QRelScore: Better Evaluating Generated Questions with Deeper Understanding of Context-aware Relevance

Xiaoqiang Wang[1*], **Bang Liu**[2*†], **Siliang Tang**[1‡] and **Lingfei Wu**[3‡]
[1]Zhejiang University, [2]Université de Montréal & Mila, [3]Pinterest
{xq.wang, siliang}@zju.edu.cn
bang.liu@umontreal.ca, lwu@email.wm.edu

Abstract

Existing metrics for assessing question generation not only require costly human reference but also fail to take into account the input context of generation, rendering the lack of deep understanding of the relevance between the generated questions and input contexts. As a result, they may wrongly penalize a legitimate and reasonable candidate question when it (i) involves complicated reasoning with the context or (ii) can be grounded by multiple evidences in the context. In this paper, we propose **QRelScore**, a context-aware **Rel**evance evaluation metric for **Q**uestion Generation. Based on off-the-shelf language models such as BERT and GPT2, QRelScore employs both word-level hierarchical matching and sentence-level prompt-based generation to cope with the complicated reasoning and diverse generation from multiple evidences, respectively. Compared with existing metrics, our experiments demonstrate that QRelScore is able to achieve a higher correlation with human judgments while being much more robust to adversarial samples.

1 Introduction

Question generation (QG) systems aim to generate natural language questions that are relevant to and usually can be answered by a given piece of input text (Chen et al., 2019c; Liu et al., 2019a, 2020). QG can be used to improve various applications, such as question answering (QA) (Chen et al., 2019a; Fabbri et al., 2020; Yu et al., 2020b; Cheng et al., 2021), conversational systems (Wang et al., 2018; Chen et al., 2019b), and information retrieval (IR) (Yu et al., 2020a; Zamani et al., 2020). Meanwhile, it has long been criticized that QG models usually suffer from the semantic drift problem owing to the widely adopted likelihood-based training, *i.e.* the models ask questions that are not relevant to and can not be supported by the context (Zhang and Bansal, 2019; Chen et al., 2020). Thus, how to accurately evaluate the relevance between generated questions and the context is attracting more and more attention. One of the most accurate evaluation methods is human evaluation. However, human evaluation is expensive, time-consuming, and non-reproducible. Therefore, it is necessary to develop automatic evaluation metrics for question generation systems.

Traditional automatic metrics (*e.g.* BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005)) measure the n-gram overlap between the candidate and corresponding reference question, but they often fail to robustly match paraphrases. More recently, Q-BLEU (Nema and Khapra, 2018) and BERT-based metrics such as BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019) and LS_Score (Wu et al., 2020) were proposed to evaluate the answerability and semantic similarity of a candidate question, achieving better correlation with human judgments. However, on the one hand, they compute the similarity between the system output and the reference without considering the crucial input context of generation. Therefore, they cannot properly capture the reasoning relationship between the generated output and input context. On the other hand, comparing with a reference question omits the incompleteness of the reference: we can ask different questions based on the same context by paying attention to different information (or evidence) in it, while the reference question only represents one possible output. As a result, existing QG or text generation metrics struggle in evaluating the quality of candidate questions that (i) involve complicated reasoning with the context, or (ii) are generated from the evidence in the context that differs from

[*]Equal contribution.
[†]Canada CIFAR AI Chair.
[‡]Corresponding authors.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 562 - 581
December 7-11, 2022

CONTEXT. [. . .] in 1987, when some students believed that the observer began to show a conservative bias, a liberal newspaper, Common Sense was published [. . .]				
REFERENCE. *when was Common Sense published for the first time?*	**BLEU4**	**ROUGE-L**	**Q-BLEU**	**BERTScore**
Q_1 **Candidate.** when was Common Sense first published?	0.325	0.643	0.800	0.791
Q_2 **Unanswerable.** who was Common Sense published for the first time?	0.863	0.888	0.417	0.998
Q_3 **Paraphrasing.** in what year did Common Sense begin publication?	0.000	0.232	0.276	0.671
Q_4 **Coreference.** in what year did the student liberal newspaper begin publication?	0.000	0.106	0.053	0.291
Q_5 **Other evidences.** when did the observer begin to show a conservative bias?	0.000	0.212	0.427	0.265

Table 1: Five generated questions, the context, the ground-truth answer span (colored in green) that the question is generated for, and the human reference. We box the cases where the well-formed and meaningful candidates are scored much lower than the candidate Q_1. In contrast, the unanswerable adversarial example with a higher score than the candidate Q_1 is marked in red.

the reference questions.

Table 1 exemplifies some weaknesses of previous metrics. As shown in the table, BLEU4 and ROUGE-L cannot detect the unanswerable question (Q_2) and wrongly score the other well-formed candidates (Q_3 - Q_5) significantly lower than the candidate Q_1. Although Q-BLEU successfully penalizes the unanswerable question, it fails to discern the complicated but beneficial paraphrasing candidate (Q_3). BERTScore leverages contextualized embeddings from BERT (Devlin et al., 2019) and shows some degree of ability to distinguish the paraphrasing candidate, but it cannot perform linguistic reasoning related to the context (such as coreference resolution for Q_4) and scores the legitimate novel generation from other evidence (Q_5) much lower than the candidate Q_1.

In this paper, we present `QRelScore` , an automatic reference-free evaluation metric for question generation (QG). `QRelScore` addresses the weaknesses above by considering the context-aware relevance in a word- and sentence-level manner. On the one hand, inspired by the hierarchical procedure taken by masked language models such as BERT to understand a question (van Aken et al., 2019), `QRelScore` understands the word-level relevance by explicitly capturing the reasoning relationship between the candidate tokens and the context tokens. On the other hand, based on the benefit of intra-sentence coherence in the autoregressive language models such as GPT2 that originates from the word-by-word nature of human language production, the sentence-level relevance is measured by the overall factual consistency between the candidate and all the possible evidences in the context.

We verify the effectiveness and efficiency of `QRelScore` through various experiments. First, we demonstrate that `QRelScore` can improve the performance of question answering: by serving as a reward to train a QG model with reinforcement learning and then use it to augment a QA dataset (*e.g.* the SQuAD dataset (Rajpurkar et al., 2016)), the performance of a QA model can be improved by fine-tuning on the augmented dataset. Second, `QRelScore` achieves a state-of-the-art correlation with human judgments on the candidates generated by the existing QG models. Furthermore, when considering the available human reference of the dataset in `QRelScore`, we present a reference-augmented version, `Ref-QRelScore`, which achieves an even higher correlation. Last, extensive experiments on the robustness test also demonstrate that `QRelScore` has a stronger ability to discriminate against adversarial samples when compared to existing metrics.

2 `QRelScore` Metric

In this section, we formulate our reference-free evaluation metric `QRelScore` based on the off-the-shelf pre-trained language models. Specifically, `QRelScore` consists of two scoring components: the local relevance matching ($\texttt{QRel}_{LRM}$) component and the global relevance generation ($\texttt{QRel}_{GRG}$) component. The former is used to handle the candidates involving complicated reasoning with the contexts by computing word-level similarity using layer-wise embeddings and cross attention, while the latter is responsible for measuring the factual consistency between the candidate and all evidences of a given answer by comparing the difference in the confidence of generating the context with or without a prompt. Based on the local and global relevance measurement, `QRelScore` can not only handle the candidate involving complicated reasoning with the context but also pay equal attention to all evidences of a given answer in the context and ensure the fluency of generation.

Figure 1 illustrates the computation of $\texttt{QRel}_{LRM}$ and $\texttt{QRel}_{GRG}$. Given a candidate question

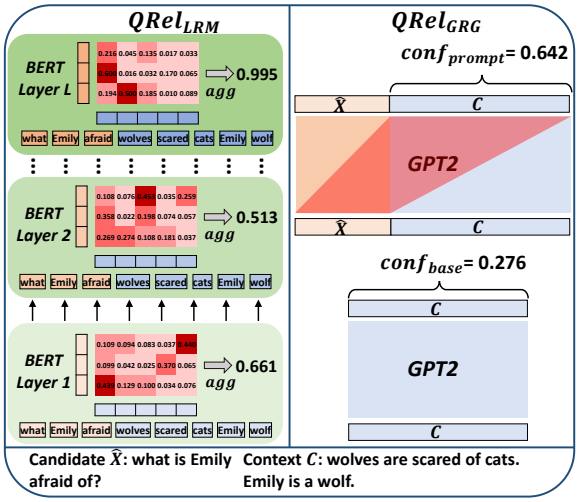

Figure 1: Illustration of the computation of our QRel_{LRM} (*left* part) and QRel_{GRG} (*right* part). QRel_{LRM} is based on the layer-wise cross attention by feeding the candidate and context together into BERT, while QRel_{GRG} is formulated as the confidence gain obtained by employing the candidate as a prompt of GPT2. Several stop words that give uniform attention to context tokens are omitted in the attention maps.

$\hat{X} = \langle \hat{x}_1, \cdots, \hat{x}_m, \cdots, \hat{x}_M \rangle$ and its context $C = \langle c_1, \cdots, c_n, \cdots, c_N \rangle$, $\texttt{QRelScore}$ is computed as the harmonic mean of $\texttt{QRel}_{LRM}$ and $\texttt{QRel}_{GRG}$:

$$\texttt{QRelScore}(\hat{X}, C) = 2\frac{\texttt{QRel}_{LRM} \cdot \texttt{QRel}_{GRG}}{\texttt{QRel}_{LRM} + \texttt{QRel}_{GRG}} \tag{1}$$

2.1 Local Relevance Matching

$\texttt{QRel}_{LRM}$ computes the word-level similarity between the candidate and context using the layer-wise contextualized embeddings and the cross attention between them. We firstly obtain the dynamic contextualized embeddings as follows:

$$\left\{ a^l_{mn}, f\left(\hat{x}_m\right)^l, f\left(c_n\right)^l \right\}_{l=1}^{L} = \text{BERT}([\hat{X}, C]) \tag{2}$$

where a^l_{mn} represents the maximum of normalized attention scores among all heads at the l-th layer of BERT between the m-th token in the candidate and the n-th token in the context, while $f\left(\hat{x}_m\right)^l$ and $f\left(c_n\right)^l$ denote the contextualized embeddings of corresponding tokens at the same layer, and $[\hat{X}, C]$ means the concatenation of the candidate and the context. Then, our $\texttt{QRel}_{LRM}$ is computed as the precision-based cosine similarity of tokens from the candidate and the context, where each token in the candidate is matched to the token in the context with an aggregate function and dynamic weighting. After that, our $\texttt{QRel}_{LRM}$ merges the layer-wise relevance score with power means (Rücklé et al., 2018), which is an effective generalization of pooling techniques for multi-level information.

$$\texttt{QRel}_{LRM} = \sqrt[p]{\frac{1}{L}\sum_{l=1}^{L} \text{Prec}^p_{l\text{-th}}} \tag{3}$$

$$\text{Prec}_{l\text{-th}} = \frac{1}{M}\sum_{m}^{M} \underset{c_n \in C}{agg}\left(a^l_{mn} f\left(\hat{x}_m\right)^l \odot f\left(c_n\right)^l \right) \tag{4}$$

where p, $\odot$ and $agg(\cdot)$ represent the exponent of power means, the cosine similarity and an aggregate function, respectively. Empirically, we set $p = 1$ and the $agg(\cdot)$ as a max function.

In practice, we observe that the layer-wise scores are in a more limited range (around $0.7 \sim 1.0$), potentially because of the learned geometry of contextualized embeddings from language models. Following the widely adopted solutions (Zhang et al., 2019; Hessel et al., 2021), we linearly rescale[1] $\texttt{QRel}_{LRM}$ with its lower bound b_{LRM} as a baseline to put it between 0 and 1.

$$\texttt{QRel}_{LRM} = \frac{\texttt{QRel}_{LRM} - b_{LRM}}{1 - b_{LRM}} \tag{5}$$

Empirically, we compute the b_{LRM} by averaging $\texttt{QRel}_{LRM}$ on the random $\langle$ *candidate*, *context* $\rangle$ pairs on the corresponding dataset.

Although the contextualized embeddings have been introduced in the evaluation of the text generation task, there are two critical differences in its utilization between our $\texttt{QRel}_{LRM}$ and previous works such as BERTScore (Zhang et al., 2019) and Moverscore (Zhao et al., 2019). First, we feed the candidate and the context into the model together, whereas previous works feed them in a 2-step division, first for the candidate and then for the context. Therefore, we can leverage cross attention between the candidate and the context to weigh the importance of every token better than previous works, whose weighting are based on the hand-crafted inverse document frequency (IDF). Because the IDF weighting only considers the static and independent token-level distribution over the whole candidate set, ignoring the specificity of certain a sample, they may wrongly encourage a token that is rare in the candidate set but occurs many times in the sample (*e.g.* proper nouns). Besides, Yi et al. (2020)

[1] Notice that the max-min normalization has the same effect as this baseline re-scaling. Please refer to Appendix F for more details and justification for our re-scaling.

Context. *Jack drove his car to the bazaar to purchase milk and honey for his large family.*
Reference (0.905). Where did Jack buy his milk and honey?
Entity swap (0.816). Where did Jack buy his car?
Pronoun swap (0.847). Where did Jack buy your milk and honey?
Sentence negation (0.803). Where didn't Jack buy his milk and honey?

Figure 2: Three unanswerable example questions constructed by perturbing only the individual words. Their $\texttt{QRel}_{LRM}$ scores (marked in the round brackets) do not reflect the factual inconsistency ideally.

demonstrates that the tokens with high IDF are not always indicative of semantic similarity due to the co-occurrences. Second, attention from different representation layers of BERT has been proven with different semantic and reasoning abilities (van Aken et al., 2019). For example, the shallow layer is used for named entity labeling, the middle layer for coreference resolution, and the deep layer for relation classification. Thereby, layer-wise contextualized embeddings and attention of BERT can be engaged to capture different relationships between tokens to evaluate the word-level relevance reasonably and hierarchically, *i.e.* approximately from superficial relationships to complicated ones.

2.2 Global Relevance Generation

Although $\texttt{QRel}_{LRM}$ can measure the word-level relevance of QG, candidates that contain a group of semantically similar tokens to the context, but ungrammatical or incoherent, can also receive a relatively high score. In this case, $\texttt{QRel}_{LRM}$ fails to ideally penalize the factual inconsistency arising from the individual words and capture multiple evidences in the context. Figure 2 shows some pitfalls of $\texttt{QRel}_{LRM}$. To mitigate this problem and achieve a robust measure of the global relevance, we further devise $\texttt{QRel}_{GRG}$ based on the prompt of causal language models (CLMs) such as GPT2.

Prompt-based learning maximizes the generalization capability of language models and is becoming a new paradigm in natural language processing (Liu et al., 2021a). In this paper, we formulate our $\texttt{QRel}_{GRG}$ as the confidence gain by *comparing the likelihood of generating the context with or without the candidate as a prompt*. Our $\texttt{QRel}_{GRG}$ appropriately encourages the candidate that is highly relevant to the context because a question inconsistent with the context is pretty likely to make a limited or even negative difference to the unidirectional generation. Based on the confidence difference caused by the candidate, $\texttt{QRel}_{GRG}$ measures the overall relevance between the candidate and all the possible evidences in the context.

More precisely, causal language modeling, also known as autoregressive language modeling, is a classic probabilistic density estimation problem. Given an input sequence $S = \langle s_1, \cdots, s_t, \cdots, s_T \rangle$, its joint distribution $p(S)$ or $p(s_{1:T})$ can be decomposed as:

$$p(S) = \prod_{t=1}^{T} p\left(s_t \mid s_{0:t-1}\right) \tag{6}$$

where s_0 is a special token indicating the begin of sequence and $p\left(s_t \mid s_{0:t-1}\right)$ represents the tractable conditional probabilities $p\left(s_t \mid s_0, \cdots, s_{t-1}\right)$. Abbreviating $p\left(s_t \mid s_{0:t-1}\right)$ as p_{s_t}, we feed the C and $[\hat{X}, C]$ into the GPT2 successively to obtain the conditional probability of every token in the context as follows:

$$\{p_{c_n}\}_{n=1}^{N} = \text{GPT2}\left(C\right) \tag{7}$$

$$\{p'_{\hat{x}_m}\}_{m=1}^{M}, \{p'_{c_n}\}_{n=1}^{N} = \text{GPT2}([\hat{X}, C]) \tag{8}$$

After that, the baseline confidence Conf_{base} and prompted confidence Conf_{prompt} are computed as: $\text{Conf}_{base} = \sum_{n=1}^{N} \log p_{c_n}$ and $\text{Conf}_{prompt} = \sum_{n=1}^{N} \log p'_{c_n}$, respectively. Finally, our $\texttt{QRel}_{GRG}$ is quantified as the gain ratio of the confidence caused by the candidate.

$$\texttt{QRel}_{GRG} = \max\left\{\frac{\text{Conf}_{prompt} - \text{Conf}_{base}}{|\text{Conf}_{base}|}, 0\right\} \tag{9}$$

For the same reason as $\texttt{QRel}_{LRM}$, we rescale the $\texttt{QRel}_{GRG}$ with b_{GRG} to increase the readability of this score and without its ranking ability or correlation with human judgments.

2.3 Reference-augmented QRelScore

QRelScore can additionally be extended to incorporate references if they are available. Specifically, given a set of human references R, Ref-QRelScore is computed as the arithmetic mean of QRelScore between the candidate and context, and maximal QRelScore between the candidate and reference.

$$\texttt{Ref-QRelScore}(\hat{X}, C, R) = \frac{1}{2}(\texttt{QRelScore}(\hat{X}, C) + \max_{r \in R} \texttt{QRelScore}(\hat{X}, r)) \tag{10}$$

3 Experiments

Datasets. We employ two widely-used QG datasets to validate QRelScore, including SQuADv1 (Rajpurkar et al., 2016) and HotpotQA (Yang et al., 2018). We re-divide the SQuADv1 dataset into train/dev/test splits following Zhou et al. (2017). For the HotpotQA dataset, we utilize the official train/dev/test splits.

Candidate questions. We obtain two candidate sets of shallow questions (*i.e.* factoid questions) respectively from NQG++ (Zhou et al., 2017) and BART-QG (Lewis et al., 2020) on the SQuADv1 dataset, and another two candidate sets of more complicated questions that require reasoning over multiple pieces of information respectively from DP-Graph (Pan et al., 2020) and DCQG (Cheng et al., 2021) on the HotpotQA dataset.

Implementation details. Our $\texttt{QRel}_{LRM}$ and $\texttt{QRel}_{GRG}$ are implemented by BERT-base and OpenAI GPT2 English models, respectively. The contextualized embeddings and attention scores of BERT-base and generation likelihood of GPT2 are extracted by the HuggingFace Transformers package (Wolf et al., 2020). In case of the input exceeding the maximum length acceptable to the language models (*i.e.* 512 and 1024 tokens for BERT and GPT2, respectively), we first cut the long context into several text chunks with maximum acceptable length. They are then fed into the model one by one, along with the candidate question. After that, the final score is calculated by averaging the relevance scores across all chunks. To perform rigorous analysis, we adopt the bootstrapping method (p-value $<$ 0.05) (Koehn, 2004) for pair-wise statistical significance tests in the following experiments. *Please refer to Appendix F for more details.*

Baselines. We verify the effectiveness of QRelScore by comparing it to the following three types of evaluation metrics. Firstly, we choose traditional n-gram matching based metrics including BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005). Furthermore, we also extend more recent reference-based methods as baselines such as Q-BLEU (Nema and Khapra, 2018), BERTScore (Zhang et al., 2019), Moverscore (Zhao et al., 2019), BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020). Among them, the last two baselines are supervised metrics optimized by the regression and ranking objective, respectively. In addition, we construct two reference-free baselines by replacing the reference input of Q-BLEU and BERTScore with the corresponding context, which is denoted as Q-BLEU$_{free}$ and BERTScore$_{free}$, respectively. At last, we adopt two state-of-the-art reference-free factuality evaluation metrics in the abstractive summarization task as our baselines, including the embedding-based consistency dimension of CTC (Deng et al., 2021) and the faithfulness dimension of BARTScore (Yuan et al., 2021).

Human annotation. Because the examined QG models do not release corresponding human evaluation results on the quality of their generated questions, we first evaluate the quality of the generated candidate via voluntary human evaluation. Following the human criteria of QG elaborated by Rus et al. (2010) and Nema and Khapra (2018), we annotate each sample in terms of grammaticality, answerability, and relevance. Specifically, we ask five annotators to rate the quality of 1,600 $\langle$ *passage*, *question*, *answer* $\rangle$ candidates from the four models, including NQG++, BART-QG, DP-Graph and DCQG , with 400 candidates per model. All the samples are randomly shuffled and anonymized. The annotators are informed of the detailed annotation instruction with clear scoring examples and evaluate the grammaticality, answerability and relevance on a three-point Likert scale (1 for "poor", 2 for "average", and 3 for "good"). *Please refer to Appendix A for more details about the annotation.*

3.1 Main Results

Human vs. human correlation. The inter-annotator Krippendorff's α for the three dimensions are 82.81, 85.25, and 87.39, respectively, which demonstrates an acceptable level of agreement ($>$ 80%) between annotators (Krippendorff, 2004). We use the average of five corresponding annotator ratings as the final human judgment for a specific dimension of a given candidate question.

Human vs. metrics correlation. Table 2 presents segment-level correlation to human judgments on SQuADv1. We observe that QRelScore consistently outperforms all the baselines in terms of answerability and relevance, which indicates the effectiveness of incorporating context-aware relevance into the evaluation of QG. In addition, the better grammaticality correlations can be attributed to the autoregressive language model in QRelScore, which measures the naturalness and fluency of the candidate more accurately by consid-

Metrics	Grammaticality			Answerability			Relevance		
	r	ρ	τ_b	r	ρ	τ_b	r	ρ	τ_b
BLEU-4	0.153	0.145	0.144	0.198	0.179	0.139	0.135	0.111	0.102
ROUGE-L	0.186	0.178	0.177	0.227	0.208	0.163	0.162	0.140	0.125
METEOR	0.200	0.191	0.190	0.241	0.221	0.173	0.174	0.153	0.135
Q-BLEU	0.317	0.308	0.305	0.347	0.326	0.259	0.273	0.258	0.219
BERTScore	0.352	0.345	0.341	0.380	0.360	0.285	0.303	0.289	0.244
MoverScore	0.372	0.364	0.359	0.396	0.375	0.301	0.319	0.306	0.257
BLEURT	0.391	0.383	0.377	0.412	0.391	0.315	0.334	0.322	0.269
COMET	0.446	0.433	0.432	0.461	0.442	0.353	0.381	0.370	0.307
Q-BLEU$_{free}$	0.379	0.371	0.367	0.402	0.384	0.306	0.324	0.313	0.260
BERTScore$_{free}$	0.415	0.408	0.403	0.434	0.414	0.332	0.356	0.344	0.286
CTC	0.448	0.440	0.435	0.466	0.444	0.355	0.384	0.375	0.309
BARTScore	0.454	0.447	0.444	0.472	0.454	0.360	0.391	0.378	0.316
QRelScore	0.497	0.488	0.485	0.513	0.494	0.394	0.424	0.417	0.347
Ref-QRelScore	**0.517**	**0.508**	**0.504**	**0.529**	**0.510**	**0.405**	**0.442**	**0.436**	**0.359**

Table 2: Segment-level correlation in Pearson's r, Spearman's ρ, and Kendall's τ_b with human judgments on the SQuADv1 dataset. The best and second-best results are **bold** and underlined, respectively.

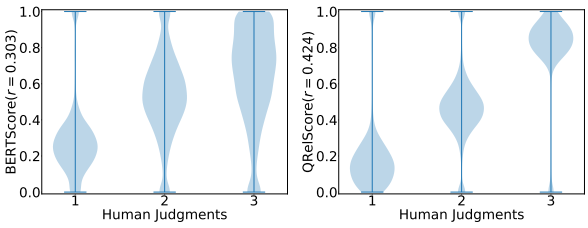

Figure 3: Score distributions of BERTScore and QRelScore under different relevance ratings (*i.e.* 1-3) of human judgments.

ering the word-by-word human language properties. When incorporating the available human reference into our metric, Ref-QRelScore achieves an even higher correlation with human judgments.

Qualitative results. In Figure 3, we take a closer look at the correlation results by the distribution of scores. Results reveal that previous metrics such as BERTScore can correctly assign lower scores to the candidates of low quality (rating "1"), but it performs poorly in the candidates of high quality (rating "2-3"). Moreover, it is worth noting that these underrated samples make up the majority of the whole candidate sets (*i.e.* more than 60% in the average of the candidate sets, see Appendix B for details). Conversely, QRelScore can clearly distinguish the candidates with different qualities. In Figure 4, we further show several qualitative examples that are annotated with high relevance and quality but scored significantly different by other metrics and QRelScore. We observe that QRelScore provides a consistent gauge with human judgments (relevance ratings), whereas other metrics cannot to handle the reasoning relationship (*i.e. separation of powers* refers to *the principle* in Example 1) and novel generation from multiple evidences (*i.e.* the answer is relevant to two facts, *the movie Obsessed* and *the two actors in it* in Example 2). In summary, these findings agree with the motivation of our work, namely that lacking a deep understanding of the context-aware relevance may lead to a wrong penalization to the legitimate and reasonable candidate. *Please refer to Appendix B for more experimental results.*

Example 1. *During the age of enlightenment, philosophers such as **John Locke** advocated the principle in their writings [...] separating the legislature, the executive, and the judiciary.*
Reference. Who was an advocate of separation of powers?
Candidate. Who advocated the principle in the age of enlightenment?
Human: 1.000 , **QRelScore:** 0.915 , **BLEU4:** 0.000 , **BERTScore:** 0.445 , **BARTScore:** 0.403

Example 2. *The fight scene finale between Sharon and the character played by Ali Larter, from the movie Obsessed, won **the 2010 MTV Movie Award** for Best Fight.*
Reference. A fight scene from the movie, Obsessed, won which award?
Candidate. Which award did the fight scene between Sharon and the role of Ali Larter win?
Human: 1.000 , **QRelScore:** 0.924 , **BLEU4:** 0.000 , **BERTScore:** 0.342 , **BARTScore:** 0.768

Figure 4: Randomly sampled qualitative candidates evaluated by QRelScore and other metrics, all of which have been re-scaled to $[0, 1]$ on the candidate sets.

3.2 Ablation Analysis

We conduct our ablation experiments and summarize the quantitative results in Table 3 on a basis of the two scoring components of QRelScore, *i.e.* QRel$_{LRM}$ and QRel$_{GRG}$. The experiments involve the following three aspects, including the variants of QRel$_{LRM}$, the variants of QRel$_{GRG}$, and their combinations.

First, we study the easiest combination of the two scoring components and find out whether QRel$_{LRM}$ or QRel$_{GRG}$ alone is sufficient to evaluate the relevance of QG, verifying the individual contributions of QRel$_{LRM}$ and QRel$_{GRG}$, respectively.

The first two baselines compute the relevance score by QRel$_{LRM}$ or QRel$_{GRG}$ only, denoted as "QRel$_{LRM}$ (M_1)" and "QRel$_{GRG}$ (M_8)", respectively. As shown in the table, both QRel$_{LRM}$ and QRel$_{GRG}$ make significant contributions to the final performance. For example, both M_1 and M_8 also outperform previous metrics (in Table 2) in terms of three dimensions. This result attributes to

Name	Metrics	Grammaticality			Answerability			Relevance		
		r	ρ	τ_b	r	ρ	τ_b	r	ρ	τ_b
M_1	**QRel$_{LRM}$**	**0.478**	**0.471**	**0.467**	**0.494**	**0.477**	**0.380**	**0.412**	**0.402**	**0.332**
M_2	w/ first	0.370	0.364	0.362	0.394	0.376	0.300	0.319	0.304	0.256
M_3	w/ middle	0.406	0.397	0.395	0.430	0.410	0.326	0.349	0.337	0.281
M_4	w/ last	0.425	0.417	0.413	0.446	0.425	0.340	0.366	0.355	0.296
M_5	w/ specific	0.442	0.436	0.431	0.463	0.443	0.352	0.380	0.370	0.309
M_6	w/ average	0.444	0.437	0.431	0.462	0.441	0.354	0.381	0.370	0.309
M_7	w/ mover	0.464	0.456	0.448	0.478	0.457	0.368	0.395	0.386	0.321
M_8	**QRel$_{GRG}$**	**0.464**	**0.451**	**0.450**	**0.478**	**0.458**	**0.367**	**0.397**	**0.386**	**0.320**
M_9	w/ absolute	0.390	0.381	0.378	0.412	0.394	0.314	0.334	0.323	0.268
	QRelScore	**0.497**	**0.488**	**0.485**	**0.513**	**0.494**	**0.394**	**0.424**	**0.417**	**0.347**

Table 3: Segment-level correlation in terms of Pearson's r, Spearman's ρ, and Kendall's τ_b with human judgments on the SQuADv1 dataset. In the table, the *upper* part is for the ablation analysis of QRel$_{LRM}$, while the *lower* part is for QRel$_{GRG}$. The best results are highlighted in **bold**.

the incorporation of the word- and sentence-level relevance into the evaluation metrics.

Second, we study the variants of QRel$_{LRM}$ by considering the layers of cross-attention scores and the way it aggregates the semantically similar tokens in the context for a token in the candidate. Therefore, on the one hand, "QRel$_{LRM}$ w/ first (M_2)", "QRel$_{LRM}$ w/ middle (M_3)", "QRel$_{LRM}$ w/ last (M_4)" and "QRel$_{LRM}$ w/ specific (M_5)" use the first four layers (0, 1, 2, 3), the middle four layers (4, 5, 6, 7), the last four layers (8, 9, 10, 11) and specific four layers (0, 3, 7, 11) of BERT attention, respectively. The experimental results reveal that M_2, M_3, M_4 and M_5 degrade the performance w.r.t. M_1 in three dimensions, demonstrating the attention at different layers plays an irreplaceable role in final results. Among them, M_5 achieves the best correlation, which shows the necessity of evaluating the relevance in a progressive manner, that is, from the shallow layer to the deep one. On the other hand, "QRel$_{LRM}$ w/ average (M_6)" and "QRel$_{LRM}$ w/ mover (M_7)" substitute the max operation in Eq. 4 with an avg function and a sum function weighted by the probability transitive matrix, which is obtained by optimizing earth mover's distance (EMD) (Rubner et al., 1998) from the candidate to the context on each layer. According to the results in Table 3, M_6 and M_7 show worse correlation than M_1, verifying that the averaging aggregation and optimal transportation optimization result in a biased relevance evaluation. A possible reason is that they fail to capture the token-wise specificity because average-based aggregation weakens the effects of irrelevant tokens and hinders the discriminative ability of the metrics.

Third, we study the variants of QRel$_{GRG}$ by calculating the confidence gain in different approaches, *i.e.* direct subtraction of these two confidence probabilities or their relative value to baseline confidence. Hence, "QRel$_{GRG}$ w/ absolute (M_9)" computes the global relevance by directly subtracting the Conf$_{base}$ from Conf$_{prompt}$ in Eq. 9. From the results in Table 3, M_9 degrades performance w.r.t. M_8 significantly, showing that the absolute confidence gain is not a proper measurement for sentence-level relevance since it takes account of the factors unrelated to the generation quality, such as the length of the candidate and the domain effects of pre-trained language models.

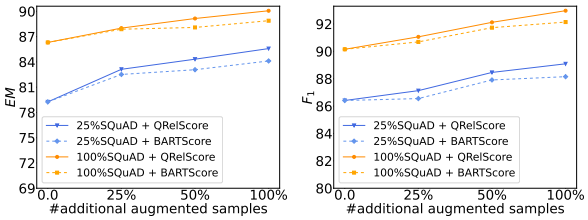

Figure 5: Performance of the DistilBERT-based QA system on the SQuADv1 dataset, augmented with the data generated by different QG rewards.

3.3 Evaluating QRelScore Rewards for QG with Reinforcement Learning (RL)

To further demonstrate the superiority of QRelScore, we employ the QRelScore as a reward to optimize an RL-based QG system and evaluate the quality of generated questions with a QA system. Specifically, we embed BART-QG into a self-critical sequence training (SCST) framework (Rennie et al., 2017) and compute the reward using QRelScore. After that, the whole pipeline is trained on the train split of the SQuADv1 dataset to generate questions conditioned on the context and answer. During the inference stage, the model is not fed into the unseen paragraphs (*i.e.* the paragraphs in the dev or test split) and generates diverse questions for the existing paragraphs in the SQuADv1 training set by keeping all beam search (size$=8$) outputs for each sample.

Furthermore, we filter out the obviously low-quality questions if their word counts are not between $6 \sim 30$, or if the answers directly appear in the questions. Finally, we randomly sample 90,000 QA pairs and augment the SQuADv1 training dataset with them. As a comparison, following the same setting as above, we design a baseline by employing BARTScore as the RL reward, which is one of the most competitive metrics in Table 2.

A DistilBERT-based (Sanh et al., 2019) QA model is trained on this augmented dataset to evalu-

Type	Method	SQuADv1	HotpotQA
Supervised models	DecAtt	0.791	0.641
	DIIN	0.852	0.718
	BERT	**0.943**	**0.801**
Metrics	BLEU-4	0.698	0.527
	ROUGE-L	0.703	0.533
	METEOR	0.712	0.542
	Q-BLEU	0.733	0.566
	BERTScore	0.740	0.575
	MoverScore	0.751	0.588
	BLEURT	0.773	0.612
	COMET	0.798	0.643
	Q-BLEU$_{free}$	0.767	0.606
	BERTScore$_{free}$	0.788	0.630
	CTC	0.808	0.653
	BARTScore	0.815	0.661
	QRelScore	**0.844**	**0.690**

Table 4: Area under the ROC curve (AUC) of classifying adversarial samples on SQuADv1 and HotpotQA datasets. The best results are highlighted in **bold**.

ate the quality of generated questions. The comparisons of QA performance in a high-resource setting (using the whole training set of SQuADv1) and a low-resource setting (using 25% of data sampled from SQuADv1) are illustrated in Figure 5. We can observe that BART-QG with QRelScore as the reward achieves better performance than BARTScore under both settings. As more and more of our generated data is added to the training set, the QA performance gets better and better and reaches a 4.36% / 3.13% improvement of EM/F_1 when the number of additional augmented samples reaches the size of the SQuADv1 training set. *Please refer to Appendix F for more implementation details.*

3.4 Robustness Analysis

A competent evaluation metric can not only distinguish between good and bad systems but also help analyze the samples (Zhang et al., 2019; Zhao et al., 2019). Therefore, we test the robustness of QRelScore by detecting adversarial samples. Specifically, inspired by the major types of relevance and factuality errors in the text generation (Goyal and Durrett, 2020; Chen et al., 2021; Pagnoni et al., 2021), we construct the positive samples by paraphrasing transformation. In contrast, negative samples are generated by the swapping and negation perturbations, including entity, pronoun swapping, and sentence negation. We generate 10,000 positive and 10,000 negative samples using the randomly chosen samples from the SQuADv1 and HotpotQA dev set as the positive anchors and employ the QRelScore to classify them based on the relevance scores. In addition to existing automatic metrics, we also fine-tune three supervised baselines, including DecAtt (Parikh et al., 2016), DIIN (Gong et al., 2018) and BERT (Devlin et al., 2019). We train them on the adversarial samples in a 5-fold cross-validation and report the results of validation sets as the final performance. *Please refer to Appendix C for the details on the adversarial examples.*

Table 4 reports the area under the ROC curve (AUC) for QRelScore and other baselines. As shown in the table, compared to the supervised BERT classifier, most of the metrics degrade performance significantly. However, some metrics, including QRelScore, outperform a relatively weak model (*i.e.* DecAtt). This suggests that these metrics have a certain level of ability to detect adversarial samples. Among all the metrics, QRelScore achieves the best results and shows the slightest performance drop on both datasets, showing more robustness than the other metrics.

4 Related Work

Aspect-specific evaluation. Some works measured semantic similarity between text by leveraging static word representations (Kusner et al., 2015; Lo, 2017), contextualized embedding (Zhang et al., 2019; Zhao et al., 2019), or fine-tuning on human-rated quality scores for different tasks to aggregate multiple features (Sellam et al., 2020; Rei et al., 2020). In a more unified formulation, the recent approaches devised a family of metrics to evaluate different text generation tasks. CTC (Deng et al., 2021) evaluated the information alignment between text from three aspects, including compression, transduction, and creation, while BARTScore (Yuan et al., 2021) gauged the text quality in a generative fashion and presented different evaluation aspects based on different generation directions. Although it was similar to QRel$_{GRG}$ of QRelScore in some way, it employed the absolute likelihood of generation and required extra fine-tuning to reduce the domain effects.

Relevance and factual consistency evaluation. Relevance is widely investigated in the response coherence of dialogue system (Huang et al., 2020) and factuality of document summarization (Gabriel et al., 2021) besides question generation. Kryscinski et al. (2020) proposed a weakly-supervised approach for verifying the factual consistency of a summary and identifying conflicts between source

documents and a generated summary. On a broader scale, Maynez et al. (2020) conducted an extensive human evaluation of several summarization systems and analyzed the types of factual hallucinations they produced. More recently, MARS (Liu et al., 2021b) was proposed to evaluate relevance by augmented references , which was generated by filling in the cloze templates according to the context. We considered lessons of context-awareness from these works when designing QRelScore.

5 Conclusion

Existing evaluation metrics for question generation are still reference-based and ignore the crucial input context of generation, lacking a deep understanding of the relevance between the generated questions and context. To address these issues, we propose QRelScore, which measures the word- and sentence-level relevance through the off-the-shelf language models. Extensive experiments demonstrate that QRelScore achieves start-of-the-art correlation with human judgments and makes up for the shortcomings of existing reference-based metrics.

Limitations

Our work proposes a new metric, namely QRelScore, to evaluate the quality of generated questions. The limitations are two-fold:

On the one hand, QRelScore is built on the pre-trained language models (PLMs) of general domains. Firstly, it is a black-box model that lacks interpretability in how the model predicts these evaluation scores. It might also perform biased evaluation because these models are pre-trained on heterogeneous web data and are shown to encode representational harms such as gender, race, and religion (Gonen and Goldberg, 2019; Liang et al., 2021). Moreover, herein we only aim to propose a general-purpose metric for the QG task and ignore some domain-specific analysis. We regard it as our future work and think that employing the domain-specific PLMs is a promising direction, *i.e.* MedBERT (Rasmy et al., 2021), a PLM on large-scale electronic health records, is used for the evaluation of medical questions, which can not only mitigate the human rating efforts in the medical domain but also improves the domain specialty of our metric. Last but not least, experimental results reveal significant room for improvement, *i.e.* ≈ 0.4 correlations of our proposed metrics to human judgments in Table 2, although it outperforms other baselines consistently. Appendix E provides several examples where QRelScore and human judgments are substantially different. In Appendix B, we improve the results through the larger model size (*i.e.* BERT-large and GPT2-medium) and more superior models (*i.e.* RoBERTa and XLNet). How to improve the efficiency of our QRelScore by using smaller PLMs but retaining similar performance, or how to boost the effectiveness of existing metrics by co-evolving the metric and corresponding generation systems, could be two interesting research topics.

On the other hand, following Nema and Khapra (2018), we verify the reasonability and superiority of our proposed metric by human evaluation on two typical datasets and limited PLM backbones. The QG tasks for cross-language or multi-language scenarios and framing additional evaluation protocols are left for our future work. Although we also conduct extra verification on downstream tasks, we advocate cautious and responsible practices in real-world deployment.

Acknowledgements

We would like to thank anonymous reviewers for their valuable comments and suggestions. This work has been supported in part by the Zhejiang NSF (LR21F020004), Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Chinese Knowledge Center of Engineering Science and Technology (CKCEST), the Canada CIFAR AI Chair Program, and the NSERC Discovery Grant (RGPIN-2021-03115).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2082–2095, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2019a. Bidirectional attentive memory networks for question answering over knowledge bases. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2913–2923, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2019b. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. *arXiv preprint arXiv:1908.00059*.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2019c. Reinforcement learning based graph-to-sequence model for natural question generation. In *The Eighth International Conference on Learning Representations (ICLR 2020)*.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2020. Toward subgraph guided knowledge graph question generation with graph neural networks. *arXiv preprint arXiv:2004.06015*.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5968–5978, Online. Association for Computational Linguistics.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. 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*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. 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*, pages 4508–4513, Online. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *International Conference on Learning Representations*.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

M Honnibal and I Montani. 2017. Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *https://spacy. io*.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings*

of ACL 2018, System Demonstrations, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043.

Bang Liu, Mingjun Zhao, Di Niu, Kunfeng Lai, Yancheng He, Haojie Wei, and Yu Xu. 2019a. Learning to generate questions by learning what not to generate. In *The World Wide Web Conference*, pages 1106–1118.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Ruibo Liu, Jason Wei, and Soroush Vosoughi. 2021b. Language model augmented relevance score. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6677–6690, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention

model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Christian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Bruce Thompson. 1995. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1823–1832.

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2193–2203, Melbourne, Australia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, Online. Association for Computational Linguistics.

Qian Yu, Lidong Bing, Qiong Zhang, Wai Lam, and Luo Si. 2020a. Review-based question generation with adaptive instance transfer and augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 280–290, Online. Association for Computational Linguistics.

Wenhao Yu, Lingfei Wu, Yu Deng, Ruchi Mahindru, Qingkai Zeng, Sinem Guven, and Meng Jiang. 2020b. A technical question answering system with transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 92–99, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 418–428.

Shiyue Zhang and Mohit Bansal. 2019. 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)*, pages 2495–2509, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. 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)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 662–671. Springer.

A Annotation Details

A total of five annotators participated in our study. The annotators were Computer Science graduates competent in English and kindly offered their help as volunteers without being compensated in any form. All the samples from the three examined models are randomly shuffled and anonymized, and each sample is evaluated by the following three dimensions:

- **Grammaticality.** It checks whether a question is well-formed. Annotators are asked to rate a sample as 3 for "no grammatical errors", 2 for "not grammatically correct but able to infer actual meaning", and 1 for "unacceptable".
- **Answerability.** As elaborated by Nema and Khapra (2018), this dimension checks whether a question is answerable according to the presence and correctness of important information such as named entities, content (relation) words, and question types. Annotators are asked to rate a sample as 3 for "all important information is present", 2 for "some important information is missing", and 1 for "all important information is missing".
- **Relevance.** Following the human criteria used in QG-STEC Task B (Rus et al., 2010), this dimension checks whether a question is consistent with the context and the given answer span. Annotators are asked to rate a sample as 3 for "Completely relevant to the context and given answer", 2 for partially relevant but unable to be grounded by the context, and 1 for "totally irrelevant".

In addition to the detailed annotation instruction, the annotators were also informed of the clear scoring examples as summarized in Table 6. As shown in Figure 9, we develop a web application to collect the evaluation results automatically. The software will provide candidate questions to the human annotators, guide them to perform annotation, and post their ratings back to our server. After that, we can analyze the final human judgments based on the results on our server.

B More Experimental Results

Human evaluation ratings of different candidate question sets are illustrated in Figure 6, which reflects how well the existing QG models perform

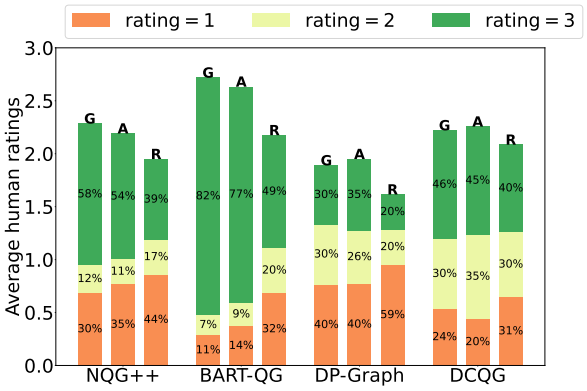

Figure 6: Bar illustration on human rating distributions of different candidate question sets in terms of grammaticality (**G**), answerability (**A**), relevance (**R**). The total length and coloring part of the bar respectively represent the average human ratings and the ratio of the corresponding rating on 1-3 scale (*i.e.* plotted in three colors).

Metrics	Grammaticality			Answerability			Relevance		
	r	ρ	τ_b	r	ρ	τ_b	r	ρ	τ_b
BLEU-4	0.117	0.108	0.108	0.165	0.146	0.112	0.104	0.078	0.076
ROUGE-L	0.150	0.142	0.141	0.194	0.175	0.136	0.131	0.107	0.099
METEOR	0.164	0.155	0.154	0.208	0.188	0.147	0.143	0.120	0.109
Q-BLEU	0.280	0.272	0.270	0.314	0.293	0.233	0.243	0.225	0.193
BERTScore	0.316	0.309	0.305	0.347	0.327	0.258	0.272	0.257	0.219
MoverScore	0.336	0.328	0.323	0.363	0.342	0.274	0.288	0.273	0.232
BLEURT	0.355	0.346	0.341	0.379	0.358	0.288	0.303	0.289	0.244
COMET	0.409	0.397	0.396	0.428	0.409	0.327	0.351	0.337	0.282
Q-BLEU$_{free}$	0.343	0.334	0.331	0.369	0.351	0.279	0.294	0.281	0.235
BERTScore$_{free}$	0.379	0.372	0.367	0.401	0.381	0.305	0.326	0.312	0.260
CTC	0.411	0.404	0.399	0.433	0.411	0.329	0.353	0.342	0.283
BARTScore	0.417	0.410	0.408	0.439	0.421	0.334	0.360	0.346	0.290
QRelScore	0.461	0.452	0.449	0.480	0.461	0.367	0.394	0.385	0.321
Ref-QRelScore	**0.481**	**0.472**	**0.468**	**0.496**	**0.477**	**0.379**	**0.411**	**0.403**	**0.333**

Table 5: Segment-level correlation in terms of Pearson's r, Spearman's ρ, and Kendall's τ_b with human judgments on the HotpotQA dataset. The best and second-best results are **bold** and underlined, respectively.

in terms of grammaticality, answerability, and relevance. We can see that most of the candidates ($>$ 70%) are annotated as high quality ("2-3" ratings), so a competent evaluation metric should encourage this kind of high-quality candidates. `QRelScore` serves as an automatic metric to evaluate the quality of candidate questions, then we conduct correlation analysis between the metric scores and corresponding human ratings.

Table 5 presents the segment-level correlation to human judgments on the HotpotQA dataset. We observe that `QRelScore` consistently outperforms all the baselines, which indicates the effectiveness of incorporating language models into the relevance evaluation of QG.

Figure 7 qualitatively illustrates the score distributions of COMET, BLEURT, Q-BLEU, and BLEU-4 under different relevance ratings of hu-

	Instruction	Context	Candidate question
Grammaticality	3 = No grammatical errors	[. . .] Denver linebacker Von Miller was named Super Bowl MVP, recording **five** solo tackles [. . .]	How many solo tackles did Von Miller make at Super Bowl 50?
	2 = Not grammatically correct but able to infer actual meaning	[. . .] Miami's Sun Life Stadium and the San Francisco Bay Area's **Levi's Stadium** [. . .]	What site is locate in the San Francisco Bay Area?
	1 = Unacceptable grammaticality	[. . .] Kubiak replacing **Elway** at the end of the Broncos' defeats in Super Bowls XXI [. . .]	Why was replaced of Kubiak in Super Bowl XXIV?
Answerability	3 = All important information is present	[. . .] **six-time** Grammy winner and Academy Award nominee Lady Gaga [. . .]	How many Grammys has Lady Gaga won?
	2 = Some important information is missing	[. . .] and one of the largest in East-Central Europe, employing **2,000** professors [. . .]	How many ~~professors~~ does the Warsaw University of Technology employ?
	1 = All important information is missing	[. . .] liberated by **Napoleon's** army in 1806, Warsaw was made the capital [. . .]	Whose ~~army liberated~~ Warsaw in ~~1806~~?
Relevance	3 = Completely relevant to the context and given answer	[. . .] **the Vistula River** is the specific axis of Warsaw, which divides the city into two parts [. . .]	What is the axis of Warsaw which divides it into two parts?
	2 = Partially relevant but unable to be grounded by the context	[. . .] within a greater metropolitan area of **2.666 million** residents [. . .]	How big is the greater metropolitan area?
	1 = Totally irrelevant	[. . .] transmitting mechanical energy with minimal loss over **any terrestrial distance** [. . .]	Who received a bid in 1935?

Table 6: Human annotation instructions along with the scoring examples for the grammaticality, answerability, and relevance dimension. The given answers and problematic words in corresponding candidate questions are marked in **bold** and red, respectively.

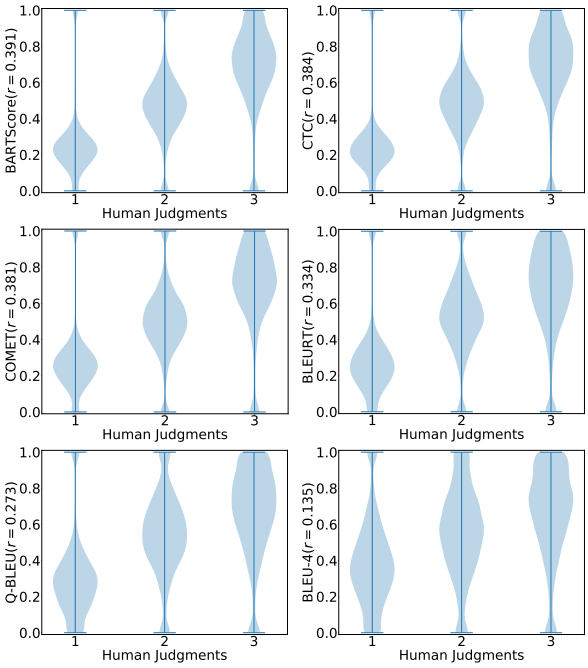

Figure 7: Score distributions of BARTScore, CTC, COMET, BLEURT, Q-BLEU and BLEU-4 under different relevance ratings (*i.e.* 1-3) of human judgments.

man judgments. These metrics poorly correlate with human judgments because they either assign relatively low scores to the candidates of high quality or score the candidates of a certain level of quality with high variance.

Last, we conduct additional experiments with other types of pre-trained language models, consisting of RoBERTa (Liu et al., 2019b) and XLNet (Yang et al., 2019). As shown in the Figure 8, the larger model size (*i.e.* BERT-large and GPT2-medium) and more superior models (*i.e.* RoBERTa and XLNet) improve the correlations with human judgments by a significant margin, showing that the stronger generalization ability of adopted language models contributes to a more robust and accurate evaluation of `QRelScore`. For a fair comparison with BERT-based baseline metrics, we report the final results using BERT-base and GPT2.

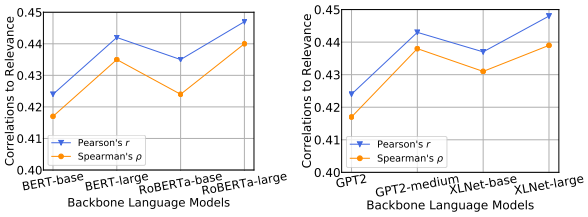

Figure 8: Segment-level correlations with human judgments when using different backbone language models for $\texttt{QRel}_{LRM}$ and $\texttt{QRel}_{GRG}$, respectively. When we change one of them, the others are fixed. Since both Spearman's ρ and Kendall's τ_b are rank-based correlation coefficients, we omit Kendall's τ_b for simplicity and report the results in terms of Pearson's r and Spearman's ρ.

C Adversarial Examples

As shown in Table 7, on the one hand, positive samples are constructed by **paraphrasing transformation**, which is implemented by back-translation with the multi-lingual MarianMTModel (Junczys-Dowmunt et al., 2018). The original sentence was translated to an intermediate language and translated back to English, yielding a semantically-equivalent sentence with minor syntactic and lexical changes. French, German, Chinese, Spanish,

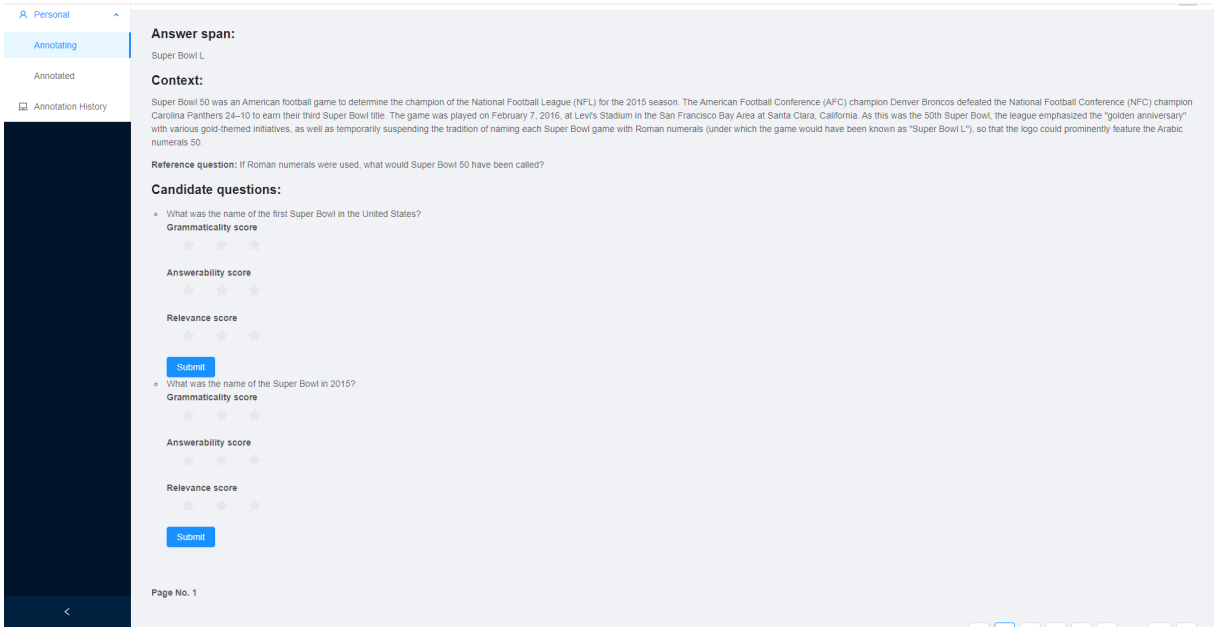

Figure 9: A screenshot of our human annotation process.

Transformation	Original question	Transformed question
Paraphrasing	On what date did the NFL announce that Coldplay would headline the half-time show?	When did the NFL announce that Coldplay would mark the title of the half-time program?
Entity swap	Into what language did Marlee Matlin translate the national anthem?	Into what language did Lady Gaga translate the national anthem?
Pronoun swap	In 2005, what did Doctor Who think the condition of his home planet was?	In 2005, what did Doctor Who think the condition of your home planet was?
Sentence negation	What controls wages in a purely capitalist mode of production?	What doesn't control wages in a purely capitalist mode of production?

Table 7: Examples of text transformations used to generate adversarial samples. Green and red text highlight the changes made by the transformation. Among these transformations, paraphrasing is a semantically invariant transformation, while sentence negation, entity swap, and pronoun swap are semantically variant transformations.

and Russian were used as intermediate languages. These languages were chosen based on the performance of current NMT systems with the expectation that well-performing languages could ensure better translation quality. On the other hand, negative samples are generated by the following perturbations:

- **Entity and pronoun swapping.** For entity extraction, a named entity recognition (NER) system is applied to both the reference question and the context to extract all mentioned entities. It divides them into four groups comprising named entities, covering persons, location/institution/organization names, and number entities. After that, the random entity sampled from the entity set is swapped within its corresponding group. In this work, we use the spaCy NER tagger (Honnibal and Montani, 2017). For pronouns, all gender-specific pronouns were first extracted from the reference question. Next, a randomly chosen pronoun was swapped with a different one from the same pronoun group to ensure syntactic correctness.
- **Sentence negation.** In the first step, the reference question is scanned in search of auxiliary verbs and modal verbs. Then, we randomly choose a verb and add *not* after it or use WordNet (Miller, 1995) wrapped in the NLTK (Bird et al., 2009) package to find its antonym to negate the sentence.

D Redundancy Analysis

Although `QRelScore` achieves a better correlation with human judgments than other metrics, it is unclear if individual metrics capture distinct or redundant dimensions of human judgment. For example, while $\texttt{QRel}_{LRM}$ and BERTScore both produce

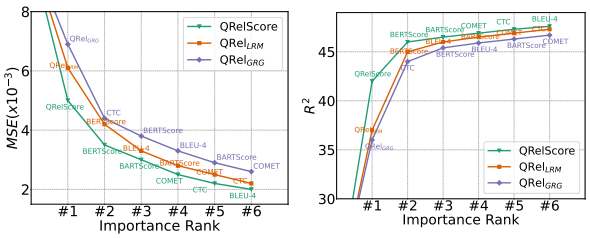

Figure 10: MES and R^2 for the forward-selection regression of metrics on the SQuADv1 dataset. Its horizontal axis represents which metric is most commonly chosen at each selection iteration, and a metric that is chosen earlier means more informativeness than the remaining metrics. Only the top-6 metrics are illustrated in this diagram.

relatively high correlation, are they redundant or complementary? This redundancy arises from the difference in the gold-standard input of `QRelScore` and other metrics. That is, we use the context as the input while others use the reference, and the content of the reference is usually contained within the corresponding context. Following Hessel et al. (2021), we seek a minimal set of metrics that explains the most variance in human judgment and fits it approximately. To be precise, we undertake a forward selection algorithm (Thompson, 1995) on the metrics set consisting of the baselines, `QRelScore`, $\texttt{QRel}_{LRM}$ and $\texttt{QRel}_{GRG}$. This algorithm performs an iterative greedy selection by picking the most informative additional metric from the metrics set and adding it to the target set, which is initially empty. In this work, we use the implementation of sklearn package (Pedregosa et al., 2011) and repeat the forward selection algorithm ten times in 5-fold cross-validation to perform rigorous analysis.

Figure 10 shows the information gain obtained by different metrics in terms of both mean squared error (MSE) and determination coefficient (R^2). On the on hand, we can see that `QRelScore`, $\texttt{QRel}_{LRM}$ and $\texttt{QRel}_{GRG}$ tend to be chosen early by the forward selection and make significant improvements to MSE and R^2. This result shows that our reference-free metrics contribute substantial information gain to fitting the human judgments. On the other hand, reference-based metrics such as BERTScore, BLEU-4, and BLEURT are chosen closely after our reference-free metrics, demonstrating that reference-free evaluation plays a complementary and not redundant role in measuring the overall relevance of QG.

Error	Example
Out of Vocabulary	**Context:** [. . .] The 2012 Washington State Cougars football team was coached by first-year head coach Mike Leach. [. . .] **Candidate:** *Where does UNK UNK currently coach at?* **Human:** 0.600 , 0.667 , 0.800 **QRelScore:** 0.198
Confusion	**Context:** [. . .] Jacob put the marbles in the box and the bowl on the table. [. . .] **Candidate:** *Where did he put the marbles?* **Human:** 1.000 , 0.333 , 0.867 **QRelScore:** 0.821
Domain-specific Knowledge	**Context:** [. . .] Denver continued to pound away as RB Cecil Sapp got a 4-yard TD run, while kicker Jason Elam got a 23-yard field goal. [. . .] **Candidate:** *Which position scored the shortest touchdown of the game?* **Human:** 1.000 , 1.000 , 0.933 **QRelScore:** 0.206

Table 8: Three typical types of errors found in the samples which received significant differences between the QRelScore and human judgments.

E Error Analysis

We analyze cases where the `QRelScore` substantially differs from human judgments. As shown in Table 8, these errors can be categorized into one of three types: (1) Out of vocabulary errors, often induced by unknown tokens in the candidates, (2) Confusion errors, the scope of coordination may be interpreted differently and thus lead to a syntactic ambiguity, *e.g.* in showing cases, the marbles were either put both in the box and in the bowl that was on the table, or the marbles were put in the box and the bowl was put on the table, and (3) Knowledge errors, where the candidates are further inferences based on the commonsense knowledge or domain-specific knowledge, *e.g.* in showing cases, both running back (RB) and kicker (K) are the positions of a player on an American football team. These errors reveal the limitations of `QRelScore` and give us directions for future improvement by engaging language models with a larger capacity.

F Implementation Details

Hyperparameters of `QRelScore`. The hyperparameters of `QRelScore`, *i.e.* b_{LRM} and b_{GRG}, are devised as a monotonic rescaling operation, which does not affect the ranking results and human correlations of `QRelScore`. For example, the layer-wise $\texttt{QRel}_{LRM}$ is inherently computed as the precision-

	Model	#Params	SQuADv1	HotpotQA
b_{LRM}	bert-base-cased	110M	0.691	0.541
	bert-large-cased	340M	0.612	0.505
	roberta-base	125M	0.678	0.556
	roberta-large	355M	0.642	0.549
b_{GRG}	gpt2	117M	0.546	0.327
	gpt2-medium	345M	0.435	0.303

Table 9: Baseline scores for different configurations of pre-trained language models and datasets.

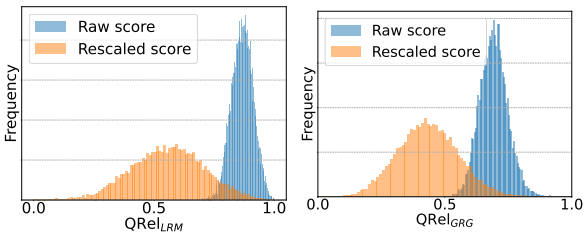

Figure 11: Relative frequency distribution of raw and rescaled metric scores on the SQuADv1 dataset. The exemplified $\texttt{QRel}_{LRM}$ and $\texttt{QRel}_{GRG}$ are computed with the BERT and GPT2, respectively. The rescaled metric scores range from $[0, 1]$ and show better readability.

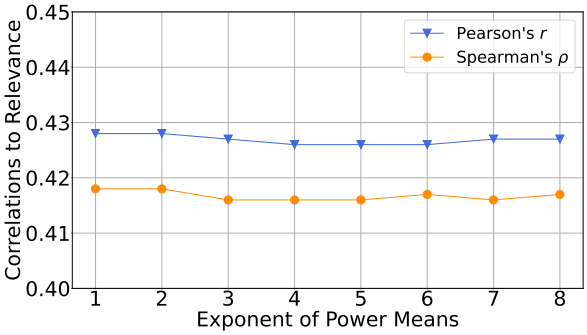

Figure 12: Segment-level correlations with human judgments when using different exponents for power means of $\texttt{QRel}_{LRM}$.

based similarity of tokens from the candidate and the context. While the cosine similarity, in theory, can range from $[-1, 1]$, we generally observe values ranging from roughly 0.7 to roughly 1.0. A possible reason for this observation is the learned geometry of contextualized embeddings. Following the widely adopted solutions (Zhang et al., 2019; Hessel et al., 2021), we seek to linearly rescale $\texttt{QRel}_{LRM}$ with its lower bound b_{LRM} in order to increase the readability of the metric score. We compute the b_{LRM} by averaging $\texttt{QRel}_{LRM}$ on the random $\langle$ *question*, *context* $\rangle$ pairs on the corresponding dataset. Specifically, for each dataset and language model, we create candidate $\langle$ *question*, *context* $\rangle$ pairs by grouping two different samples, one provides the question, the other provides the context. Then, we filter out the pairs with significantly high lexical overlapping (BLEU1 > 0.05) between the question and context and compute the mean $\texttt{QRel}_{LRM}$ on these random pairs as the b_{LRM}. In addition, $\texttt{QRel}_{GRG}$ has a similar observation as $\texttt{QRel}_{LRM}$ due to the incorporation of language models and computation of generation likelihood. The baseline scores for different datasets and language models are summarized in Table 9. Figure 11 shows the raw and rescaled metric scores of the SQuADv1 dataset. We can observe that the metric scores are linearly transformed from a more limited range to approximate $[0, 1]$ and show better readability.

Hyperparameters of power means. In Section 2, $\texttt{QRel}_{LRM}$ compute the overall relevance by aggregating the precision-based similarity of all the embedding layers by power means. Empirically, we perform ablation experiments using different exponents to calculate the power means and report the correlation results between QRelScore and human judgments. As shown in Figure 12, different exponents have a marginal effect on the correlation to human judgments, *i.e.* less than 0.002 correlation changes. In this work, following Rücklé et al. (2018); Zhao et al. (2019), we report the results by setting the exponent as $p = 1$.

Baseline metrics. Our baseline metrics encompass BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), Q-BLEU (Nema and Khapra, 2018), BERTScore (Zhang et al., 2019), Moverscore (Zhao et al., 2019), BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), Q-BLEU$_{free}$, BERTScore$_{free}$, CTC (Deng et al., 2021), and BARTScore (Yuan et al., 2021). The first three metrics are implemented by the Microsoft COCO evaluation scripts (Chen et al., 2015).

- **Q-BLEU** implementation is from the official repository at https://github.com/PrekshaNema25/Answerability-Metric. Following the paper's suggestion, we set the hyperparameters w_r, w_n, w_q and w_f as 0.1, 0.6, 0.2 and 0.1, respectively.
- **BERTScore** and **Moverscore** are computed using the released Python packages v0.3.11 https://pypi.org/project/bert-score/ and official repository at https://github.com/AIPHES/

emnlp19-moverscore, respectively. Their BERT embeddings are extracted with the Huggingface Transformers package (Wolf et al., 2020).

- **BLEURT** is a training-based metric, the architecture files and pre-trained parameters are from the official implementation at https://github.com/google-research/bleurt. The reported results are computed using the backbone bleurt-base-128.
- **COMET** original is a training-based metric that is devised for machine translation (MT). The architecture files and pre-trained parameters are from the official Python package v1.1.0 https://pypi.org/project/unbabel-comet/. The reported results are computed using the backbone wmt21-comet-qe-mqm.
- **Q-BLEU**$_{free}$ and **BERTScore**$_{free}$ replace the reference input of Q-BLEU and BERTScore with the corresponding context and adopt the same hyperparameters with the original metrics.
- **CTC** proposes a unified framework for different natural language generation (NLG) tasks from three categories, consisting of *compression*, *transduction*, and *creation*. The metric is trained to detect hallucinated tokens generated by a BART model in a self-supervised manner. We regard question generation as the *compression* task and report the corresponding CTC scores. Its implementation is from the released Python package v0.1.1 https://pypi.org/project/ctc-score/.
- **BARTScore** evaluates three different aspects corresponding to three different generative direction, including *faithfulness*, *precision*, and *recall*. Among them, the first aspect is a reference-free metric, while the others are reference-based. Considering the relevance aspect we concentrate on in this work, we report the *faithfulness* scores as the final results of BARTScore. Its implementation is based on the official repository at https://github.com/neulab/BARTScore. We use the version fine-tuned on the ParaBank2 dataset (Hu et al., 2019). Its original evaluating results are based on the log-likelihood and are negative values. To improve its readability, we report the BARTScore metrics score using max-min normalization, which does not affect its correlation with human judgments.

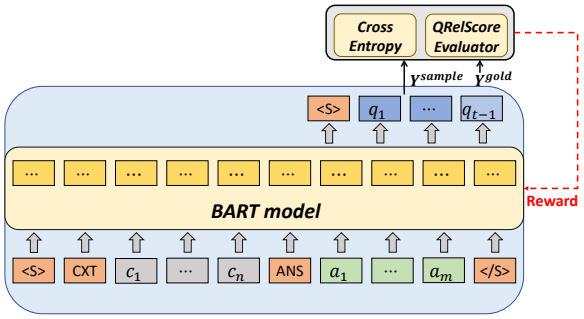

Figure 13: Illustration of BART-QG pipeline. It is optimized by a reinforcement learning algorithm, regarding QRelScore as the rewards.

QRelScore Rewards for QG. In Section 3.3, we employ the QRelScore as a reward to optimize a reinforcement learning-based QG system and evaluate the quality of generated questions with a QA system. As shown in Figure 13, we embed BART-QG into a self-critical sequence training (SCST) framework (Rennie et al., 2017) and compute the reward using QRelScore. Formally, given context c, answer a, and generated question $q = \langle q_1, \ldots, q_t, \ldots \rangle$, the loss function of SCST is defined as following policy gradients.

$$\mathcal{L}_{scst} = (r(\hat{Y}) - r(Y^s)) \sum_t \log P(q_t|c, a, q_{<t}) \tag{11}$$

$$r(\hat{Y}) = \text{QRelScore}(\hat{Y}, c) \tag{12}$$

$$r(Y^s) = \text{QRelScore}(Y^s, c) \tag{13}$$

where Y^s is the sampled output and $\hat{Y}$ is the baseline output, obtained by greedy search, that is, by maximizing the output probability distribution at each decoding step. Following the SCST setting (), We train the BART-QA in two stages. In the first state, we train the model using regular cross-entropy loss as:

$$\mathcal{L}_{lm} = \sum_t -\log P(q_t|c, a, q_{<t}) \tag{14}$$

In the second stage, we fine-tune the model by optimizing a mixed loss function combining cross-entropy loss and SCST loss as:

$$\mathcal{L} = \lambda \mathcal{L}_{scst} + \mathcal{L}_{lm} \tag{15}$$

where λ is a scaling factor controlling the trade-off between cross-entropy loss and SCST loss, which is linearly scheduled from 0.0 to 1.0 based on the training process.

DistilBERT QA model. In Section 3.3, A DistilBERT-based (Sanh et al., 2019) QA model is trained on this augmented dataset to evaluate the quality of generated questions. According to the common fine-tuning strategy of language models (Devlin et al., 2019), we use the pooled output of the DistilBERT-base model following a linear layer and sigmoid function as a pointer network. We use two pointer networks of the same structure to predict the beginning and ending position of an answer, respectively.