes.

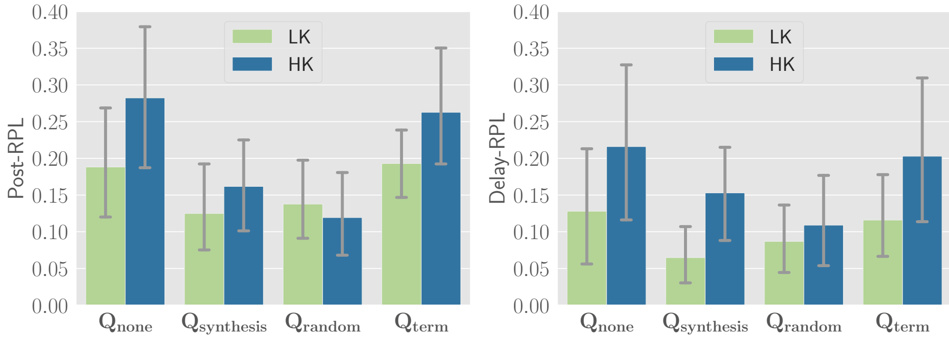
To sum up, our study revealed that participants who received adjunct questions exhibited more fine-grained reading behavior but had lower retention. However, when appropriate questions were posed, these participants achieved comparable short-term and long-term learning gains while reading significantly fewer documents. This highlights the importance of understanding learners' knowledge requirements and time constraints for presenting adjunct questions.

6.5.2 Factors that Influence Automatically Generated Adjunct Questions' Effects

Impacts of Question Types

Syed *et al.* [80] found that participants spent more time reading with additional synthesis questions while having similar learning gains with those who received only factoid questions. As shown in Row **XIV** of Table 6.4, compared to participants in factoid question conditions ($\mathbf{Q}_{\text{random}}$ and $\mathbf{Q}_{\text{term}}$), $\mathbf{Q}_{\text{synthesis}}$ participants had significantly longer average document dwell time (183s, $p < 0.05$). Similarly, $\mathbf{Q}_{\text{synthesis}}$ participants executed significantly more scrolls (Row **XV**, $M. = 98$, $p < 10^{-4}$) and spent more time reading before answering the adjunct question (Row **XVII**, $M. = 116s$ vs. $M. = 98s$ and 90s respectively) than participants who received factoid questions. Additionally, participants in the $\mathbf{Q}_{\text{synthesis}}$ condition produced significantly longer answers for adjunct questions (Row **XVI**, $p < 10^{-18}$) and spent the longest time writing their answers (Row **XVIII**, $p < 10^{-9}$). These results indicate that compared to factoid questions, the generated synthesis questions cause more cognitive burden. $\mathbf{Q}_{\text{synthesis}}$ participants have to spend more time reading, rewinding, and writing answers, which aligns with the previous work [80].

Regarding the learning outcomes, $\mathbf{Q}_{\text{synthesis}}$ participants had similar RPL to $\mathbf{Q}_{\text{random}}$ in both post-test ($M. = 0.14$ vs. $M. = 0.13$) and delay-test ($M. = 0.11$ vs. $M. = 0.10$). In addition, we also measured participants' learning with essay writing (at least 100 words) in the post-



(a) (b)

Figure 6.5: Distribution of post-test (a) and delay-test (b) RPL scores of HK and LK participants with all conditions.

test. Specifically, we consider T-Depth (measuring the number of subtopics covered) and F-Fact (measuring the number of atomic facts). As seen in Table 6.4 (Row **VI** and Row **VII**), $\mathbf{Q}_{\text{synthesis}}$ participants exhibited the highest T-Depth and F-Fact scores among all adjunct question conditions. These results showed that although essays created by participants in $\mathbf{Q}_{\text{synthesis}}$ were the most difficult to read (with the lowest Flesch reading ease score of 49.67), they covered a comparable fraction of assigned topic terms and provided better topic coverage and a greater number of facts. Thus, we conclude that compared with factoid questions, synthesis questions may cause a higher cognitive burden and higher performance on tests (i.e., essay writing) that require higher cognitive complexity than factoid questions regarding random text spans.

Impacts of Question Target Selection

As target selection is an essential procedure for generating questions, especially factoid questions, we investigate the effects of question target selection via the conditions $\mathbf{Q}_{\text{random}}$ and $\mathbf{Q}_{\text{term}}$. Recall that the answers (for which to generate questions) in $\mathbf{Q}_{\text{random}}$ were text spans extracted from the document and the answers for $\mathbf{Q}_{\text{term}}$ were the vocabulary terms of the assigned topic. To this end, we collected 377 document viewings in condition $\mathbf{Q}_{\text{term}}$, 63.4% of which targeted the assigned topic's terms. 58.8% of all topic terms were covered. As seen in rows from **IX** to **XV** of Table 6.4, participants in $\mathbf{Q}_{\text{random}}$ and $\mathbf{Q}_{\text{term}}$ did not show significant differences in the activity measures, although $\mathbf{Q}_{\text{random}}$ spent more time between queries on average. When it came to answering the questions, we found $\mathbf{Q}_{\text{term}}$ participants spent less time reading before writing answers (Row **XVII**, $M. = 89.73$ vs. $M. = 98.32$, $p = 0.89$) and spent longer time writing answers (Row **XVIII**, $M. = 13.49$ vs. $M. = 9.82$, $p = 0.64$) than $\mathbf{Q}_{\text{random}}$, but these differences were also not statistically significant. In contrast, $\mathbf{Q}_{\text{term}}$ participants' answers to adjunct questions showed significantly lower quality in terms of the F1 score (Row **XIX**, $M. = 0.39$ vs. $M. = 0.58$, $p < 10^{-4}$) and EM score (Row **XX**, $M. = 0.3$ vs. $M. = 0.58$, $p < 10^{-4}$). This may be due to the different complexity of the target answers. Vocabulary terms of the assigned topic tend to be more complex than the randomly chosen answer spans, such as named entities in the document. Notably, $\mathbf{Q}_{\text{term}}$ participants had better short-term learning outcomes (Row **III**, Post-test

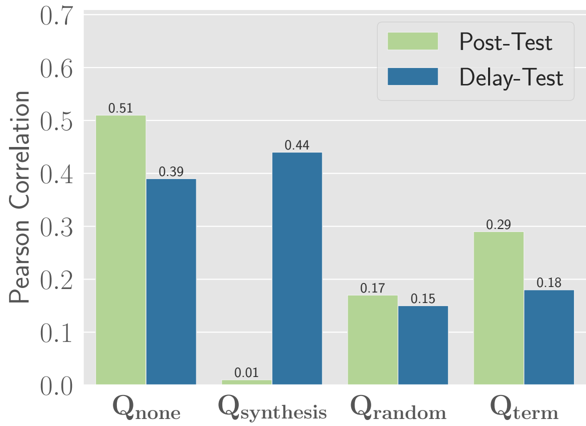


Figure 6.6: The Pearson correlation values between RPL and the average pre-test score in post-test and delay-test.

RPL, $M.=0.23$ vs. $M.=0.13$, $p=0.02$) and long-term retention (Row **IV**, Delay-test RPL, $M=0.16$ vs. $M=0.10$, $p=0.06$) than $\mathbf{Q}_{\text{random}}$ participants. We also found that $\mathbf{Q}_{\text{term}}$ participants had better yet not significant essay quality in all evaluated measures compared to $\mathbf{Q}_{\text{random}}$ participants. These observations suggest that compared with random target answer selection, guiding the users according to their learning goals achieves significantly better learning gains despite similar observed search behaviour, indicating the importance of learning goal-aware adaptive AQG for adjunct questions.

6

Impacts of Prior-Knowledge

Participants' prior knowledge may influence their behaviour, like the reading time [268] and their ability to identify the answer without reading. Thus, the effects of adjunct questions are sensitive to participants' prior knowledge levels [80] and may cause contrasting effects. In this chapter, we considered a participant as high-knowledge (HK) for a topic if her average pre-test score was higher than the median and otherwise low-knowledge (LK). We classified 71 participants as HK participants and 73 as LK participants. Table 6.5 reports the results and statistics along the evaluation metrics of the LK participants, and Table 6.6 the HK participants. Figure 6.5 compares the RPL of HK and LK participants in each condition in post-test (Figure 6.5a) and delay-test (Figure 6.5b). On average, HK participants exhibited higher RPL scores in all conditions during both the post-test and delay-test except $\mathbf{Q}_{\text{random}}$ in the post-test. Specifically, in the delay-test, $\mathbf{Q}_{\text{synthesis}}$ HK participants had significantly higher RPL than the LK group ($M.=0.15$ vs. $M.=0.06$, $p=0.044$). Moreover, compared with the RPL decrease from the post-test to the delay-test in other conditions, $\mathbf{Q}_{\text{synthesis}}$ HK participants show a slighter RPL decrease ($0.16 \Longrightarrow 0.15$). These results indicate that synthetic questions that require higher-level cognition lead to better long-term retention for more knowledgeable learners than other conditions.

We further investigated the correlation between the learning outcomes (both short-term and long-term) measured by RPL and participants' prior-knowledge measured by the pre-test vocabulary knowledge evaluation. Figure 6.6 shows the Pearson correlation scores. First, across all conditions, both post- and delay-test RPL scores were positively related to participants' prior knowledge. Furthermore, we found that $\mathbf{Q}_{\text{none}}$ participants' post-test RPL was strongly correlated with their prior-knowledge. However, adjunct questions mitigated the correlation, particularly in $\mathbf{Q}_{\text{synthesis}}$ condition, where there was no correlation between the post-test RPL and prior-knowledge. Lastly, we noted that the correlation with prior-knowledge was generally weaker in the delay-test than in the post-test, except in the $\mathbf{Q}_{\text{synthesis}}$ condition, where in contrast participants exhibited a much stronger correlation in the delay-test than in the post-test. These findings indicate the importance of adapting different AQG strategies based on learners' prior-knowledge levels.

6.6 Limitations

The experiment setup of the study on the automatically generated questions' effects on human learning behaviour and learning outcomes is constrained by specific learning resources and a limited time frame. This experiment setup may impede the generalizability of the results to informal learning by search contexts, where learners have access to a plethora of web resources and the duration of learning varies.

6.7 Conclusions

This chapter explored the effects of automatically generated adjunct questions in the complex search as the learning scenario through a user study on the customized open-source searching system. The empirical results confirm previous findings adjunct questions significantly influence participants' behaviour and learning outcomes, though in our study these effects vary across different conditions. We found evidence that with adjunct questions, participants were more engaged with the search results than those without adjunct questions. As a potential consequence of longer reading time, adjunct questions may negatively affect learning gains if the learning time is the same across all conditions. Furthermore, our results demonstrate the importance of adopting different types of adjunct questions for learning tasks with different cognitive complexity. Selecting targeting answers for adjunct questions according to participants' learning goals can significantly improve participants' learning outcomes. Lastly, we found participants' prior-knowledge had essential impacts on their learning gains, especially when they were posed with adjunct questions that required higher cognitive levels. Adjunct questions may mitigate the correlation between learning outcomes and the prior-knowledge. In future work, we aim to move beyond vocabulary knowledge tests and incorporate a wider variety of adjunct questions.

7
Conclusion

In this thesis, we examined automatic question generation from three aspects: (i) the transformers-based question generation methods, evaluation metrics, and domain adaptation, (ii) the noise reduction problem in dataset creation for question generation, and (iii) the effects of automatically generated questions on human learning. In this conclusion chapter, we first revisit the main research questions introduced in Chapter 1, and summarize the main findings. The main research questions we have addressed in this thesis are:

> **RQ1:** What metrics should be used for evaluating the quality of generated questions? How to compare the effectiveness of these metrics?
>
> **RQ2:** How to use deep neural models to facilitate dataset creation for question generation and reduce the noise of created datasets?
>
> **RQ3:** How does domain shift in context affect AQG and how to improve AQG performance on out-of-distribution unlabeled target domains?
>
> **RQ4:** How do automatically generated questions impact learners' behaviour and learning outcomes?

Then, later in this chapter, we discuss the recent progress in question generation research and future research directions in question generation and its applications in Section 7.2.

7.1 Main Findings

In this section, we summarize the main research findings.

7.1.1 Question Quality Evaluation Metrics and The Effects on Question Generation

In order to answer **RQ1**, in Chapter 2, we evaluated the question evaluation metrics by using them as rewards for RL-based AQG model training and comparing their effects on training the question generation model. We designed a common framework that provides a fair testbed that consisted of a Seq2Seq question generation model for paragraph-level question generation and a reward evaluator. We systematically categorized existing rewards into four categories, namely *Fluency, Similarity, Answerability, and Relevance*, and further proposed three novel evaluation metrics. We then implemented all metrics based on the same base model and directly optimized the AQG model with these metrics as the rewards. We performed a thorough empirical evaluation of their effects on model quality. We first confirmed the effectiveness of applying reinforcement learning with the evaluation metrics as rewards on AQG model training. We then showed that optimizing on one metric always led to the improvement of the corresponding evaluation metric, and that the magnitude of improvement differed significantly. We then examined the correlation of the rewards' effects and found that rewards within the same category were strongly correlated, and the fluency reward was not correlated with others. These results can guide the selection of rewards since adopting different uncorrelated rewards could evaluate the generations from diverse aspects and provide comprehensive feedback. This study also showed that the lexical method Meteor was strongly correlated with the BERT-based similarity metrics, especially BERTScore, and achieved the best human evaluation score on syntax, implying that the learned similarity metrics like BERTScore may focus on lexical similarity. Their effects need further examination in future research. In addition, using answerability evaluation metrics as rewards performed best, highlighting the importance of considering both the context and the target answer for evaluation question quality evaluation.

7.1.2 Dataset Creation

In order to answer **RQ2**, we provided two studies to investigate automatic noise reduction in dataset creation, in Chapter 3 and Chapter 4.

We began by examining the noise reduction in datasets created by collecting discussions in MOOC forums, with a particular case on video clip recommendations for MOOC forum questions in Chapter 3. Discussion forums serve as the principal communication medium among MOOC learners and instructors, compensating for the absence of physical interaction inherent in MOOCs. Despite the importance, MOOC discussion forums suffered from the information overload issue because of unuseful questions and unstructured and unorganized discussions with context scattering in the videos.

We first classified useful questions from noisy, unuseful questions. Then, we created a novel dataset of about 274K MOOC forum discussions and video transcripts from 6 topics to address the lack of training data issue. Unlike the previous research, we introduced dense retrieval-based methods to the MOOC forum video clip recommendation research problem and systematically compared their performance in extracting related video clips. We found that the pre-trained dense rankers achieve good effectiveness by fine-tuning with limited labeled data, and cross-encoders perform best. We further developed a distant supervision method and showed that the effectiveness of the video clip recommendation improved by 2.9% by fine-tuning with weakly labeled data created by cross-encoders over fine-tuning only with labeled data. This research demonstrated the framework for data preparation, useful question classification, clip recommendation, and distant supervision for creating datasets from MOOC forums. The results can be further used in future research on extracting context for creating discussion questions and thread recommendations for learners, providing a base for further research over MOOC forums.

Besides questions and answers created by learners and instructors in MOOC discussion forums, another prevalent method for creating QA datasets is crowdsourcing which employs crowd workers instead of domain experts for data annotation following specific instructions. However, crowdsourced data often contain a substantial amount of noise, and it is critical to infer the true label from crowd workers' noisy labels. Therefore, we then studied the QA dataset creation by crowdsourcing in Chapter 4. Answers are essential for training and evaluating both QA and AQG models. Previous research has shown the difficulty and adopted a multi-stage paradigm in aggregating crowdsourced answers. Therefore, we focused on aggregating multiple answer annotations for extractive question-answering datasets. We proposed a novel answer aggregation method that considered both the answer annotations' contextual representations and the quality of individual answers. Instead of treating the answer annotations as free text, we encoded the passage containing the answer with transformers-based PLM. We then used the mean of answer tokens' embeddings as the answer annotation's contextual representation. We then inferred the true answer annotation in two ways: (i) the annotation's similarity to the mean contextual representation of all answer annotations and (ii) the annotation's sum of similarity with other answer annotations. The effectiveness of these methods implies the crowdsourced answers contained common knowledge or contextual information that can be further explored. In addition to the relative similarity among the answer annotations, the quality of each answer annotation can also be evaluated based on whether it can answer the question. Therefore, we further evaluated the answer annotations using the Natural Language Inference (NLI) model and the QA model and proposed the algorithm to perform answer aggregation by joint consideration of the answer quality and the contextual representation. We evaluated the proposed method on several QA datasets. We showed that the automatic answer aggregation method can effectively infer the correct answer annotation compared to multi-stage human selection. The QA model trained on the automatically inferred answer annotation achieved similar performance with the human-selected answer annotations. We used the automatic answer aggregation method as a voter in addition to human experts for answer aggregation. The cleaned crowdsourced data led to 1.15% higher QA F1 performance than the crowdsourced data.

7.1.3 Domain Adaptation for Question Generation

We studied the domain adaptation problem of AQG models in Chapter 5 for answering **RQ3**. Our results first showed that the AQG performance degraded by half due to the domain shifts between the inputs to the neural AQG model and the training data, which prevented applying AQG models to low-resource domains. Then, we demonstrated that BERT-based context representation can be used for robust domain data clustering. We proposed an answer-type (including *time, location, numeric, person, and noun, etc.* based on the NER and POS tagger.) aware pseudo-in-domain data selection method based on their distance to the domain clustering center. Then, we proposed to re-train the AQG model with the pseudo-in-domain data and pseudo-labeled data with self-training. Our results showed that pseudo-in-domain data selection and self-training with filtering can effectively help improve and generalize the AQG model on out-of-distribution domains.

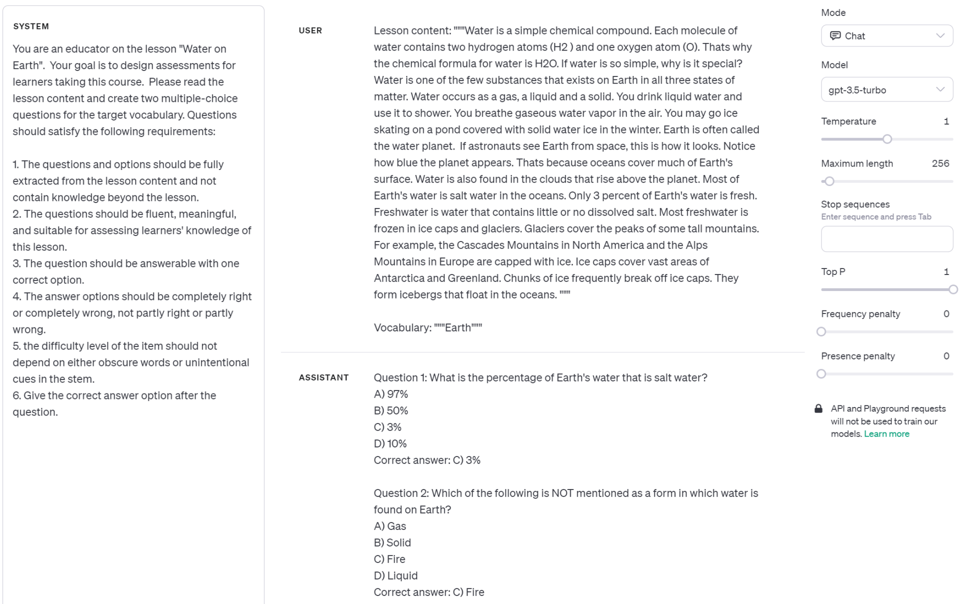
7.1.4 Effects of Automatically Generated Questions on Human Learning

Finally, we presented one study in Chapter 6 to better understand the effects of automatically generated questions on users' learning behaviour and learning outcomes in order to answer **RQ4**. Specifically, we investigated the impacts of automatically generated questions that were adjunct to the snippets users clicked and read. We implemented the search and reading interface, the AQG pipeline for generating factoid and synthesis types of questions, and then performed a user study that involved 144 participants. In contrast to prior research [80] which showed adjunct questions' potential long-term benefit for low prior-knowledge learners in the reading comprehension scenario, we found participants who did not receive questions achieved better learning outcomes than those with random or synthesis questions. Our study showed that the search phase has critical impacts on learning outcomes, and the conclusion in prior research cannot translate to the scenario with the search phase. The results further revealed questions' significant influence on learners' behaviour. Specifically, receiving adjunct questions resulted in longer dwell time and the number of mouse scrolls over the opened documents than participants who did not receive adjunct questions. In addition, we investigated the impacts of the questions' characteristics, including question types (factoid or synthesis) and the selection of questioning targets. The results showed that synthesis questions can cause a higher cognitive burden and higher performance on tests that require higher cognitive complexity, such as easy writing. However, synthesis questions cannot improve learning gains concerning vocabulary knowledge. Compared to random factoid questions and synthesis questions, questions targeting the learning goals can significantly improve learning gains. Besides, our study found participants' prior knowledge has critical impacts on learning gains. On average, participants with high prior knowledge had higher learning gains. The results of this study showed that it is arguable whether to pose adjunct questions for learners in the SAL scenario, especially with limited learning time, and it is important to jointly consider both the learning goals and users' prior knowledge to decide when and what to ask with the adjunct questions.

7.2 Future Directions

Currently, with hundreds of billions of parameters and pre-trained with massive amounts of data, the LLMs such as GPT-3 [43], GPT-4 [269], and LLaMA [270], have demonstrated not only the ability to produce semantically correct and coherent texts but also the ability to follow humans instructions after being fine-tuned with humans feedbacks towards aligning the model's responses to human's preferences [271]. Despite the advantages, LLMs are still prone to limitations such as the tendency to generate outputs that are not faithful to the source context [272], or producing hallucinated content that deviates from factual accuracy [273].

The progress in LLMs has brought both great opportunities and challenges for question generation research and application. For example, as shown in Figure 7.1, we ask GPT-3.5/4 to create assessment questions for the lesson "Water on Earth" about the vocabulary term *Earth*. It is notable that this example demonstrates a simple reading comprehension problem instead of complex problems that require multiple reasoning steps or mathemat-



(a)

(b)

Figure 7.1: An illustration of creating questions for the course materials with GPT-3.5 and GPT-4. We use OpenAI's playground interface[1] with results obtained on 18-10-2023 for illustration. In the system prompt, we specify the guidelines for creating multiple-choice questions and the option items [274]. The lesson content and vocabulary terms are taken from a textbook QA dataset [275].

ical reasoning. Both GPT-3.5 and GPT-4 can correctly follow the instructions and create two clear and concise multiple-choice questions to assess learners' understanding of water on Earth. On the one hand, they can create correct questions with suitable options for users to choose from, even without any in-context examples, like the second question in Figure 7.1a and the second question in Figure 7.1b. On the other hand, we can observe the factual inconsistency in the first question in Figure 7.1a where 3% of Earth's water should be fresh instead of salt water. We can also observe the inaccurate answer option in the first question in Figure 7.1b where the correct explanation should be "oceans cover much of Earth's surface".

Based on the research findings of this thesis and the recent progress in LLMs, we provide the following possible future directions towards explainable and reliable automatic question generation and evaluation and effects of applying AQG on education.

7.2.1 Human-in-the-Loop Assessment Generation

Creating assessments for learning involves instructors, learners, and AI tools [276]. As we show in Chapter 6, the types and targets of questions can significantly influence users' behaviour and learning outcomes. The instructors' expertise in the subject matter, objective design, and assessment strategies is crucial for creating effective assessments [277]. LLMs have good potential in generating assessment questions. Still, the questions may not align with the instructional goals or may not be optimally challenging for the learners, and there are concerns about their reliability, like the hallucination problem that may confuse learners. Therefore, an important research area is combining human experts' knowledge and LLMs systems for creating questions for learners. This collaborative approach could potentially harness the strengths of both entities, with AI systems generating a bulk of questions swiftly and human experts selecting and editing them to ensure alignment with learning objectives and to foster deeper understanding and critical thinking among learners. The human-AI collaboration can happen by actively involving humans in designing proper prompts for LLMs, such as the human-in-the-loop through chain-of-thought [278]. The collaboration can also happen by evaluating the quality of questions and filtering the potential low-quality questions with toxic, hallucinatory, fact-inconsistent text. LLMs often present long responses, making it difficult for people to comprehend, evaluate, and interact with them [279]. It is necessary to conduct Human Computer Interaction (HCI) research on designing new human-in-the-loop systems in this regard to explore the interface and the procedures for this collaboration [280], aiming to develop a synergistic model that maximizes the benefits of both human expertise and AI capabilities in the domain of educational question generation.

7.2.2 Explainable Question Quality Evaluation

As mentioned in Chapter 2, automatic question quality evaluation has been extensively studied, especially in recent years as researchers have endeavored to design metrics for evaluating the questions' *Grammaticality, Relevance*, and *Answerability* with pre-trained language models. One line of research focuses on better evaluating the generations' similarity compared with human-curated reference questions such as BARTScore [72] and PRISM [73]. Another line of research focuses on reference-free metrics that use the learned models' ability to score the generated questions directly without references to mimic human judgments on question quality, such as QuestEval [74], QRelScore [77] and UniEval [79]. However, these metrics are limited by the backbone language models and the data used for training them. First, it is necessary to study whether or not the metrics computed by cheap, smaller models like BERT can provide appropriate evaluation for the questions generated by LLMs. Second, when the pre-trained models are fine-tuned for the evaluation tasks (most are classification tasks), these models may rely on spurious attributes that are correlated to the labels in the training data [281, 282]. The learned metrics tend to be brittle and vulnerable to Out of Distribution (OOD) data and changes in the spurious attributes. Despite the research on spurious attributes detection and removal for better OOD generalization in NLP keeps attracting a lot of attention [283], such as from the data augmentation by retrieval and counterfactual generation [284–288], the OOD problem in question generation evaluation metrics is understudied. Therefore, it requires further research on both the training reliable and robust evaluation metrics and evaluating the trustworthiness of the evaluation metrics. Third, although current metrics can provide numeric scores on dimensions such as fluency, relevance, and answerability, they rely on black-box language models and cannot provide reasoning and explanation for the scores, raising concerns about the credibility of the evaluation results and the ability to guide generation selection. Therefore, in addition to quantitative metrics, future metrics should take explainability into account. The explanations can have a direct influence on improving the AQG systems. For example, counterfactual examples to the generated questions, which have minimum edition but higher evaluation scores, can provide rich information on how to select the generated questions or further improve the AQG method. The most recent research [289] discussed key properties and the latest state-of-the-art approaches to explainable metrics based on generative LLMs for machine translation. The question evaluation involves distinct evaluation criteria and requires further research.

7.2.3 Data Augmentation for Question Generation with LLMs

Training question generation models requires high-quality labeled data. With recent advantages, LLMs exhibit the potential to perform commonsense reasoning over texts and the capability to generate high-quality synthetic data through in-context examples and even zero-shot learning [290]. Recent research has explored the methods and effects of creating synthetic augmenting datasets for downstream tasks such as NLI [287] and sentiment analysis [291]. The results show that the smaller classification models can benefit from knowledge distilled from augmentation data and achieve better Domain Adaptation (DA) performance with better generalization to OOD data. In addition to creating synthesis data by perturbation inputs and creating soft labels, the LLMs are capable of further creating high-quality novel context as well as question-answer pairs, which can be critical for specific applications and domains where the unlabeled contexts are also scarce. As Samuel *et al.* [292] demonstrates, the generated synthesis augment data can be used to improve downstream smaller QA models' performance on low-resource domains. The previous research demonstrates great opportunities in data augmentation with LLMs for AQG. However, on the one hand, some current methods of data augmentation question generation, such as [293, 294] are based on smaller PLMs which lacks the emergent abilities of LLMs. On the other hand, less has been investigated in the literature concerning applying LLMs in the online learning context. It would be a promising line of research to explore the effective data augmentation methods with LLMs to distill LLMs' knowledge for downstream applications, for example, data augmentation for adapting to course domains or data augmentation for personalized AQG considering learners' profile such as their knowledge level and education goals.

LLMs may also suffer from limitations such as the lack of up-to-date knowledge not seen in the training corpora and lack of correct memorization of long-tail sparse long-tail knowledge in their parameters [295]. An active research area in language generation with LLMs is retrieval-augmented language modeling (RALM) [296] or retrieval-augmented generation (RAG) [297, 298], which combines the LLM and the external memory with relevant documents retrieved from the exterior knowledge sources. The language models are grounded in the relevant context of generation. These methods consist of two major high-level components: document retrieval, which is about selecting the most relevant documents while alleviating irrelevant noises, and document reading, which is about incorporating the retrieved document into the generation process. In this way, the generation results are more attributable and can achieve better results even with the frozen language models. To this end, in the AQG research area, these two major procedures would involve determining contexts that are question-worthy and incorporating them into the prompts for the LLM effectively for reliable and more transparent application [299]. These methods can also be adopted for mitigating the domain adaptation [298, 300] issue.

7.2.4 Question Generation over Heterogeneous Sources

The assessment design should be centered on supporting learners' learning while keeping learners engaged and motivated, starting by understanding learners' knowledge states and the intrinsic difficulties in making the inferences necessary [301]. As the results in Chapter 6, learners with different levels of prior knowledge demonstrates different reflection on the generated questions. It is critical to adapt the questions posed to learners based on their background knowledge and learning progress [7, 302]. Knowledge tracing [280, 303, 304] is commonly used to predict learners' knowledge mastery based on their interaction with questions, which cannot utilize rich information in the learning environment. For example, in MOOCs, there is rich content like video clips, related textbook materials, and discussion forums. Beyond these contents related to knowledge, the logs of learners' activities also contain rich information on learners' knowledge states, interactions among learners, learning progress, etc. [305]. Furthermore, the social communities and social networks that arise around the courses and the learners are also essential to understanding learners' background knowledge, fostering participation and peer support [306–308]. Therefore, it would be essential to *learn from heterogeneous sources* to determine when and what questions should be posed to learners in order to help learners learn and evaluate the knowledge within different modalities. To achieve this, future research also needs to investigate how to incorporate the knowledge tracing model with AQG in order to develop AI systems that are capable of understanding and adapting to individual learners' profiles, thereby personalizing the learning experience and promoting more effective learning outcomes. AQG based on heterogeneous sources could pave the way for more nuanced and effective learning strategies, fostering a richer and more engaging learning environment for students.

7.2.5 The Effects of on-the-fly Constant Evaluation and Feedback on Learners

The rapid progress of LLMs has great potential impacts on education. GPT-3.5 and its successor GPT-4 are reportedly to pass qualification exams designed for humans, such as the neurosurgery practice board exams at rates comparable to neurosurgery residents [309], indicating traditional assessment methods could be easily cheated by LLMs, especially in MOOCs where physical restrictions are absent to prevent students from using LLMs. The easy access to LLM tools may have the danger of making learners less likely to learn. One way to address this challenge is to deploy *formative assessment* [310], i.e., provide learners questions periodically to help space their studying and give them feedback about what they know and do not know [311], by taking advantage of AQG and LLMs to conduct constant evaluation and feedback on-the-fly [312]. By further incorporation with fine-grained user activity monitoring and knowledge tracing.

In order to design such systems, firstly, it is essential to investigate how to incorporate AQG tools in the education systems, such as the impacts of the position of showing the interaction tools, the formats of assessments, and the frequency of engaging learners with assessments. Further, it would be essential to investigate what are the LLM tools' impacts on human learning and what are the effects of constant evaluation and feedback in order to provide empirical guidelines for building future education systems. Despite some most recent works have proposed to investigate some aspects of applying formative assessments with LLMs, such as the differences in learners learning gains between ChatGPT and human tutor curated hints [313], the effects of applying ChatGPT for evaluating learners' responses and creating feedbacks automatically [314–316]. The results suggest LLMs' high assess accuracy and high feedback quality, and also the struggle with decimal values and the potential over-reliance of LLMs. These results emphasize the importance of further investigating LLMs' effectiveness and effects.

To sum up, LLMs would play an essential role in the future systems designed for learning. It requires significant research efforts to build safe, trustworthy, intelligent, and personalized educational systems with them. These efforts are two-fold. On the one hand, new LLMs and tools are necessary to understand knowledge and learners better. On the other hand, it is critical to focus on the interaction and collaboration between LLMs and humans, such as better interaction interfaces and pipelines, to take humans to the center of learning material creation, evaluation, consumption, and effects tracking.

Bibliography

References

[1] A. Gates, *Recitation as a factor in memorizing*. Archives of Psychology (1917).

[2] B. E. Johnson, *The effect of written examinations on learning and on the retention of learning,* The Journal of Experimental Education **7**, 55 (1938).

[3] C. Hamaker, *The effects of adjunct questions on prose learning,* Review of educational research **56**, 212 (1986).

[4] D. Shah, *3 million sign-ups, 100,000+ courses: Class central's 2022 year in review,* (2023).

[5] D. Shah, *By the numbers: Moocs in 2021,* (2022).

[6] D. Shah, *2022 year in review: The "new normal" that wasn't,* (2023).

[7] I. Rushkin, Y. Rosen, A. M. Ang, C. Fredericks, D. Tingley, M. J. Blink, and G. Lopez, *Adaptive assessment experiment in a harvardx mooc.* in *EDM* (2017).

[8] B. S. Bloom, P. Airasian, K. Cruikshank, R. Mayer, P. Pintrich, J. Raths, and M. Wittrock, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Pearson, 2001).

[9] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, *A systematic review of automatic question generation for educational purposes,* International Journal of Artificial Intelligence in Education **30**, 121 (2020).

[10] M. Heilman and N. A. Smith, *Good question! statistical ranking for question generation,* in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2010) pp. 609–617.

[11] M. Heilman, *Automatic factual question generation from text,* Language Technologies Institute School of Computer Science Carnegie Mellon University **195** (2011).

[12] D. Lindberg, F. Popowich, J. Nesbit, and P. Winne, *Generating natural language questions to support learning on-line,* in *Proceedings of the 14th European Workshop on Natural Language Generation* (2013) pp. 105–114.

[13] X. Du, J. Shao, and C. Cardie, *Learning to ask: Neural question generation for reading comprehension,* arXiv preprint arXiv:1705.00106 (2017).

[14] D. Tang, N. Duan, T. Qin, Z. Yan, and M. Zhou, *Question answering and question generation as dual tasks,* arXiv preprint arXiv:1706.02027 (2017).

[15] Z. Wang, A. S. Lan, W. Nie, A. E. Waters, P. J. Grimaldi, and R. G. Baraniuk, *Qg-net: a data-driven question generation model for educational content,* in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (2018) pp. 1–10.

[16] Y. Zhao, X. Ni, Y. Ding, and Q. Ke, *Paragraph-level neural question generation with maxout pointer and gated self-attention networks,* in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018) pp. 3901–3910.

[17] B. Liu, M. Zhao, D. Niu, K. Lai, Y. He, H. Wei, and Y. Xu, *Learning to generate questions by learningwhat not to generate,* in *The World Wide Web Conference* (2019) pp. 1106–1118.

[18] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, *Unified language model pre-training for natural language understanding and generation,* in *Advances in Neural Information Processing Systems* (2019) pp. 13042–13054.

[19] B. Das, M. Majumder, S. Phadikar, and A. A. Sekh, *Automatic question generation and answer assessment: a survey,* Research and Practice in Technology Enhanced Learning **16**, 1 (2021).

[20] A. Rogers, M. Gardner, and I. Augenstein, *Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension,* ACM Computing Surveys **55**, 1 (2023).

[21] R. Mitkov *et al.*, *Computer-aided generation of multiple-choice tests,* in *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing* (2003) pp. 17–22.

[22] J. Mostow and W. Chen, *Generating instruction automatically for the reading strategy of self-questioning.* in *AIED* (2009) pp. 465–472.

[23] K. Mazidi and R. Nielsen, *Linguistic considerations in automatic question generation,* in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2014) pp. 321–326.

[24] A. R. Fabbri, P. Ng, Z. Wang, R. Nallapati, and B. Xiang, *Template-based question generation from retrieved sentences for improved unsupervised question answering,* in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020) pp. 4508–4513.

[25] B. Leite. and H. Cardoso., *Do rules still rule? comprehensive evaluation of a rule-based question generation system,* in *Proceedings of the 15th International Conference on Computer Supported Education - Volume 2: CSEDU,*, INSTICC (SciTePress, 2023) pp. 27–38.

[26] K. D. Dhole and C. D. Manning, *Syn-qg: Syntactic and shallow semantic rules for question generation,* arXiv preprint arXiv:2004.08694 (2020).

[27] X. Yuan, T. Wang, C. Gulcehre, A. Sordoni, P. Bachman, S. Subramanian, S. Zhang, and A. Trischler, *Machine comprehension by text-to-text neural question generation,* arXiv preprint arXiv:1705.02012 (2017).

[28] D. Tang, N. Duan, Z. Yan, Z. Zhang, Y. Sun, S. Liu, Y. Lv, and M. Zhou, *Learning to collaborate for question answering and asking,* in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (2018) pp. 1564–1574.

[29] X. Ma, Q. Zhu, Y. Zhou, X. Li, and D. Wu, *Improving question generation with sentence-level semantic matching and answer position inferring,* arXiv preprint arXiv:1912.00879 (2019).

[30] L. Murakhovs'ka, C.-S. Wu, P. Laban, T. Niu, W. Liu, and C. Xiong, *MixQG: Neural question generation with mixed answer types,* in *Findings of the Association for Computational Linguistics: NAACL 2022* (Association for Computational Linguistics, Seattle, United States, 2022) pp. 1486–1497.

[31] A. Ushio, F. Alva-Manchego, and J. Camacho-Collados, *Generative language models for paragraph-level question generation,* in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022) pp. 670–688.

[32] N. Duan, D. Tang, P. Chen, and M. Zhou, *Question generation for question answering,* in *Proceedings of the 2017 conference on empirical methods in natural language processing* (2017) pp. 866–874.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need,* in *Advances in neural information processing systems* (2017) pp. 5998–6008.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding,* arXiv preprint arXiv:1810.04805 (2018).

[35] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, J. Gao, S. Piao, M. Zhou, *et al.*, *Unilmv2: Pseudo-masked language models for unified language model pre-training,* in *International Conference on Machine Learning* (PMLR, 2020) pp. 642–652.

[36] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,* arXiv preprint arXiv:1910.13461 (2019).

[37] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, *Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training,* in *Proceedings of*

the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (2020) pp. 2401–2410.

[38] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, *Exploring the limits of transfer learning with a unified text-to-text transformer,* The Journal of Machine Learning Research **21**, 5485 (2020).

[39] R. Nogueira, J. Lin, and A. Epistemic, *From doc2query to docttttquery,* Online preprint (2019).

[40] L. Bonifacio, H. Abonizio, M. Fadaee, and R. Nogueira, *Inpars: Unsupervised dataset generation for information retrieval,* in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022) pp. 2387–2392.

[41] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, *Improving language understanding by generative pre-training,* (2018).

[42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, *Language models are unsupervised multitask learners,* OpenAI blog **1**, 9 (2019).

[43] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, *Language models are few-shot learners,* Advances in neural information processing systems **33**, 1877 (2020).

[44] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, *A survey of large language models,* arXiv preprint arXiv:2303.18223 (2023).

[45] S. Zhang and M. Bansal, *Addressing semantic drift in question generation for semi-supervised question answering,* in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019) pp. 2495–2509.

[46] Y. Chen, L. Wu, and M. J. Zaki, *Reinforcement learning based graph-to-sequence model for natural question generation, 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020,* (2020).

[47] P. Zhu, *Improving long content question generation with multi-level passage encoding,* in *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18* (Springer, 2021) pp. 140–152.

[48] P. Zhu and C. Hauff, *Evaluating bert-based rewards for question generation with reinforcement learning,* in *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (2021) pp. 261–270.

[49] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, *The curious case of neural text degeneration, 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020,* (2020).

[50] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, *Squad: 100,000+ questions for machine comprehension of text,* arXiv preprint arXiv:1606.05250 (2016).

[51] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, *Race: Large-scale reading comprehension dataset from examinations,* arXiv preprint arXiv:1704.04683 (2017).

[52] M. Gupta, N. Kulkarni, R. Chanda, A. Rayasam, and Z. C. Lipton, *Amazonqa: A review-based question answering task,* arXiv preprint arXiv:1908.04364 (2019).

[53] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, *Eli5: Long form question answering,* arXiv preprint arXiv:1907.09190 (2019).

[54] G. Chen, J. Yang, C. Hauff, and G.-J. Houben, *Learningq: a large-scale dataset for educational question generation,* in *Twelfth International AAAI Conference on Web and Social Media* (2018).

[55] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, *Hotpotqa: A dataset for diverse, explainable multi-hop question answering,* arXiv preprint arXiv:1809.09600 (2018).

[56] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, *et al.*, *Natural questions: a benchmark for question answering research,* Transactions of the Association for Computational Linguistics **7**, 453 (2019).

[57] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, *Ms marco: A human generated machine reading comprehension dataset,* in *CoCo@ NIPS* (2016).

[58] B. Plank, *What to do about non-standard (or non-canonical) language in nlp,* arXiv preprint arXiv:1608.07836 (2016).

[59] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, *Domain-specific language model pretraining for biomedical natural language processing,* ACM Transactions on Computing for Healthcare (HEALTH) **3**, 1 (2021).

[60] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, *A survey on deep transfer learning,* in *International conference on artificial neural networks* (Springer, 2018) pp. 270–279.

[61] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, *A kernel method for the two-sample-problem,* Advances in neural information processing systems **19**, 513 (2006).

[62] T. Hosking and S. Riedel, *Evaluating Rewards for Question Generation Models,* , 2278 (2019), arXiv:1902.11049 .

[63] P. Laban, C.-S. Wu, L. Murakhovs' ka, W. Liu, and C. Xiong, *Quiz design task: Helping teachers create quizzes with automated question generation,* arXiv preprint arXiv:2205.01730 (2022).

[64] V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan, *The first question generation shared task evaluation challenge,* in *Proceedings of the 6th International Natural Language Generation Conference* (2010).

[65] P. Nema and M. M. Khapra, *Towards a better metric for evaluating question generation systems,* arXiv preprint arXiv:1808.10192 (2018).

[66] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, *Bleu: a method for automatic evaluation of machine translation,* in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (2002) pp. 311–318.

[67] C.-Y. Lin, *Rouge: A package for automatic evaluation of summaries,* in *Text summarization branches out* (2004) pp. 74–81.

[68] M. Denkowski and A. Lavie, *Meteor universal: Language specific translation evaluation for any target language,* in *Proceedings of the ninth workshop on statistical machine translation* (2014) pp. 376–380.

[69] M. Stanojević and K. Sima'an, *Beer: Better evaluation as ranking,* in *Proceedings of the Ninth Workshop on Statistical Machine Translation* (2014) pp. 414–419.

[70] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, *MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance,* in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics, Hong Kong, China, 2019) pp. 563–578.

[71] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *Bertscore: Evaluating text generation with bert,* arXiv preprint arXiv:1904.09675 (2019).

[72] W. Yuan, G. Neubig, and P. Liu, *Bartscore: Evaluating generated text as text generation,* in *Advances in Neural Information Processing Systems*, Vol. 34, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021) pp. 27263–27277.

[73] B. Thompson and M. Post, *Automatic machine translation evaluation in many languages via zero-shot paraphrasing,* in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Online, 2020) pp. 90–121.

[74] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang, and P. Gallinari, *QuestEval: Summarization asks for fact-based evaluation,* in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021) pp. 6594–6604.

[75] C. Rebuffel, T. Scialom, L. Soulier, B. Piwowarski, S. Lamprier, J. Staiano, G. Scoutheeten, and P. Gallinari, *Data-QuestEval: A referenceless metric for data-to-text semantic evaluation,* in *Proceedings of the 2021 Conference on Empirical Methods*

in Natural Language Processing (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021) pp. 8029–8036.

[76] M. Deng, B. Tan, Z. Liu, E. Xing, and Z. Hu, *Compression, transduction, and creation: A unified framework for evaluating natural language generation,* in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021) pp. 7580–7605.

[77] X. Wang, B. Liu, S. Tang, and L. Wu, *QRelScore: Better evaluating generated questions with deeper understanding of context-aware relevance,* in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022) pp. 562–581.

[78] A. Mohammadshahi, T. Scialom, M. Yazdani, P. Yanki, A. Fan, J. Henderson, and M. Saeidi, *Rquge: Reference-free metric for evaluating question generation by answering the question,* arXiv preprint arXiv:2211.01482 (2022).

[79] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han, *Towards a unified multi-dimensional evaluator for text generation,* in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022) pp. 2023–2038.

[80] R. Syed, K. Collins-Thompson, P. N. Bennett, M. Teng, S. Williams, D. W. W. Tay, and S. Iqbal, *Improving learning outcomes with gaze tracking and automatic question generation,* in *Proceedings of The Web Conference 2020* (2020) pp. 1693–1703.

[81] O. H. Lu, A. Y. Huang, D. C. Tsai, and S. J. Yang, *Expert-authored and machine-generated short-answer questions for assessing students learning performance,* Educational Technology & Society **24**, 159 (2021).

[82] X. Fu, K. Lokesh Krishna, and R. Sabitha, *Artificial intelligence applications with e-learning system for china's higher education platform,* Journal of Interconnection Networks , 2143016 (2021).

[83] T. Steuer, A. Filighera, T. Tregel, and A. Miede, *Educational automatic question generation improves reading comprehension in non-native speakers: A learner-centric case study,* Frontiers in Artificial Intelligence **5** (2022).

[84] R. Van Campenhout, N. Brown, B. Jerome, J. S. Dittel, and B. G. Johnson, *Toward effective courseware at scale: Investigating automatically generated questions as formative practice,* in *Proceedings of the Eighth ACM Conference on Learning@ Scale* (2021) pp. 295–298.

[85] P. Zhu and C. Hauff, *Evaluating bert-based rewards for question generation with reinforcement learning,* in *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, edited by F. Hasibi, Y. Fang, and A. Aizawa (ACM, 2021) pp. 261–270.

[86] P. Zhu, C. Hauff, and J. Yang, *MOOC-Rec: Instructional video clip recommendation for MOOC forum questions,* in *Proceedings of the 15th International Conference on Educational Data Mining*, edited by A. Mitrovic and N. Bosch (International Educational Data Mining Society, Durham, United Kingdom, 2022) pp. 705–709.

[87] P. Zhu, Z. Wang, C. Hauff, J. Yang, and A. Anand, *Answer quality aware aggregation for extractive QA crowdsourcing,* in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022) pp. 6147–6159.

[88] P. Zhu and C. Hauff, *Unsupervised domain adaptation for question generation with DomainData selection and self-training,* in *Findings of the Association for Computational Linguistics: NAACL 2022* (Association for Computational Linguistics, Seattle, United States, 2022) pp. 2388–2401.

[89] Y. Wang, C. Liu, M. Huang, and L. Nie, *Learning to ask questions in open-domain conversational systems with typed decoders,* arXiv preprint arXiv:1805.04843 (2018).

[90] Q. Yu, L. Bing, Q. Zhang, W. Lam, and L. Si, *based question generation with adaptive instance transfer and augmentation,* arXiv preprint arXiv:1911.01556 (2019).

[91] H. Zamani, S. Dumais, N. Craswell, P. Bennett, and G. Lueck, *Generating clarifying questions for information retrieval,* in *Proceedings of The Web Conference 2020* (2020) pp. 418–428.

[92] R. Nogueira, W. Yang, J. Lin, and K. Cho, *Document expansion by query prediction,* arXiv preprint arXiv:1904.08375 (2019).

[93] K. Collins-Thompson, P. Hansen, and C. Hauff, *Search as learning (dagstuhl seminar 17092),* in *Dagstuhl reports*, Vol. 7 (Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017).

[94] P. Rajpurkar, R. Jia, and P. Liang, *Know what you don't know: Unanswerable questions for squad,* arXiv preprint arXiv:1806.03822 (2018).

[95] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, *Sequence level training with recurrent neural networks,* arXiv preprint arXiv:1511.06732 (2015).

[96] Z. Fan, Z. Wei, S. Wang, Y. Liu, and X. Huang, *A Reinforcement Learning Framework for Natural Question Generation using Bi-discriminators,* Coling , 1763 (2018).

[97] G. Yasui, Y. Tsuruoka, and M. Nagata, *Using semantic similarity as reward for reinforcement learning in sentence generation,* in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (2019) pp. 400–406.

[98] L. Song, Z. Wang, and W. Hamza, *A unified query-based generative model for question generation and question answering,* arXiv preprint arXiv:1709.01058 (2017).

[99] V. Kumar, G. Ramakrishnan, and Y.-F. Li, *A framework for automatic question generation from text using deep reinforcement learning,* arXiv preprint arXiv:1808.04961 (2018).

[100] Y. Xie, L. Pan, D. Wang, M.-Y. Kan, and Y. Feng, *Exploring question-specific rewards for generating deep questions,* arXiv preprint arXiv:2011.01102 (2020).

[101] T. Hosking and S. Riedel, *Evaluating rewards for question generation models,* arXiv preprint arXiv:1902.11049 (2019).

[102] B. Dhingra, D. Pruthi, and D. Rajagopal, *Simple and effective semi-supervised question answering,* arXiv preprint arXiv:1804.00720 (2018).

[103] B. Liu, H. Wei, D. Niu, H. Chen, and Y. He, *Asking questions the human way: Scalable question-answer generation from text corpus,* in *Proceedings of The Web Conference 2020* (2020) pp. 2032–2043.

[104] X. Yao, E. Tosch, G. Chen, E. Nouri, R. Artstein, A. Leuski, K. Sagae, and D. Traum, *Creating conversational characters using question generation tools,* Dialogue & Discourse **3**, 125 (2012).

[105] Y. Ling, F. Cai, H. Chen, and M. de Rijke, *Leveraging context for neural question generation in open-domain dialogue systems,* in *Proceedings of The Web Conference 2020* (2020) pp. 2486–2492.

[106] Y. Gao, P. Li, I. King, and M. R. Lyu, *Interconnected question generation with coreference alignment and conversation flow modeling,* arXiv preprint arXiv:1906.06893 (2019).

[107] J. Gu, M. Mirshekari, Z. Yu, and A. Sisto, *Chaincqg: Flow-aware conversational question generation,* arXiv preprint arXiv:2102.02864 (2021).

[108] I. Labutov, S. Basu, and L. Vanderwende, *Deep questions without deep understanding,* in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2015) pp. 889–898.

[109] G. Danon and M. Last, *A syntactic approach to domain-specific automatic question generation,* arXiv preprint arXiv:1712.09827 (2017).

[110] L. Ma and Y. Ma, *Automatic question generation based on mooc video subtitles and knowledge graph,* in *Proceedings of the 2019 7th International Conference on Information and Education Technology* (2019) pp. 49–53.

[111] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, *Self-critical sequence training for image captioning,* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017) pp. 7008–7024.

[112] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser, *Why we need new evaluation metrics for nlg,* arXiv preprint arXiv:1707.06875 (2017).

[113] J. Pennington, R. Socher, and C. D. Manning, *Glove: Global vectors for word representation,* in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014) pp. 1532–1543.

[114] E. Clark, A. Celikyilmaz, and N. A. Smith, *Sentence mover's similarity: Automatic evaluation for multi-sentence texts,* in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, 2019) pp. 2748–2760.

[115] Z. Yang, Z. Hu, C. Dyer, E. P. Xing, and T. Berg-Kirkpatrick, *Unsupervised text style transfer using language models as discriminators,* arXiv preprint arXiv:1805.11749 (2018).

[116] L. Dong, J. Mallinson, S. Reddy, and M. Lapata, *Learning to paraphrase for question answering,* arXiv preprint arXiv:1708.06022 (2017).

[117] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, *Ppdb: The paraphrase database,* in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2013) pp. 758–764.

[118] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, *Gptscore: Evaluate as you desire,* arXiv preprint arXiv:2302.04166 (2023).

[119] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, *Gpteval: Nlg evaluation using gpt-4 with better human alignment,* arXiv preprint arXiv:2303.16634 (2023).

[120] P. Adamopoulos, *What makes a great mooc? an interdisciplinary analysis of student retention in online courses,* (2013).

[121] A. Agrawal, J. Venkatraman, S. Leonard, and A. Paepcke, *Youedu: Addressing confusion in mooc discussion forums by recommending instructional video clips.* International Educational Data Mining Society (2015).

[122] D. A. Wiley and E. K. Edwards, *Online self-organizing social systems: The decentralized future of online learning,* Quarterly review of distance education **3**, 33 (2002).

[123] D. Yang, M. Wen, I. Howley, R. Kraut, and C. Rose, *Exploring the effect of confusion in discussion forums of massive open online courses,* in *Proceedings of the second (2015) ACM conference on learning@ scale* (2015) pp. 121–130.

[124] A. Ntourmas, N. Avouris, S. Daskalaki, and Y. Dimitriadis, *Evaluation of a massive online course forum: design issues and their impact on learners' support,* in *IFIP conference on human-computer interaction* (Springer, 2019) pp. 197–206.

[125] P. Trirat, S. Noree, and M. Y. Yi, *Intellimooc: Intelligent online learning framework for mooc platforms,* in *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)* (2020) pp. 682–685.

[126] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, *Dense passage retrieval for open-domain question answering,* arXiv preprint arXiv:2004.04906 (2020).

[127] O. Khattab and M. Zaharia, *Colbert: Efficient and effective passage search via contextualized late interaction over bert,* in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (2020) pp. 39–48.

[128] A. Ntourmas, S. Daskalaki, Y. Dimitriadis, and N. Avouris, *Classifying mooc forum posts using corpora semantic similarities: a study on transferability across different courses,* Neural Computing and Applications , 1 (2021).

[129] B. Kulis *et al.*, *Metric learning: A survey,* Foundations and Trends® in Machine Learning **5**, 287 (2013).

[130] B. Mitra, N. Craswell, *et al.*, *An introduction to neural information retrieval* (Now Foundations and Trends, 2018).

[131] M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil, *Efficient natural language response suggestion for smart reply,* arXiv preprint arXiv:1705.00652 (2017).

[132] J. Lin, R. Nogueira, and A. Yates, *Pretrained transformers for text ranking: Bert and beyond,* arXiv preprint arXiv:2010.06467 (2020).

[133] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, *Mpnet: Masked and permuted pre-training for language understanding,* Advances in Neural Information Processing Systems **33**, 16857 (2020).

[134] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, *Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,* Advances in Neural Information Processing Systems **33**, 5776 (2020).

[135] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks,* arXiv preprint arXiv:1908.10084 (2019).

[136] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization,* arXiv preprint arXiv:1412.6980 (2014).

[137] R. Nogueira and K. Cho, *Passage re-ranking with bert,* arXiv preprint arXiv:1901.04085 (2019).

[138] B. McFee and G. R. Lanckriet, *Metric learning to rank,* in *ICML* (2010).

[139] C. C. Aggarwal *et al.*, *Recommender systems*, Vol. 1 (Springer, 2016).

[140] C. Van Gysel and M. de Rijke, *Pytrec_eval: An extremely fast python interface to trec_eval,* in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018) pp. 873–876.

[141] C. Parsing, *Speech and language processing,* (2009).

[142] P. Lewis, L. Denoyer, and S. Riedel, *Unsupervised question answering by cloze translation,* arXiv preprint arXiv:1906.04980 (2019).

[143] Z. Zhang, J. Yang, and H. Zhao, *Retrospective reader for machine reading comprehension,* arXiv preprint arXiv:2001.09694 (2020).

[144] A. Rogers, M. Gardner, and I. Augenstein, *Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension,* ArXiv **abs/2107.12708** (2021).

[145] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, *Newsqa: A machine comprehension dataset,* arXiv preprint arXiv:1611.09830 (2016).

[146] A. Talmor, J. Herzig, N. Lourie, and J. Berant, *Commonsenseqa: A question answering challenge targeting commonsense knowledge,* arXiv preprint arXiv:1811.00937 (2018).

[147] G. Kazai, J. Kamps, and N. Milic-Frayling, *Worker types and personality traits in crowdsourcing relevance labels,* in *Proceedings of the 20th ACM international conference on Information and knowledge management* (2011) pp. 1941–1944.

[148] M. Geva, Y. Goldberg, and J. Berant, *Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets,* arXiv preprint arXiv:1908.07898 (2019).

[149] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, *Truth inference in crowdsourcing: Is the problem solved?* Proceedings of the VLDB Endowment **10**, 541 (2017).

[150] R. J. Passonneau and B. Carpenter, *The benefits of a model of annotation,* Transactions of the Association for Computational Linguistics **2**, 311 (2014).

[151] H. Lakkaraju, J. Leskovec, J. Kleinberg, and S. Mullainathan, *A bayesian framework for modeling human evaluations,* in *Proceedings of the 2015 SIAM International Conference on Data Mining* (SIAM, 2015) pp. 181–189.

[152] A. T. Nguyen, B. C. Wallace, J. J. Li, A. Nenkova, and M. Lease, *Aggregating and predicting sequence labels from crowd annotations,* in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2017 (NIH Public Access, 2017) p. 299.

[153] X. Zhang, G. Xu, Y. Sun, M. Zhang, and P. Xie, *Crowdsourcing learning as domain adaptation: A case study on named entity recognition,* arXiv preprint arXiv:2105.14980 (2021).

[154] J. Li, *Crowdsourced text sequence aggregation based on hybrid reliability and representation,* in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020) pp. 1761–1764.

[155] J. Li and F. Fukumoto, *A dataset of crowdsourced word sequences: Collections and answer aggregation for ground truth creation,* in *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP* (2019) pp. 24–28.

[156] J. Welbl, N. F. Liu, and M. Gardner, *Crowdsourcing multiple choice science questions,* arXiv preprint arXiv:1707.06209 (2017).

[157] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, *Transformers: State-of-the-art natural language processing,* in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics, Online, 2020) pp. 38–45.

[158] R. Snow, B. O'connor, D. Jurafsky, and A. Y. Ng, *Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks,* in *Proceedings of the 2008 conference on empirical methods in natural language processing* (2008) pp. 254–263.

[159] J. Yang, A. Smirnova, D. Yang, G. Demartini, Y. Lu, and P. Cudré-Mauroux, *Scalpel-cd: leveraging crowdsourcing and deep probabilistic modeling for debugging noisy training data,* in *The World Wide Web Conference* (2019) pp. 2158–2168.

[160] B. Sayin, E. Krivosheev, J. Yang, A. Passerini, and F. Casati, *A review and experimental analysis of active learning over crowdsourced data,* Artificial Intelligence Review , 1 (2021).

[161] C. Sun, N. Rampalli, F. Yang, and A. Doan, *Chimera: Large-scale classification using machine learning, rules, and crowdsourcing,* Proceedings of the VLDB Endowment **7**, 1529 (2014).

[162] E. Simpson, J. Pfeiffer, and I. Gurevych, *Low resource sequence tagging with weak labels,* in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 (2020) pp. 8862–8869.

[163] J. Lin, T. Yu, and Z. J. Wang, *Rethinking crowdsourcing annotation: Partial annotation with salient labels for multi-label image classification,* arXiv preprint arXiv:2109.02688 (2021).

[164] L. Aroyo and C. Welty, *Truth is a lie: Crowd truth and the seven myths of human annotation,* AI Magazine **36**, 15 (2015).

[165] C. Northcutt, L. Jiang, and I. Chuang, *Confident learning: Estimating uncertainty in dataset labels,* Journal of Artificial Intelligence Research **70**, 1373 (2021).

[166] S. Min, D. Chen, H. Hajishirzi, and L. Zettlemoyer, *A discrete hard em approach for weakly supervised question answering,* arXiv preprint arXiv:1909.04849 (2019).

[167] N. Chen, L. Shou, M. Gong, and J. Pei, *From good to best: Two-stage training for cross-lingual machine reading comprehension,* in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36 (2022) pp. 10501–10508.

[168] A. Braylan and M. Lease, *Modeling and aggregation of complex annotations via annotation distances,* in *Proceedings of The Web Conference 2020* (2020) pp. 1807–1818.

[169] A. Braylan and M. Lease, *Aggregating complex annotations via merging and matching,* in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021) pp. 86–94.

[170] H. Kobayashi, *Frustratingly easy model ensemble for abstractive summarization,* in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018) pp. 4165–4176.

[171] J. Chen, E. Choi, and G. Durrett, *Can nli models verify qa systems' predictions?* arXiv preprint arXiv:2104.08731 (2021).

[172] S. Harabagiu and A. Hickl, *Methods for using textual entailment in open-domain question answering,* in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (2006) pp. 905–912.

[173] M. Richardson, C. J. Burges, and E. Renshaw, *Mctest: A challenge dataset for the open-domain machine comprehension of text,* in *Proceedings of the 2013 conference on empirical methods in natural language processing* (2013) pp. 193–203.

[174] A. Fisch, A. Talmor, R. Jia, M. Seo, E. Choi, and D. Chen, *MRQA 2019 shared task: Evaluating generalization in reading comprehension,* in *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP* (2019).

[175] T. Tian and J. Zhu, *Max-margin majority voting for learning from crowds,* Advances in neural information processing systems **28** (2015).

[176] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, *ERASER: A benchmark to evaluate rationalized NLP models,* in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online, 2020) pp. 4443–4458.

[177] Z. Zhang, K. Rudra, and A. Anand, *Explain and predict, and then predict again,* in *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021* (ACM, 2021) pp. 418–426.

[178] M. A. Sultan, S. Chandel, R. F. Astudillo, and V. Castelli, *On the importance of diversity in question generation for qa,* in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020) pp. 5651–5656.

[179] J. Hu, M. Xia, G. Neubig, and J. Carbonell, *Domain adaptation of neural machine translation by lexicon induction,* arXiv preprint arXiv:1906.00376 (2019).

[180] R. Aharoni and Y. Goldberg, *Unsupervised domain clusters in pretrained language models,* arXiv preprint arXiv:2004.02105 (2020).

[181] S. Rennie, E. Marcheret, N. Mallinar, D. Nahamoo, and V. Goel, *Unsupervised adaptation of question answering systems via generative self-training,* in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020) pp. 1148–1157.

[182] Y. Cao, M. Fang, B. Yu, and J. T. Zhou, *Unsupervised domain adaptation on reading comprehension,* in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 (2020) pp. 7480–7487.

[183] M. Van Der Wees, A. Bisazza, and C. Monz, *Dynamic data selection for neural machine translation,* arXiv preprint arXiv:1708.00712 (2017).

[184] S. A. Rauf, K. Kiani, A. Zafar, R. Nawaz, *et al.*, *Exploring transfer learning and domain data selection for the biomedical translation,* in *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)* (2019) pp. 156–163.

[185] J. Niehues and N.-Q. Pham, *Modeling confidence in sequence-to-sequence models,* arXiv preprint arXiv:1910.01859 (2019).

[186] D. Britz, Q. Le, and R. Pryzant, *Effective domain mixing for neural machine translation,* in *Proceedings of the Second Conference on Machine Translation* (2017) pp. 118–126.

[187] R. Sennrich, B. Haddow, and A. Birch, *Improving neural machine translation models with monolingual data,* arXiv preprint arXiv:1511.06709 (2015).

[188] A. Currey, A. V. Miceli-Barone, and K. Heafield, *Copied monolingual data improves low-resource neural machine translation,* in *Proceedings of the Second Conference on Machine Translation* (2017) pp. 148–156.

[189] G. Druck, G. Mann, and A. McCallum, *Learning from labeled features using generalized expectation criteria,* in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (2008) pp. 595–602.

[190] X. Ma, P. Xu, Z. Wang, R. Nallapati, and B. Xiang, *Domain adaptation with bert-based domain classification and data selection,* in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)* (2019) pp. 76–83.

[191] A. Gretton, K. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola, *A kernel method for the two-sample problem,* arXiv preprint arXiv:0805.2368 (2008).

[192] A. Ramponi and B. Plank, *Neural unsupervised domain adaptation in nlp—a survey,* arXiv preprint arXiv:2006.00632 (2020).

[193] A. Axelrod, X. He, and J. Gao, *Domain adaptation via pseudo in-domain data selection,* in *Proceedings of the 2011 conference on empirical methods in natural language processing* (2011) pp. 355–362.

[194] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, *Don't stop pretraining: adapt language models to domains and tasks,* arXiv preprint arXiv:2004.10964 (2020).

[195] H. Guo, R. Pasunuru, and M. Bansal, *Multi-source domain adaptation for text classification via distancenet-bandits,* in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34 (2020) pp. 7830–7838.

[196] S. Ruder and B. Plank, *Learning to select data for transfer learning with bayesian optimization,* arXiv preprint arXiv:1707.05246 (2017).

[197] R. C. Moore and W. Lewis, *Intelligent selection of language model training data,* (2010).

[198] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach,* arXiv preprint arXiv:1907.11692 (2019).

[199] D. McClosky, E. Charniak, and M. Johnson, *Effective self-training for parsing,* in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (2006) pp. 152–159.

[200] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, *Multi-source domain adaptation and its application to early detection of fatigue,* ACM Transactions on Knowledge Discovery from Data (TKDD) **6**, 1 (2012).

[201] H. S. Bhatt, D. Semwal, and S. Roy, *An iterative similarity based adaptation technique for cross-domain text classification,* in *Proceedings of the Nineteenth Conference on Computational Natural Language Learning* (2015) pp. 52–61.

[202] M. Sachan and E. Xing, *Self-training for jointly learning to ask and answer questions,* in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (2018) pp. 629–640.

[203] D.-H. Lee *et al.*, *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,* in *Workshop on challenges in representation learning, ICML*, Vol. 3 (2013) p. 896.

[204] Y.-H. Liao and J.-L. Koh, *Question generation through transfer learning,* in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (Springer, 2020) pp. 3–17.

[205] D. Yu, K. Sun, D. Yu, and C. Cardie, *Self-teaching machines to read and comprehend with large-scale multi-subject question-answering data,* arXiv preprint arXiv:2102.01226 (2021).

[206] P. Banerjee, T. Gokhale, and C. Baral, *Self-supervised test-time learning for reading comprehension,* arXiv preprint arXiv:2103.11263 (2021).

[207] D. Kulshreshtha, R. Belfer, I. V. Serban, and S. Reddy, *Back-training excels self-training at unsupervised domain adaptation of question generation and passage retrieval,* arXiv preprint arXiv:2104.08801 (2021).

[208] X. Jia, W. Zhou, X. Sun, and Y. Wu, *Eqg-race: Examination-type question generation,* arXiv preprint arXiv:2012.06106 (2020).

[209] M. Liu, Y. Song, H. Zou, and T. Zhang, *Reinforced training data selection for domain adaptation,* in *Proceedings of the 57th annual meeting of the association for computational linguistics* (2019) pp. 1957–1968.

[210] L. Lee, *On the effectiveness of the skew divergence for statistical language analysis,* in *International Workshop on Artificial Intelligence and Statistics* (PMLR, 2001) pp. 176–183.

[211] R. Mihalcea, *Co-training and self-training for word sense disambiguation,* in *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004* (2004) pp. 33–40.

[212] S. Abney, *Semisupervised learning for computational linguistics* (CRC Press, 2007).

[213] X. Cui and D. Bollegala, *Self-adaptation for unsupervised domain adaptation,* Proceedings-Natural Language Processing in a Deep Learning World (2019).

[214] R. Řehůřek and P. Sojka, *Software Framework for Topic Modelling with Large Corpora,* in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (ELRA, Valletta, Malta, 2010) pp. 45–50, `http://is.muni.cz/publication/884893/en`.

[215] D. Sculley, *Web-scale k-means clustering,* in *Proceedings of the 19th international conference on World wide web* (2010) pp. 1177–1178.

[216] J.-F. Rouet and E. Vidal-Abarca, *Mining for meaning: Cognitive effects of inserted questions in learning from scientific text,* The psychology of science text comprehension , 417 (2002).

[217] M. M. Dornisch, *Adjunct questions: Effects on learning,* Encyclopedia of the Sciences of Learning , 128 (2012).

[218] S. K. Carpenter, H. Pashler, J. T. Wixted, and E. Vul, *The effects of tests on learning and forgetting,* Memory & Cognition **36**, 438 (2008).

[219] C. I. Johnson and R. E. Mayer, *A testing effect with multimedia learning.* Journal of Educational Psychology **101**, 621 (2009).

[220] E. Z. Rothkopf, *Some theoretical and experimental approaches to problems in written instruction,* Learning and the educational process. Chicago: Rand McNally **965** (1965).

[221] E. Z. Rothkopf, *Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events,* American Educational Research Journal **3**, 241 (1966).

[222] G. Marchionini, *Exploratory search: from finding to understanding,* Communications of the ACM **49**, 41 (2006).

[223] A. Câmara, N. Roy, D. Maxwell, and C. Hauff, *Searching to learn with instructional scaffolding,* in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (2021) pp. 209–218.

[224] J. Von Hoyer, A. Hoppe, Y. Kammerer, C. Otto, G. Pardi, M. Rokicki, R. Yu, S. Dietze, R. Ewerth, and P. Holtz, *The search as learning spaceship: Toward a comprehensive model of psychological and technological facets of search as learning.* Frontiers in Psychology **13**, 827748 (2022).

[225] S. R. Putra, K. Grashoff, F. Moraes, and C. Hauff, *On the development of a collaborative search system.* in *DESIRES* (2018) pp. 76–82.

[226] M. Wesche and T. S. Paribakht, *Assessing second language vocabulary knowledge: Depth versus breadth,* Canadian Modern Language Review **53**, 13 (1996).

[227] N. Roy, M. V. Torre, U. Gadiraju, D. Maxwell, and C. Hauff, *Note the highlight: incorporating active reading tools in a search as learning environment,* in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (2021) pp. 229–238.

[228] H. Liu, C. Liu, and N. J. Belkin, *Investigation of users' knowledge change process in learning-related search tasks,* Proceedings of the Association for Information Science and Technology **56**, 166 (2019).

[229] N. Roy, F. Moraes, and C. Hauff, *Exploring users' learning gains within search sessions,* in *Proceedings of the 2020 conference on human information interaction and retrieval* (2020) pp. 432–436.

[230] M. J. Bates, *The design of browsing and berrypicking techniques for the online search interface,* Online Review **13**, 407 (1989).

[231] N. J. Belkin, *The cognitive viewpoint in information science,* Journal of Information Science **16**, 11 (1990).

[232] K. Byström and K. Järvelin, *Task complexity affects information seeking and use,* Information Processing & Management **31**, 191 (1995).

[233] F. Moraes, S. R. Putra, and C. Hauff, *Contrasting search as a learning activity with instructor-designed learning,* in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (2018) pp. 167–176.

[234] R. Syed and K. Collins-Thompson, *Retrieval Algorithms Optimized for Human Learning,* in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval,* SIGIR '17 (Association for Computing Machinery, New York, NY, USA, 2017) pp. 555–564.

[235] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag, *Personalizing web search results by reading level,* in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management,* CIKM '11 (Association for Computing Machinery) pp. 403–412.

[236] K. Athukorala, A. Medlar, A. Oulasvirta, G. Jacucci, and D. Glowacka, *Beyond Relevance: Adapting Exploration/Exploitation in Information Retrieval,* in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16 (Association for Computing Machinery) pp. 359–369.

[237] S. Salimzadeh, D. Maxwell, and C. Hauff, *The Impact of Entity Cards on Learning-Oriented Search Tasks,* in *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21 (Association for Computing Machinery) pp. 63–72.

[238] S. Qiu, A. Bozzon, and U. Gadiraju, *Conversational Interfaces for Search As Learning,* in *Proceedings of the First International Workshop on Investigating Learning During Web Search*, IWILDS 2020, p. 4.

[239] C. Liu, J. Gwizdka, J. Liu, T. Xu, and N. J. Belkin, *Analysis and evaluation of query reformulations in different task types,* in *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T'10 (American Society for Information Science) pp. 1–10.

[240] R. W. White, S. T. Dumais, and J. Teevan, *Characterizing the influence of domain expertise on web search behavior,* in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09 (Association for Computing Machinery) pp. 132–141.

[241] K. Umemoto, T. Yamamoto, and K. Tanaka, *ScentBar: A Query Suggestion Interface Visualizing the Amount of Missed Relevant Information for Intrinsically Diverse Search,* in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16 (Association for Computing Machinery) pp. 405–414.

[242] H. Liu, C. Liu, and N. J. Belkin, *Investigation of users' knowledge change process in learning-related search tasks,* Proceedings of the Association for Information Science and Technology **56**, 166 (2019).

[243] A. M. Shapiro, *Promoting Active Learning: The Role of System Structure in Learning From Hypertext,* Human–Computer Interaction **13**, 1 (1998).

[244] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham, *Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology,* Psychological Science in the Public Interest: A Journal of the American Psychological Society **14**, 4 (2013).

[245] W. Fass and G. M. Schumacher, *Effects of motivation, subject activity, and readability on the retention of prose materials,* Journal of Educational Psychology **70**, 803 (1978).

[246] W. Kintsch, *The role of knowledge in discourse comprehension: a construction-integration model.* Psychological review **95**, 163 (1988).

[247] R. C. Anderson and W. B. Biddle, *On asking people questions about what they are reading,* in *Psychology of learning and motivation*, Vol. 9 (Elsevier, 1975) pp. 89–132.

[248] S. Bulathwela, H. Muse, and E. Yilmaz, *Scalable educational question generation with pre-trained language models,* arXiv preprint arXiv:2305.07871 (2023).

[249] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, S. Piao, J. Gao, M. Zhou, and H.-W. Hon, *Unilmv2: Pseudo-masked language models for unified language model pre-training,* in *Preprint* (2020).

[250] P. Lewis, Y. Wu, L. Liu, P. Minervini, H. Küttler, A. Piktus, P. Stenetorp, and S. Riedel, *Paq: 65 million probably-asked questions and what you can do with them,* Transactions of the Association for Computational Linguistics **9**, 1098 (2021).

[251] W. Qi, Y. Gong, Y. Yan, C. Xu, B. Yao, B. Zhou, B. Cheng, D. Jiang, J. Chen, R. Zhang, *et al.*, *Prophetnet-x: Large-scale pre-training models for english, chinese, multi-lingual, dialog, and code generation,* arXiv preprint arXiv:2104.08006 (2021).

[252] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, *A survey of natural language generation,* ACM Computing Surveys **55**, 1 (2022).

[253] J. Brown, G. Frishkoff, and M. Eskenazi, *Automatic question generation for vocabulary assessment,* in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (2005) pp. 819–826.

[254] C.-H. Lee, T.-Y. Chen, L.-P. Chen, P.-C. Yang, and R. T.-H. Tsai, *Automatic question generation from children's stories for companion chatbot,* in *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (IEEE, 2018) pp. 491–494.

[255] T. Steuer, A. Filighera, and C. Rensing, *Remember the facts? investigating answer-aware neural question generation for text comprehension,* in *International Conference on Artificial Intelligence in Education* (Springer, 2020) pp. 512–523.

[256] M. Srivastava and N. Goodman, *Question generation for adaptive education,* arXiv preprint arXiv:2106.04262 (2021).

[257] A. Ahadi, A. Singh, M. Bower, and M. Garrett, *Text mining in education—a bibliometrics-based systematic review,* Education Sciences **12**, 210 (2022).

[258] D. Maxwell and C. Hauff, *Logui: contemporary logging infrastructure for web-based experiments,* in *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43* (Springer, 2021) pp. 525–530.

[259] U. Gadiraju, R. Yu, S. Dietze, and P. Holtz, *Analyzing knowledge gain of users in informational search sessions on the web,* in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (2018) pp. 2–11.

[260] L. Dietz, *Trec car y3: Complex answer retrieval overview,* (2019).

[261] C. R. Fox and J. Levav, *Familiarity bias and belief reversal in relative likelihood judgment,* Organizational Behavior and Human Decision Processes **82**, 268 (2000).

[262] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, *G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences,* Behavior research methods **39**, 175 (2007).

[263] J. L. Shefelbine, *Student factors related to variability in learning word meanings from context,* Journal of Reading Behavior **22**, 71 (1990).

[264] H. G. Colt, M. Davoudi, S. Murgu, and N. Zamanian Rohani, *Measuring learning gain during a one-day introductory bronchoscopy course,* Surgical endoscopy **25**, 207 (2011).

[265] OpenAI, *Gpt-4 technical report,* ArXiv **abs/2303.08774** (2023).

[266] M. J. Wilson and M. L. Wilson, *A comparison of techniques for measuring sensemaking and learning within participant-generated summaries,* Journal of the American Society for Information Science and Technology **64**, 291 (2013).

[267] K. Collins-Thompson, S. Y. Rieh, C. C. Haynes, and R. Syed, *Assessing learning outcomes in web search: A comparison of tasks and query strategies,* in *Proceedings of the 2016 ACM on conference on human information interaction and retrieval* (2016) pp. 163–172.

[268] M. J. Cole, J. Gwizdka, C. Liu, N. J. Belkin, and X. Zhang, *Inferring user knowledge level from eye movement patterns,* Information Processing & Management **49**, 1075 (2013).

[269] OpenAI, *Gpt-4 technical report,* (2023), arXiv:2303.08774 [cs.CL] .

[270] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, *Llama: Open and efficient foundation language models,* arXiv preprint arXiv:2302.13971 (2023).

[271] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, *Training language models to follow instructions with human feedback,* Advances in Neural Information Processing Systems **35**, 27730 (2022).

[272] S. Zheng, J. Huang, and K. C.-C. Chang, *Why does chatgpt fall short in answering questions faithfully?* arXiv preprint arXiv:2304.10513 (2023).

[273] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, *Survey of hallucination in natural language generation,* ACM Computing Surveys **55**, 1 (2023).

[274] F. J. Kelly, *The kansas silent reading tests.* Journal of Educational Psychology **7**, 63 (1916).

[275] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi, *Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension,* in *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition* (2017) pp. 4999–5007.

[276] S. Moore, R. Tong, A. Singh, Z. Liu, X. Hu, Y. Lu, J. Liang, C. Cao, H. Khosravi, P. Denny, *et al.*, *Empowering education with llms-the next-gen interface and content generation,* in *International Conference on Artificial Intelligence in Education* (Springer, 2023) pp. 32–37.

[277] R. J. Osborne and M. C. Wittrock, *Learning science: a generative process.* Science education **67**, 489 (1983).

[278] Z. Cai, B. Chang, and W. Han, *Human-in-the-loop through chain-of-thought,* arXiv preprint arXiv:2306.07932 (2023).

[279] P. Jiang, J. Rayan, S. P. Dow, and H. Xia, *Graphologue: Exploring large language model responses with interactive diagrams,* arXiv preprint arXiv:2305.11473 (2023).

[280] Z. A. Pardos, M. Tang, I. Anastasopoulos, S. K. Sheel, and E. Zhang, *Oatutor: An open-source adaptive tutoring system and curated content library for learning sciences research,* in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23 (Association for Computing Machinery, New York, NY, USA, 2023).

[281] R. T. McCoy, E. Pavlick, and T. Linzen, *Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference,* in *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019* (Association for Computational Linguistics (ACL), 2020) pp. 3428–3448.

[282] Y. Wang, L. Lyu, and A. Anand, *Bert rankers are brittle: a study using adversarial document perturbations,* in *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval* (2022) pp. 115–120.

[283] L. Yuan, Y. Chen, G. Cui, H. Gao, F. Zou, X. Cheng, H. Ji, Z. Liu, and M. Sun, *Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations,* ArXiv **abs/2306.04618** (2023).

[284] S. Mu, Y. Li, W. X. Zhao, J. Wang, B. Ding, and J.-R. Wen, *Alleviating spurious correlations in knowledge-aware recommendations through counterfactual generator,* in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022) pp. 1401–1411.

[285] W. Xu, X. Qian, M. Wang, L. Li, and W. Y. Wang, *Sescore2: Learning text generation evaluation via synthesizing realistic mistakes,* in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023) pp. 5166–5183.

[286] X. Deng, W. Wang, F. Feng, H. Zhang, X. He, and Y. Liao, *Counterfactual active learning for out-of-distribution generalization,* in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023) pp. 11362–11377.

[287] T. Dixit, B. Paranjape, H. Hajishirzi, and L. Zettlemoyer, *CORE: A retrieve-then-edit framework for counterfactual data generation, Findings of the Association for Computational Linguistics: EMNLP 2022,* , 2964 (2022).

[288] A. Feder, Y. Wald, C. Shi, S. Saria, and D. Blei, *Causal-structure driven augmentations for text ood generalization,* arXiv preprint arXiv:2310.12803 (2023).

[289] C. Leiter, P. Lertvittayakumjorn, M. Fomicheva, W. Zhao, Y. Gao, and S. Eger, *Towards explainable evaluation metrics for machine translation,* ArXiv **abs/2306.13041** (2023).

[290] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, and X. Li, *Auggpt: Leveraging chatgpt for text data augmentation,* (2023), arXiv:2302.13007 [cs.CL] .

[291] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, *GPT3Mix: Leveraging large-scale language models for text augmentation,* in *Findings of the Association for Computational Linguistics: EMNLP 2021* (Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021) pp. 2225–2239.

[292] V. Samuel, H. Aynaou, A. G. Chowdhury, K. V. Ramanan, and A. Chadha, *Can llms augment low-resource reading comprehension datasets? opportunities and challenges,* arXiv preprint arXiv:2309.12426 (2023).

[293] H. Zeng, B. Wei, J. Liu, and W. Fu, *Synthesize, prompt and transfer: Zero-shot conversational question generation with pre-trained language model,* in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Toronto, Canada, 2023) pp. 8989–9010.

[294] X. L. Do, B. Zou, S. Joty, T. Tai, L. Pan, N. Chen, and A. T. Aw, *Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation,* in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Toronto, Canada, 2023) pp. 10785–10803.

[295] A. Asai, S. Min, Z. Zhong, and D. Chen, *Retrieval-based language models and applications,* in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)* (Association for Computational Linguistics, Toronto, Canada, 2023) pp. 41–46.

[296] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, *In-context retrieval-augmented language models,* arXiv preprint arXiv:2302.00083 (2023).

[297] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, *Retrieval-augmented generation for knowledge-intensive nlp tasks,* Advances in Neural Information Processing Systems **33**, 9459 (2020).

[298] B. Xu, C. Zhao, W. Jiang, P. Zhu, S. Dai, C. Pang, Z. Sun, S. Wang, and Y. Sun, *Retrieval-augmented domain adaptation of language models,* in *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)* (2023) pp. 54–64.

[299] Y. Wu, A. Henriksson, M. Duneld, and J. Nouri, *Towards improving the reliability and transparency of chatgpt for educational question answering,* in *European Conference on Technology Enhanced Learning* (Springer, 2023) pp. 475–488.

[300] N. Stylianou, D. Chatzakou, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, *Domain-aligned data augmentation for low-resource and imbalanced text classification,* in *European Conference on Information Retrieval* (Springer, 2023) pp. 172–187.

[301] J. Goodell and J. Kolodner, *Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond* (Taylor & Francis, 2022).

[302] H. A. Nguyen, S. Bhat, S. Moore, N. Bier, and J. Stamper, *Towards generalized methods for automatic question generation in educational domains,* in *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-TEL 2022, Toulouse, France, September 12–16, 2022, Proceedings* (Springer, 2022) pp. 272–284.

[303] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, *Deep knowledge tracing,* Advances in neural information processing systems **28** (2015).

[304] A. Ghosh, J. Raspat, and A. Lan, *Option tracing: Beyond correctness analysis in knowledge tracing,* in *International Conference on Artificial Intelligence in Education* (Springer, 2021) pp. 137–149.

[305] J. Yu, Y. Wang, Q. Zhong, G. Luo, Y. Mao, K. Sun, W. Feng, W. Xu, S. Cao, K. Zeng, *et al.*, *MOOCCubeX: A large knowledge-centered repository for adaptive learning in MOOCs,* in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (2021) pp. 4643–4652.

[306] P. Oleksandra and D. Shane, *Untangling mooc learner networks,* in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (2016) pp. 208–212.

[307] J. Cruz-Benito, O. Borras-Gene, F. J. Garcia-Penalvo, A. F. Blanco, and R. Theron, *Learning communities in social networks and their relationship with the moocs,* IEEE Revista Iberoamericana de Tecnologias del Aprendizaje **12**, 24 (2017).

[308] X. Yi, Z. Zhan, and S. Ma, *An evaluation model of mooc learning engagement based on learning behaviors and social network analysis,* in *2022 IEEE 2nd International Conference on Educational Technology (ICET)* (IEEE, 2022) pp. 165–169.

[309] R. Ali, O. Y. Tang, I. D. Connolly, P. L. Z. Sullivan, J. H. Shin, J. S. Fridley, W. F. Asaad, D. Cielo, A. A. Oyelese, C. E. Doberstein, Z. L. Gokaslan, and A. E. Telfeian, *Performance of chatgpt and gpt-4 on neurosurgery written board examinations,* medRxiv (2023), 10.1101/2023.03.25.23287743, https://www.medrxiv.org/content/early/2023/03/29/2023.03.25.23287743.full.pdf .

[310] R. Berry, *Assessment for learning*, Vol. 1 (Hong Kong University Press, 2008).

[311] H. L. Roediger III and J. D. Karpicke, *The power of testing memory: Basic research and implications for educational practice,* Perspectives on psychological science **1**, 181 (2006).

[312] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, *et al.*, *Chatgpt for good? on opportunities and challenges of large language models for education,* Learning and individual differences **103**, 102274 (2023).

[313] Z. A. Pardos and S. Bhandari, *Learning gain differences between chatgpt and human tutor generated algebra hints,* arXiv preprint arXiv:2302.06871 (2023).

[314] A. Bewersdorff, K. Seßler, A. Baur, E. Kasneci, and C. Nerdel, *Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters,* arXiv preprint arXiv:2308.06088 (2023).

[315] M. Pankiewicz and R. S. Baker, *Large language models (gpt) for automating feedback on programming assignments,* arXiv preprint arXiv:2307.00150 (2023).

[316] H. A. Nguyen, H. Stec, X. Hou, S. Di, and B. M. McLaren, *Evaluating chatgpt's decimal skills and feedback generation in a digital learning game,* in *European Conference on Technology Enhanced Learning* (Springer, 2023) pp. 278–293.

List of Figures

[2]https://platform.openai.com/playground

List of Tables

Summary

Questions are critical for information-seeking and learning. Automatic Question Generation (AQG) involves the subjects of Information Retrieval (IR) and Natural Language Processing (NLP), and focuses on automatically creating questions for various applications, subjects which have been studied for decades. In this thesis, we study how to create a robust automatic question generation system from several aspects, including data creation, evaluation, and effects of question generation.

First, we contribute to the quality evaluation of the generated questions. Specifically, we introduce three new evaluation metrics and compare the effects of applying the automatic evaluation metrics as rewards for reinforcement learning-based question generation system training. Question quality evaluation is an essential part of AQG systems. It is further used in this thesis in dataset creation, question selection for self-training, and filtering automatically generated questions shown for learners.

Data are essential for building AQG systems. In Chapters 3 and 4, we focus on data quality control in two main methods of dataset creation: collecting user-generated resources from online platforms and from crowdsourcing. Specifically, we start by investigating the information overload issue in MOOC forum discussions caused by unuseful, unlabeled, and unstructured data. We propose a framework for clip recommendation that includes useful question classification and a neural ranker. We further investigate training the neural ranker with both labeled and weakly labeled data. We then study how to infer the true answer span from multiple crowdsourced annotations automatically. We propose an approach to effectively utilize the quality of each answer annotation and its relation to other answer annotations for answer aggregation. Despite the various available methods of collecting labeled data, there are many application domains where the labeled data is hard or expensive to harvest. In Chapter 5, we move to automatically adapting the AQG model trained on label-data-abundant domains to strange domains with few labeled data.

With the impressive advantages of automatic question generation methods, it is critical to understand how the generated questions on humans. Finally, in Chapter 6, we turn to study the effects of automatically generated questions on the learners' behaviours and learning outcomes when they serve as the adjunct questions in the informal search as learning scenario. We conduct an extensive user study to shed light on this topic.

Samenvatting

Vragen zijn cruciaal voor zoeken van informatie en voor leren. Automatische Vraaggeneratie (AVG) omvat de onderwerpen van Information Retrieval (IR) en Natural Language Processing (NLP), en richt zich op het automatisch creëren van vragen voor diverse toepassingen, onderwerpen die al decennia lang worden bestudeerd. In dit proefschrift bestuderen we hoe we een robuust automatisch vraaggeneratiesysteem kunnen creëren vanuit verschillende aspecten, zoals datacreatie, evaluatie en effecten van vraaggeneratie.

Ten eerste dragen we bij aan de kwaliteitsevaluatie van de gegenereerde vragen. Specifiek introduceren we drie nieuwe evaluatiemetrieken en vergelijken de effecten van het toepassen van de automatische evaluatiemetrieken als beloningen voor reinforcement learning-gebaseerde training van het vraaggeneratiesysteem. Evaluatie van de vraagkwaliteit is een essentieel onderdeel van AVG-systemen. Het wordt verder in dit proefschrift gebruikt in datasetcreatie, vraagselectie voor zelftraining en de filtering van automatisch gegenereerde vragen die aan leerlingen worden getoond.

Data zijn essentieel voor het bouwen van AVG-systemen. In Hoofdstukken 3 en 4 concentreren we ons op de controle van datakwaliteit in twee belangrijke methoden van datasetcreatie: het verzamelen van door gebruikers gegenereerde bronnen vanuit online platforms en vanuit crowdsourcing. Specifiek beginnen we met het onderzoeken van het probleem van informatieoverload in MOOC-forumdiscussies, veroorzaakt door onbruikbare, ongelabelde en ongestructureerde data. We stellen een raamwerk voor aanbeveling van clips voor dat de classificatie van nuttige vragen en een neurale rangschikker omvat. We onderzoeken verder het trainen van de neurale rangschikker met zowel gelabelde als zwak gelabelde data. Vervolgens bestuderen we hoe we automatisch het ware antwoordbereik kunnen afleiden uit meerdere crowdsourced annotaties. We stellen een aanpak voor om effectief gebruik te maken van de kwaliteit van elke antwoordannotatie en de relatie tot andere antwoordannotaties voor antwoordaggregatie. Ondanks de verschillende beschikbare methoden voor het verzamelen van gelabelde data, zijn er veel toepassingsdomeinen waar gelabelde data moeilijk of duur is om te vergaren. In Hoofdstuk 5 gaan we over tot het automatisch aanpassen van het AVG-model dat getraind is op domeinen met veel gelabelde data, naar vreemde domeinen met weinig gelabelde data.

Met de indrukwekkende voordelen van automatische vraaggeneratiemethoden is het cruciaal om te begrijpen wat voor effect de gegenereerde vragen op mensen hebben. Ten slotte, in Hoofdstuk 6, keren we ons naar het bestuderen van de effecten van automatisch gegenereerde vragen op het gedrag van leerlingen en leerresultaten, wanneer ze dienen als toegevoegde vragen in het informele zoeken-als-leren-scenario. We voeren een uitgebreide gebruikersstudie uit om licht te werpen op dit onderwerp.

Curriculum Vitæ

Experience and Education

2016 – 2019	M.Eng in Computer Science and Technology, School of Computer Science and Technology, USTC, Hefei, China
2014 – 2015	Software Engineer, UCloud Technology, Shanghai, China
2013 – 2014	Software Engineer, Huawei Technology, Shenzhen, China
2013.3 – 6	Intern Software Engineer, Virtuos, Shanghai, China
2009 – 2013	B.Eng in Information Security, College of Computer Science and Technology, NUAA, Nanjing, China

Publications

CHIIR 2024

On the Effects of Automatically Generated Adjunct Questions for Search as Learning
Peide Zhu, Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff

MMM 2024

A New Benchmark and OCR-free Method for Document Image Topic Classification
Zhen Wang, **Peide Zhu**, Fuyang Yu, Manabu Okumura

MMM 2024

MRHF: Multi-stage Retrieval and Hierarchical Fusion for Textbook Question Answering
Peide Zhu, Zhen Wang, Jie Yang, Manabu Okumura

Arxiv 2023

ControversialQA: Exploring Controversy in Question Answering
Zhen Wang, **Peide Zhu**, Jie Yang

Findings@EMNLP 2022	Answer Quality Aware Aggregation for Extractive QA Crowdsourcing **Peide Zhu**, Zhen Wang, Claudia Hauff, Jie Yang, Avishek Anand
Findings@NAACL 2022	Answer Quality Aware Aggregation for Extractive QA Crowdsourcing **Peide Zhu**, Claudia Hauff
EDM 2022	MOOC-Rec: Instructional Video Clip Recommendation for MOOC Forum Questions **Peide Zhu**, Jie Yang, Claudia Hauff
ICTIR 2021	Evaluating BERT-based Rewards for Question Generation with Reinforcement Learning **Peide Zhu**, Claudia Hauff
PRICAI 2021	Improving Long Content Question Generation with Multi-level Passage Encoding **Peide Zhu**
BIGCOM 2019	Control with gestures: A hand gesture recognition system using off-the-shelf smartwatch **Peide Zhu**, Hao Zhou, Shumin Cao, Panlong Yang, Shuangshuang Xue
IPCCC 2019	IPand: accurate gesture input with ambient acoustic sensing on hand Shumin Cao, Xin He, **Peide Zhu**, Mingshi Chen, Xiangyang Li, Panlong Yang
MOBICOM 2019	SignSpeaker: A Real-Time, High-Precision SmartWatch-Based Sign Language Translator Jiahui Hou, Xiang-Yang Li, **Zhu, Peide**, Zefan Wang, Yu Wang, Jianwei Qian, Panlong Yang

SIKS Dissertation Series

SIKS Dissertatiereeks

==== 2016 ====

2016-01 Syed Saiden Abbas (RUN) Recognition of Shapes by Humans and Machines

2016-02 Michiel Christiaan Meulendijk (UU) Optimizing medication reviews through decision support: prescribing a better pill to swallow

2016-03 Maya Sappelli (RUN) Knowledge Work in Context: User Centered Knowledge Worker Support

2016-04 Laurens Rietveld (VU) Publishing and Consuming Linked Data

2016-05 Evgeny Sherkhonov (UvA) Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

2016-06 Michel Wilson (TUD) Robust scheduling in an uncertain environment

2016-07 Jeroen de Man (VU) Measuring and modeling negative emotions for virtual training

2016-08 Matje van de Camp (TiU) A Link to the Past: Constructing Historical Social Networks from Unstructured Data

2016-09 Archana Nottamkandath (VU) Trusting Crowdsourced Information on Cultural Artefacts

2016-10 George Karafotias (VUA) Parameter Control for Evolutionary Algorithms

2016-11 Anne Schuth (UvA) Search Engines that Learn from Their Users

2016-12 Max Knobbout (UU) Logics for Modelling and Verifying Normative Multi-Agent Systems

2016-13 Nana Baah Gyan (VU) The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

2016-14 Ravi Khadka (UU) Revisiting Legacy Software System Modernization

2016-15 Steffen Michels (RUN) Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

2016-16 Guangliang Li (UvA) Socially Intelligent Autonomous Agents that Learn from Human Reward

2016-17 Berend Weel (VU) Towards Embodied Evolution of Robot Organisms

2016-18 Albert Meroño Peñuela (VU) Refining Statistical Data on the Web

2016-19 Julia Efremova (TU/e) Mining Social Structures from Genealogical Data

2016-20 Daan Odijk (UvA) Context & Semantics in News & Web Search

2016-21 Alejandro Moreno Célleri (UT) From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

2016-22 Grace Lewis (VU) Software Architecture Strategies for Cyber-Foraging Systems

2016-23 Fei Cai (UvA) Query Auto Completion in Information Retrieval

2016-24 Brend Wanders (UT) Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

2016-25 Julia Kiseleva (TU/e) Using Contextual Information to Understand Searching and Browsing Behavior

2016-26 Dilhan Thilakarathne (VU) In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

2016-27 Wen Li (TUD) Understanding Geo-spatial Information on Social Media

2016-28 Mingxin Zhang (TUD) Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

2016-29 Nicolas Höning (TUD) Peak reduction in decentralised electricity systems -Markets and prices for flexible planning

2016-30 Ruud Mattheij (TiU) The Eyes Have It

2016-31 Mohammad Khelghati (UT) Deep web content monitoring

2016-32 Eelco Vriezekolk (UT) Assessing Telecommunication Service Availability Risks for Crisis Organisations

2016-33 Peter Bloem (UvA) Single Sample Statistics, exercises in learning from just one example

2016-34 Dennis Schunselaar (TU/e) Configurable Process Trees: Elicitation, Analysis, and Enactment

2016-35 Zhaochun Ren (UvA) Monitoring Social Media: Summarization, Classification and Recommendation

2016-36 Daphne Karreman (UT) Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

2016-37 Giovanni Sileno (UvA) Aligning Law and Action - a conceptual and computational inquiry

2016-38 Andrea Minuto (UT) Materials that Matter - Smart Materials meet Art & Interaction Design

2016-39 Merijn Bruijnes (UT) Believable Suspect Agents: Response and Interpersonal Style Selection for an Artificial Suspect

2016-40 Christian Detweiler (TUD) Accounting for Values in Design

2016-41 Thomas King (TUD) Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

2016-42 Spyros Martzoukos (UvA) Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

2016-43 Saskia Koldijk (RUN) Context-Aware Support for Stress Self-Management: From Theory to Practice

2016-44 Thibault Sellam (UvA) Automatic Assistants for Database Exploration

2016-45 Bram van de Laar (UT) Experiencing Brain-Computer Interface Control

2016-46 Jorge Gallego Perez (UT) Robots to Make you Happy

2016-47 Christina Weber (UL) Real-time foresight - Preparedness for dynamic innovation networks

2016-48 Tanja Buttler (TUD) Collecting Lessons Learned

2016-49 Gleb Polevoy (TUD) Participation and Interaction in Projects. A Game-Theoretic Analysis

2016-50 Yan Wang (TiU) The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

==== 2017 ====

2017-01 Jan-Jaap Oerlemans (UL) Investigating Cybercrime

2017-02 Sjoerd Timmer (UU) Designing and Understanding Forensic Bayesian Networks using Argumentation

2017-03 Daniël Harold Telgen (UU) Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

2017-04 Mrunal Gawade (CWI) MULTI-CORE PARALLELISM IN A COLUMN-STORE

2017-05 Mahdieh Shadi (UvA) Collaboration Behavior Enhancement in Co-development Networks

2017-06 Damir Vandic (EUR) Intelligent Information Systems for Web Product Search

2017-07 Roel Bertens (UU) Insight in Information: from Abstract to Anomaly

2017-08 Rob Konijn (VU) Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

2017-09 Dong Nguyen (UT) Text as Social and Cultural Data: A Computational Perspective on Variation in Text

2017-10 Robby van Delden (UT) (Steering) Interactive Play Behavior

2017-11 Florian Kunneman (RUN) Modelling patterns of time and emotion in Twitter #anticipointment

2017-12 Sander Leemans (TU/e) Robust Process Mining with Guarantees

2017-13 Gijs Huisman (UT) Social Touch Technology - Extending the reach of social touch through haptic technology

2017-14 Shoshannah Tekofsky (TiU) You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

2017-15 Peter Berck, Radboud University (RUN) Memory-Based Text Correction

2017-16 Aleksandr Chuklin (UvA) Understanding and Modeling Users of Modern Search Engines

2017-17 Daniel Dimov (UL) Crowdsourced Online Dispute Resolution

2017-18 Ridho Reinanda (UvA) Entity Associations for Search

2017-19 Jeroen Vuurens (TUD) Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

2017-20 Mohammadbashir Sedighi (TUD) Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

2017-21 Jeroen Linssen (UT) Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

2017-22 Sara Magliacane (VU) Logics for causal inference under uncertainty

2017-23 David Graus (UvA) Entities of Interest - Discovery in Digital Traces

2017-24 Chang Wang (TUD) Use of Affordances for Efficient Robot Learning

2017-25 Veruska Zamborlini (VU) Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

2017-26 Merel Jung (UT) Socially intelligent robots that understand and respond to human touch

2017-27 Michiel Joosse (UT) Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

2017-28 John Klein (VU) Architecture Practices for Complex Contexts

2017-29 Adel Alhuraibi (TiU) From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT

2017-30 Wilma Latuny (TiU) The Power of Facial Expressions

2017-31 Ben Ruijl (UL) Advances in computational methods for QFT calculations

2017-32 Thaer Samar (RUN) Access to and Retrievability of Content in Web Archives

2017-33 Brigit van Loggem (OU) Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

2017-34 Maren Scheffel (OU) The Evaluation Framework for Learning Analytics

2017-35 Martine de Vos (VU) Interpreting natural science spreadsheets

2017-36 Yuanhao Guo (UL) Shape Analysis for Phenotype Characterisation from High-throughput Imaging

2017-37 Alejandro Montes García (TU/e) WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

2017-38 Alex Kayal (TUD) Normative Social Applications

2017-39 Sara Ahmadi (RUN) Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

2017-40 Altaf Hussain Abro (VUA) Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems”

2017-41 Adnan Manzoor (VUA) Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

2017-42 Elena Sokolova (RUN) Causal discovery from mixed and missing data with applications on ADHD datasets

2017-43 Maaike de Boer (RUN) Semantic Mapping in Video Retrieval

2017-44 Garm Lucassen (UU) Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

2017-45 Bas Testerink (UU) Decentralized Runtime Norm Enforcement

2017-46 Jan Schneider (OU) Sensor-based Learning Support

2017-47 Yie Yang (TUD) Crowd Knowledge Creation Acceleration

2017-48 Angel Suarez (OU) Collaborative inquiry-based learning

==== 2018 ====

2018-01 Han van der Aa (VUA) Comparing and Aligning Process Representations

2018-02 Felix Mannhardt (TU/e) Multi-perspective Process Mining

2018-03 Steven Bosems (UT) Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

2018-04 Jordan Janeiro (TUD) Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

2018-05 Hugo Huurdeman (UvA) Supporting the Complex Dynamics of the Information Seeking Process

2018-06 Dan Ionita (UT) Model-Driven Information Security Risk Assessment of Socio-Technical Systems

2018-07 Jieting Luo (UU) A formal account of opportunism in multi-agent systems

2018-08 Rick Smetsers (RUN) Advances in Model Learning for Software Systems

2018-09 Xu Xie (TUD) Data Assimilation in Discrete Event Simulations

2018-10 Julienka Mollee (VUA) Moving forward: supporting physical activity behavior change through intelligent technology

2018-11 Mahdi Sargolzaei (UvA) Enabling Framework for Service-oriented Collaborative Networks

2018-12 Xixi Lu (TU/e) Using behavioral context in process mining

2018-13 Seyed Amin Tabatabaei (VUA) Using behavioral context in process mining: Exploring the added value of computational models for increasing the use of renewable energy in the residential sector

2018-14 Bart Joosten (TiU) Detecting Social Signals with Spatiotemporal Gabor Filters

2018-15 Naser Davarzani (UM) Biomarker discovery in heart failure

2018-16 Jaebok Kim (UT) Automatic recognition of engagement and emotion in a group of children

2018-17 Jianpeng Zhang (TU/e) On Graph Sample Clustering

2018-18 Henriette Nakad (UL) De Notaris en Private Rechtspraak

2018-19 Minh Duc Pham (VUA) Emergent relational schemas for RDF

2018-20 Manxia Liu (RUN) Time and Bayesian Networks

2018-21 Aad Slootmaker (OU) EMERGO: a generic platform for authoring and playing scenario-based serious games

2018-22 Eric Fernandes de Mello Araújo (VUA) Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

2018-23 Kim Schouten (EUR) Semantics-driven Aspect-Based Sentiment Analysis

2018-24 Jered Vroon (UT) Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

2018-25 Riste Gligorov (VUA) Serious Games in Audio-Visual Collections

2018-26 Roelof de Vries (UT) Theory-Based And Tailor-Made: Motivational Messages for Behavior Change Technology

2018-27 Maikel Leemans (TU/e) Hierarchical Process Mining for Scalable Software Analysis

2018-28 Christian Willemse (UT) Social Touch Technologies: How they feel and how they make you feel

2018-29 Yu Gu (TiU) Emotion Recognition from Mandarin Speech

2018-30 Wouter Beek (VUA) The ”K” in ”semantic web” stands for ”knowledge”: scaling semantics to the web

==== 2019 ====

2019-01 Rob van Eijk (UL) Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

2019-02 Emmanuelle Beauxis- Aussalet (CWI, UU) Statistics and Visualizations for Assessing Class Size Uncertainty

2019-03 Eduardo Gonzalez Lopez de Murillas (TU/e) Process Mining on Databases: Extracting Event Data from Real-life Data Sources

2019-04 Ridho Rahmadi (RUN) Finding stable causal structures from clinical data

2019-05 Sebastiaan van Zelst (TU/e) Process Mining with Streaming Data

2019-06 Chris Dijkshoorn (VUA) Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

2019-07 Soude Fazeli (TUD) Recommender Systems in Social Learning Platforms

2019-08 Frits de Nijs (TUD) Resource-constrained Multi-agent Markov Decision Processes

2019-09 Fahimeh Alizadeh Moghaddam (UvA) Self-adaptation for energy efficiency in software systems

2019-10 Qing Chuan Ye (EUR) Multi-objective Optimization Methods for Allocation and Prediction

2019-11 Yue Zhao (TUD) Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

2019-12 Jacqueline Heinerman (VUA) Better Together

2019-13 Guanliang Chen (TUD) MOOC Analytics: Learner Modeling and Content Generation

2019-14 Daniel Davis (TUD) Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

2019-15 Erwin Walraven (TUD) Planning under Uncertainty in Constrained and Partially Observable Environments

2019-16 Guangming Li (TU/e) Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

2019-17 Ali Hurriyetoglu (RUN) Extracting actionable information from microtexts

2019-18 Gerard Wagenaar (UU) Artefacts in Agile Team Communication

2019-19 Vincent Koeman (TUD) Tools for Developing Cognitive Agents

2019-20 Chide Groenouwe (UU) Fostering technically augmented human collective intelligence

2019-21 Cong Liu (TU/e) Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

2019-22 Martin van den Berg (VUA) Improving IT Decisions with Enterprise Architecture

2019-23 Qin Liu (TUD) Intelligent Control Systems: Learning, Interpreting, Verification

2019-24 Anca Dumitrache (VUA) Truth in Disagreement- Crowdsourcing Labeled Data for Natural Language Processing

2019-25 Emiel van Miltenburg (TiU) Pragmatic factors in (automatic) image description

2019-26 Prince Singh (UT) An Integration Platform for Synchromodal Transport

2019-27 Alessandra Antonaci (OU) The Gamification Design Process applied to (Massive) Open Online Courses

2019-28 Esther Kuindersma (UL) Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

2019-29 Daniel Formolo (VUA) Using virtual agents for simulation and training of social skills in safety-critical circumstances

2019-30 Vahid Yazdanpanah (UT) Multiagent Industrial Symbiosis Systems

2019-31 Milan Jelisavcic (VUA) Alive and Kicking: Baby Steps in Robotics

2019-32 Chiara Sironi (UM) Monte-Carlo Tree Search for Artificial General Intelligence in Games

2019-33 Anil Yaman (TU/e) Evolution of Biologically Inspired Learning in Artificial Neural Networks

2019-34 Negar Ahmadi (TU/e) EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

2019-35 Lisa Facey-Shaw (OU) Gamification with digital badges in learning programming

2019-36 Kevin Ackermans (OU) Designing Video-Enhanced Rubrics to Master Complex Skills

2019-37 Jian Fang (TUD) Database Acceleration on FPGAs

2019-38 Akos Kadar (OU) Learning visually grounded and multilingual representations

==== 2020 ====

2020-01 Armon Toubman (UL) Calculated Moves: Generating Air Combat Behaviour

2020-02 Marcos de Paula Bueno (UL) Unraveling Temporal Processes using Probabilistic Graphical Models

2020-03 Mostafa Deghani (UvA) Learning with Imperfect Supervision for Language Understanding

2020-04 Maarten van Gompel (RUN) Context as Linguistic Bridges

2020-05 Yulong Pei (TU/e) On local and global structure mining

2020-06 Preethu Rose Anish (UT) Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

2020-07 Wim van der Vegt (OU) Towards a software architecture for reusable game components

2020-08 Ali Mirsoleimani (UL) Structured Parallel Programming for Monte Carlo Tree Search

2020-09 Myriam Traub (UU) Measuring Tool Bias & Improving Data Quality for Digital Humanities Research

2020-10 Alifah Syamsiyah (TU/e) In-database Preprocessing for Process Mining

2020-11 Sepideh Mesbah (TUD) Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models

2020-12 Ward van Breda (VUA) Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

2020-13 Marco Virgolin (CWI) Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

2020-14 Mark Raasveldt (CWI/UL) Integrating Analytics with Relational Databases

2020-15 Konstantinos Georgiadis (OU) Smart CAT: Machine Learning for Configurable Assessments in Serious Games

2020-16 Ilona Wilmont (RUN) Cognitive Aspects of Conceptual Modelling

2020-17 Daniele Di Mitri (OU) The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

2020-18 Georgios Methenitis (TUD) Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

2020-19 Guido van Capelleveen (UT) Industrial Symbiosis Recommender Systems

2020-20 Albert Hankel (VUA) Embedding Green ICT Maturity in Organisations

2020-21 Karine da Silva Miras de Araujo (VUA) Where is the robot? Life as it could be

2020-22 Maryam Masoud Khamis (RUN) Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

2020-23 Rianne Conijn (UT) The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

2020-24 Lenin da Nobrega Medeiros (VUA/RUN) How are you feeling, human? Towards emotionally supportive chatbots

2020-25 Xin Du (TU/e) The Uncertainty in Exceptional Model Mining

2020-26 Krzysztof Leszek Sadowski (UU) GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

2020-27 Ekaterina Muravyeva (TUD) Personal data and informed consent in an educational context

2020-28 Bibeg Limbu (TUD) Multimodal interaction for deliberate practice: Training complex skills with augmented reality

2020-29 Ioan Gabriel Bucur (RUN) Being Bayesian about Causal Inference

2020-30 Bob Zadok Blok (UL) Creatief, Creatiever, Creatiefst

2020-31 Gongjin Lan (VUA) Learning better - From Baby to Better

2020-32 Jason Rhuggenaath (TU/e) Revenue management in online markets: pricing and online advertising

2020-33 Rick Gilsing (TU/e) Supporting service-dominant business model evaluation in the context of business model innovation

2020-34 Anna Bon (UM) Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

2020-35 Siamak Farshidi (UU) Multi-Criteria Decision-Making in Software Production

==== 2021 ====

2021-01 Francisco Xavier Dos Santos Fonseca (TUD) Location-based Games for Social Interaction in Public Space

2021-02 Rijk Mercuur (TUD) Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

2021-03 Seyyed Hadi Hashemi (UvA) Modeling Users Interacting with Smart Devices

2021-04 Ioana Jivet (OU) The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

2021-05 Davide Dell'Anna (UU) Data-Driven Supervision of Autonomous Systems

2021-06 Daniel Davison (UT) "Hey robot, what do you think?" How children learn with a social robot

2021-07 Armel Lefebvre (UU) Research data management for open science

2021-08 Nardie Fanchamps (OU) The Influence of Sense-Reason-Act Programming on Computational Thinking

2021-09 Cristina Zaga (UT) The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

2021-10 Quinten Meertens (UvA) Misclassification Bias in Statistical Learning

2021-11 Anne van Rossum (UL) Nonparametric Bayesian Methods in Robotic Vision

2021-12 Lei Pi (UL) External Knowledge Absorption in Chinese SMEs

2021-13 Bob R. Schadenberg (UT) Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

2021-14 Negin Samaeemofrad (UL) Business Incubators: The Impact of Their Support

2021-15 Onat Ege Adali (TU/e) Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

2021-16 Esam A. H. Ghaleb (UM) Bimodal Emotion recognition from audio-visual cues

2021-17 Dario Dotti (UM) Human Behavior Understanding from motion and bodily cues using deep neural networks

2021-18 Remi Wieten (UU) Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

2021-19 Roberto Verdecchia (VUA) Architectural Technical Debt: Identification and Management

2021-20 Masoud Mansoury (TU/e) Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

2021-21 Pedro Thiago Timbó Holanda (CWI) Progressive Indexes

2021-22 Sihang Qiu (TUD) Conversational Crowdsourcing

2021-23 Hugo Manuel Proença (UL) Robust rules for prediction and description

2021-24 Kaijie Zhu (TU/e) On Efficient Temporal Subgraph Query Processing

2021-25 Eoin Martino Grua (VUA) The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

2021-26 Benno Kruit (CWI & VUA) Reading the Grid: Extending Knowledge Bases from Human-readable Tables

2021-27 Jelte van Waterschoot (UT) Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

2021-28 Christoph Selig (UL) Understanding the Heterogeneity of Corporate Entrepreneurship Programs

==== 2022 ====

2022-1 Judith van Stegeren (UT) Flavor text generation for role-playing video games

2022-2 Paulo da Costa (TU/e) Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

2022-3 Ali el Hassouni (VUA) A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

2022-4 Ünal Aksu (UU) A Cross-Organizational Process Mining Framework

2022-5 Shiwei Liu (TU/e) Sparse Neural Network Training with In-Time Over-Parameterization

2022-6 Reza Refaei Afshar (TU/e) Machine Learning for Ad Publishers in Real Time Bidding

2022-7 Sambit Praharaj (OU) Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

2022-8 Maikel L. van Eck (TU/e) Process Mining for Smart Product Design

2022-9 Oana Andreea Inel (VUA) Understanding Events: A Diversity-driven Human-Machine Approach

2022-10 Felipe Moraes Gomes (TUD) Examining the Effectiveness of Collaborative Search Engines

2022-11 Mirjam de Haas (UT) Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

2022-12 Guanyi Chen (UU) Computational Generation of Chinese Noun Phrases

2022-13 Xander Wilcke (VUA) Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented

2022-14 Michiel Overeem (UU) Evolution of Low-Code Platforms

2022-15 Jelmer Jan Koorn (UU) Work in Process: Unearthing Meaning using Process Mining

2022-16 Pieter Gijsbers (TU/e) Systems for AutoML Research

2022-17 Laura van der Lubbe (VUA) Empowering vulnerable people with serious games and gamification

2022-18 Paris Mavromoustakos Blom (TiU) Player Affect Modelling and Video Game Personalisation

2022-19 Bilge Yigit Ozkan (UU) Cybersecurity Maturity Assessment and Standardisation

2022-20 Fakhra Jabeen (VUA) Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

2022-21 Seethu Mariyam Christopher (UM) Intelligent Toys for Physical and Cognitive Assessments

2022-22 Alexandra Sierra Rativa (TiU) Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

2022-23 Ilir Kola (TUD) Enabling Social Situation Awareness in Support Agents

2022-24 Samaneh Heidari (UU) Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

2022-25 Anna L.D. Latour (UL) Optimal decision-making under constraints and uncertainty

2022-26 Anne Dirkson (UL) Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

2022-27 Christos Athanasiadis (UM) Emotion-aware cross-modal domain adaptation in video sequences

2022-28 Onuralp Ulusoy (UU) Privacy in Collaborative Systems

2022-29 Jan Kolkmeier (UT) From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

2022-30 Dean De Leo (CWI) Analysis of Dynamic Graphs on Sparse Arrays

2022-31 Konstantinos Traganos (TU/e) Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

2022-32 Cezara Pastrav (UU) Social simulation for socio-ecological systems

2022-33 Brinn Hekkelman (TUD/CWI) Fair Mechanisms for Smart Grid Congestion Management

2022-34 Nimat Ullah (VUA) Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

2022-35 Mike E.U. Ligthart (VUA) Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

==== 2023 ====

2023-01 Bojan Simoski (VUA) Untangling the Puzzle of Digital Health Interventions

2023-02 Mariana Rachel Dias da Silva (TiU) Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

2023-03 Shabnam Najafian (TUD) User Modeling for Privacy-preserving Explanations in Group Recommendations

2023-04 Gineke Wiggers (UL) The Relevance of Impact: bibliometric-enhanced legal information retrieval

2023-05 P.A. (Anton) Bouter (CWI) Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization Including Real-World Medical Applications

2023-06 António Pereira Barata (UL) Reliable and Fair Machine Learning for Risk Assessment

2023-07 Tianjin Huang (TU/e) The Roles of Adversarial Examples on Trustworthiness of Deep Learning

2023-08 Lu Yin (TU/e) Knowledge Elicitation using Psychometric Learning

2023-09 Xu Wang (VUA) Scientific Dataset Recommendation with Semantic Techniques

2023-10 Dennis J.N.J. Soemers (UM) Learning State-Action Features for General Game Playing

2023-11 Fawad Taj (VUA) Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications

2023-12 Tessel Bogaard (VUA) Using Metadata to Understand Search Behavior in Digital Libraries

2023-13 Injy Sarhan (UU) Open Information Extraction for Knowledge Representation

2023-14 Selma Čaušević (TUD) Energy resilience through self-organization

2023-15 Alvaro Henrique Chaim Correia (TU/e) Insights on Learning Tractable Probabilistic Graphical Models

2023-16 Peter Blomsma (TiU) Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters

2023-17 Meike Nauta (UT) Explainable AI and Interpretable Computer Vision – From Oversight to Insight

2023-18 Gustavo Penha (TUD) Designing and Diagnosing Models for Conversational Search and Recommendation

2023-19 George Aalbers (TiU) Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals

2023-20 Arkadiy Dushatskiy (TUD) Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning

2023-21 Gerrit Jan de Bruin (UL) Network Analysis Methods for Smart Inspection in the Transport Domain

2023-22 Alireza Shojaifar (UU) Volitional Cybersecurity

2023-23 Theo Theunissen (UU) Documentation in Continuous Software Development

2023-24 Agathe Balayn (TUD) Practices Towards Hazardous Failure Diagnosis in Machine Learning

2023-25 Jurian Baas (UU) Entity Resolution on Historical Knowledge Graphs

2023-26 Loek Tonnaer (TU/e) Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour

2023-27 Ghada Sokar (TU/e) Learning Continually Under Changing Data Distributions

2023-28 Floris den Hengst (VUA) Learning to Behave: Reinforcement Learning in Human Contexts

2023-29 Tim Draws (TUD) Understanding Viewpoint Biases in Web Search Results

==== 2024 ====

2024-01 Daphne Miedema (TU/e) On Learning SQL: Disentangling concepts in data systems education

2024-02 Emile van Krieken (VUA) Optimisation in Neurosymbolic Learning Systems

2024-03 Feri Wijayanto (RUN) Automated Model Selection for Rasch and Mediation Analysis

2024-04 Mike Huisman (UL) Understanding Deep Meta-Learning

2024-05 Yiyong Gou (UM) Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

2024-06 Azqa Nadeem (TUD) Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

2024-07 Parisa Shayan (TiU) Modeling User Behavior in Learning Management Systems

2024-08 Xin Zhou (UvA) From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation

2024-09 Giso Dal (UT) Probabilistic Inference Using Partitioned Bayesian Networks

2024-10 Cristina-Iulia Bucur (VUA) Linkflows: Towards Genuine Semantic Publishing in Science

2024-11 Mahmoud Shokrollahi-Far (TiU) Computational Reliability of Quranic Grammar

2024-12 Peide Zhu (TUD) Towards Robust Automatic Question Generation For Learning