Summary-Oriented Question Generation for Informational Queries

Xusen Yin[*]
USC/ISI
xusenyin@isi.edu

Li Zhou
Amazon
lizhouml@amazon.com

Kevin Small
Amazon
smakevin@amazon.com

Jonathan May
USC/ISI
jonmay@isi.edu

Abstract

Users frequently ask simple factoid questions for question answering (QA) systems, attenuating the impact of myriad recent works that support more complex questions. Prompting users with automatically generated *suggested questions* (SQs) can improve user understanding of QA system capabilities and thus facilitate more effective use. We aim to produce self-explanatory questions that focus on main document topics and are answerable with variable length passages as appropriate. We satisfy these requirements by using a BERT-based Pointer-Generator Network trained on the Natural Questions (NQ) dataset. Our model shows SOTA performance of SQ generation on the NQ dataset (20.1 BLEU-4). We further apply our model on out-of-domain news articles, evaluating with a QA system due to the lack of gold questions and demonstrate that our model produces better SQs for news articles – with further confirmation via a human evaluation.

1 Introduction

Question answering (QA) systems have experienced dramatic recent empirical improvements due to several factors including novel neural architectures (Chen and Yih, 2020), access to pre-trained contextualized embeddings (Devlin et al., 2019), and the development of large QA training corpora (Rajpurkar et al., 2016; Trischler et al., 2017; Yu et al., 2020). However, despite technological advancements that support more sophisticated questions (Yang et al., 2018; Joshi et al., 2017; Choi et al., 2018; Reddy et al., 2019), many consumers of QA technology in practice tend to ask simple factoid questions when engaging with these systems. Potential explanations for this phenomenon include low expectations set by previous QA systems, limited coverage for more complex questions not changing these expectations, and users simply not possessing sufficient knowledge of the subject of interest to ask more challenging questions. Irrespective of the reason, one potential solution to this dilemma is to provide users with automatically generated *suggested questions* (SQs) to help users better understand QA system capabilities.

Generating SQs is a specific form of question generation (QG), a long-studied task with many applied use cases – the most frequent purpose being data augmentation for mitigating the high sample complexity of neural QA models (Alberti et al., 2019a). However, the objective of such existing QG systems is to produce large quantities of question/answer pairs for training, which is contrary to that of SQs. The latter seeks to guide users in their research of a particular subject by producing engaging and understandable questions. To this end, we aim to generate questions that are *self-explanatory* and *introductory*.

Self-explanatory questions require neither significant background knowledge nor access to documents used for QG to understand the SQ context. For example, existing QG systems may use the text *"On December 13, 2013, Beyoncé unexpectedly released her eponymous fifth studio album on the iTunes store without any prior announcement or promotion."* to produce the question *"Where was the album released?"* This kind of question is not uncommon in crowd-sourced datasets (e.g., SQuAD (Rajpurkar et al., 2016)) but do not satisfy the self-explanatory requirement. Clark and Gardner (2018) estimate that 33 % of SQuAD questions are context-dependent. This context-dependency is not surprising, given that annotators observe the underlying documents when generating questions.

Introductory questions are best answered by a larger passage than short spans such that users can learn more about the subject, possibly inspiring follow-up questions (e.g., "Can convalescent

[*] Work was done as an intern at Amazon.

Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 81–97
August 5–6, 2021. ©2021 Association for Computational Linguistics

plasma help COVID patients?”). However, existing QG methods mostly generate questions while reading the text corpus and tend to produce narrowly focused questions with close syntactic relations to associated answer spans. TriviaQA (Joshi et al., 2017) and HotpotQA (Yang et al., 2018) also provide fine-grained questions, even though reasoning from a larger document context via multi-hop inference. This narrower focus often produces factoid questions peripheral to the main topic of the underlying document and is less useful to a human user seeking information about a target concept.

Conversely, the Natural Question (NQ) dataset (Kwiatkowski et al., 2019) (and similar ones such as MS Marco (Bajaj et al., 2016), GooAQ (Khashabi et al., 2021)) is significantly closer to simulating the desired information-seeking behavior. Questions are generated independently of the corpus by processing search query logs, and the resulting answers can be entities, spans in texts (aka short answers), or entire paragraphs (aka long answers). Thus, the NQ dataset is more suitable as QG training data for generating SQs as long-answer questions that tend to satisfy our self-explanatory and introductory requirements.

To this end, we propose a novel BERT-based Pointer-Generator Network (BERTPGN) trained with the NQ dataset to generate introductory and self-explanatory questions as SQs. Using NQ, we start by creating a QG dataset that contains questions with both short and long answers. We train our BERTPGN model with these two types of context-question pairs together. During inference, the model can generate either short- or long-answer questions as determined by the context. With automatic evaluation metrics such as BLEU (Papineni et al., 2002), we show that for long-answer question generation, our model can produce state-of-the-art performance with 20.1 BLEU-4, 6.2 higher than (Mishra et al., 2020), the current state-of-the-art on this dataset. The short answer question generation performance can reach 28.1 BLEU-4.

We further validate the generalization ability of our BERTPGN model by creating an out-of-domain test set with the CNN/Daily Mail (Hermann et al., 2015). Without human-generated reference questions, automatic evaluation metrics such as BLEU are not usable. We propose to evaluate these questions with a pretrained QA system that produces two novel metrics. The first is *answerability*, measuring the possibility to find answers from given contexts. The second is *granularity*, indicating whether the answer would be passages or short spans. Finally, we conduct a human evaluation with generated questions of the test set and demonstrate that our BERTPGN model can produce introductory and self-explanatory questions for information-seeking scenarios, even for a new domain that differs from the training data.

The novel contributions of our paper include:

- We generate questions, aiming to be both introductory and self-explanatory, to support human information seeking QA sessions.
- We propose to use the BERT-based Pointer-Generator Network to generate questions by encoding larger contexts capable of resulting in answer forms including entities, short text spans, and even whole paragraphs.
- We evaluate our method, both automatically and with human evaluation, on in-domain Natural Questions and out-of-domain news datasets, providing insights into question generation for information seeking.
- We propose a novel evaluation metric with a pretrained QA system for generated SQs when there is no reference question.

2 Related Work

QG has been studied in multiple application contexts (e.g., generating questions for reading comprehension tests (Heilman and Smith, 2010), generating questions about an image (Mostafazadeh et al., 2016), recommending questions with respect to a news article (Laban et al., 2020)), evaluating summaries (Deutsch et al., 2020; Wang et al., 2020), and using multiple methods (see (Pan et al., 2019) for a recent survey). Early neural models focused on sequence-to-sequence generation based solutions (Serban et al., 2016; Du et al., 2017). The primary directions for improving these early works generally fall into the categories of providing mechanisms to inject answer-aware information into the neural encoder-decoder architectures (Du and Cardie, 2018; Li et al., 2019; Liu et al., 2019; Wang et al., 2020; Sun et al., 2018), encoding larger portions of the answer document as context (Zhao et al., 2018; Tuan et al., 2020), and incorporating richer knowledge sources (Elsahar et al., 2018).

These QG methods and the work described in this paper focus on using single-hop QA datasets such as SQuAD (Rajpurkar et al., 2016, 2018), NewsQA (Trischler et al., 2017; Hermann et al., 2015), and MS Marco (Bajaj et al., 2016). However, there has also been recent interest in multi-hop QG settings (Yu et al., 2020; Gupta et al., 2020; Malon and Bai, 2020) by using multi-hop QA datasets including HotPotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), and FreebaseQA (Jiang et al., 2019). Finally, there has been some recent interesting work regarding *unsupervised* QG, where the goal is to generate QA training data without an existing QG corpus to train better QA models (Lewis et al., 2019; Li et al., 2020).

Most directly related to our work from a motivation perspective is recent research regarding providing SQs in the context of supporting a news chatbot (Laban et al., 2020). However, the focus of this work is not QG, where they essentially use a GPT-2 language model (Radford et al., 2019) trained on SQuAD data for QG and do not evaluate this component independently. Qi et al. (2020) generates questions for information-seeking but not focuses on introductory questions. Most directly related to our work from a conceptual perspective is regarding producing questions for long answer targets (Mishra et al., 2020), which we contrast directly in Section 3. As QG is a generation task, automated evaluation frequently uses metrics such as BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and ROUGE (Lin, 2004). As these do not explicitly evaluate the requirements of our information-seeking use case, we also evaluate using the output of a trained QA system and conduct human annotator evaluations.

3 Problem Definition

Given a context X and an answer A, we want to generate a question $\tilde{Q}$ that satisfies

$$\tilde{Q} = \arg\max_{Q} P(Q|X, A),$$

where the context X could be a paragraph or a document that contains answers, rather than sentences as used in (Du and Cardie, 2018; Tuan et al., 2020), while A could be either short spans in X such as entities or noun phrases (referred to as a *short answer*), or the entire context X (referred to as a *long answer*).

The *long answer* QG task targets generating questions that are best answered by the entire context (i.e., paragraph or document) or a summary of the context, which is notably different from most QG settings where the answer is a short text span and the context is frequently a single sentence. Mishra et al. (2020) also work on the *long answer* QG setting using the NQ dataset, but their task definition is $\arg\max_Q P(Q|X)$ where they refer to the context X as the *long answer*. We use their models as baselines.

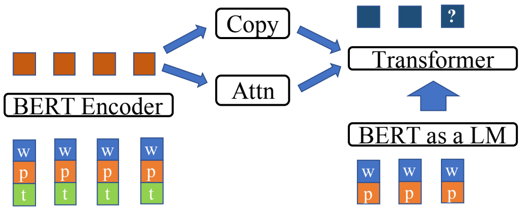

Figure 1: The BERTPGN architecture. The input for the BERT encoder is the context (w/p: word and position embeddinngs) with answer spans (or the whole context in the long answer setting) marked with the answer tagging (t: answer tagging embeddings). The decoder is a combination of BERT as a language model (i.e. has only self-attentions) and a Transformer-based pointer-generator network.

4 Methods

We use the BERT-based Pointer-Generator Network (BERTPGN) to generate questions. Tuan et al. (2020) use two-layer cross attentions between contexts and answers to encode contexts such as paragraphs when generating questions and show improved results. However, they show that three-layer cross attentions produce worse results. We will show later in the experiment that this is due to a lack of better initialization and that a higher layer is better for long answer question generation. Zhao et al. (2018) use answer tagging from the context instead of combining context and answer. Our model is motivated by these two works (Figure 1).

4.1 Context and Answer Encoding

Given context $X = \{x_i\}_{i=1}^{L}$, we add positional embeddings $P = \{p_i\}_{i=1}^{L}$ and type embeddings $T = \{t_i\}_{i=1}^{L}$ as the input for BERT. We use type embeddings to discriminate between a context and an answer, following Zhao et al. (2018); Tuan et al. (2020). We use $t_i = 0$ to represent '*context-only*' and $t_i = 1$ to represent '*both context and answer*' for token x_i. We do not apply the [CLS] in the beginning since we do not need the pooled output from BERT. We do not use the [SEP] to combine contexts and answers as inputs for BERT since we mark answers in the context with type embeddings.

The sequence output from BERT which forms our context-answer encoding is given by

$$H = f_{\text{BERT}}\left(X + P + T\right).$$

4.2 Question Decoding

The transformer-based Pointer-Generator Network is derived from (See et al., 2017) with adaptations to support transformers (Vaswani et al., 2017). Denoting $\mathrm{LN}(\cdot)$ as layer normalization, $\mathrm{MHA}(Q, K, V)$ as the multi-head attention with three parameters—query, key, and value, $\mathrm{FFN}(\cdot)$ as a linear function, and the decoder input at time t: $Y^{(t)} = \{y_j\}_{j=1}^{t}$, the decoder self-attention at time t is given by (illustrated with a single-layer transformer simplification)

$$A_S^{(t)} = \mathrm{LN}\left(\mathrm{MHA}\left(Y^{(t)}, Y^{(t)}, Y^{(t)}\right) + Y^{(t)}\right),$$

the cross-attention between encoder and decoder is

$$A_C^{(t)} = \mathrm{LN}\left(\mathrm{MHA}\left(A_S^{(t)}, H, H\right) + A_S^{(t)}\right),$$

and the final decoder output is

$$O^{(t)} = \mathrm{LN}\left(\mathrm{FFN}\left(A_C^{(t)}\right) + A_C^{(t)}\right).$$

Using the LSTM (Hochreiter and Schmidhuber, 1997) encoder-decoder model, See et al. (2017) compute a generation probability using the encoder context, decoder state, and decoder input. While the transformer decoder cross-attention $A_C^{(t)}$ already contains a linear combination between self-attention of decoder input and encoder-decoder cross attention. Thus, we use the combination of the decoder input and cross-attention to compute the generation probability

$$P_G^{(t)} = \mathrm{FFN}\left(\left[Y^{(t)}, A_C^{(t)}\right]\right).$$

To improve generalization, we also use a separate BERT model as a language model (LM) for the decoder. Even though BERT is not trained to predict the next token (Devlin et al., 2019) as with typical language models (e.g., GPT-2), we still choose BERT as our LM to ensure the COPY mechanism shares the same vocabulary between the encoder and the decoder.[1] We also do not need to process out-of-vocabulary words because we use the BPE (Sennrich et al., 2016; Devlin et al., 2019) tokenization in both the encoder and decoder.

5 Dataset

5.1 Natural Questions dataset

We use Natural Questions dataset (Kwiatkowski et al., 2019) for training as NQ questions are independent of their supporting documents. NQ has 307,000 training examples, answered and annotated from Wikipedia pages, in a format of a question, a Wikipedia link, long answer candidates, and short answer annotations. 51 % of these questions have no answer for either being invalid or non-evidence in their supporting documents. Another 36 % have long answers that are paragraphs and have corresponding short answers that either spans long answers or being masked as yes-or-no. The remaining 13 % questions only have long answers. We are most interested in the last portion of questions as they are best answered by summaries of their long answers, reflecting the coarse-grained information-seeking behavior.[2]

We use paragraphs that contain long answers or short answers in NQ as the context. We do not consider using the whole Wikipedia page, i.e., the document, as the context as most Wikipedia pages are too long to encode: In the NQ training set, there are 8407 tokens at document level on average, while for news articles in the CNN/Daily Mail that we will discuss in Section 5.2, the average document size is 583 (Tuan et al., 2020), which is not much larger than the average size of long answers in NQ (384 tokens).

We also consider the ratio between questions and the context-answer pairs to avoid generating multiple questions based on the same context-answer. After removing questions that have no answers, there are 152,148 questions and 136,450 unique long answers. The average ratio between questions and long answers is around 1.1 questions per *paragraph* (ratios are in a range of 1 to 47). The average ratio is more reasonable for question generation, comparing to the SQuAD where there are 1.4 questions per *sentence* on average (Du et al., 2017).

5.1.1 NQ Preprocessing

We extract questions, long answers, and short answer spans from the NQ dataset. We also extract the Wikipedia titles since long answers alone do not

[1] Note that we change the masking for the original BERT when using BERT as a LM, since the decoder at step t should not read inputs at steps $t + i$ where $i \geq 0$.

[2] Data annotation is a *subjective* task where different annotators could have different opinions for whether there is a short answer or not. NQ uses multi-fold annotations (e.g., a 5-fold annotation for the dev set). However, the training data only has the 1-fold annotation, so whether there is a short answer is not 100 % accurate.

data	type	count
train	mix	99,725
dev	mix	11,140
NQ-SA	long and short	3364
NQ-LA	long only	1495
News-LA	long only	3048

Table 1: QG Data summary. *-LA contains questions that only have long answers, while NQ-SA contains questions having both long and short answers.

always contain the words from their corresponding titles. We add brackets (‘[’ and ‘]’) for all possible short answer spans such that we can later extract these spans accordingly to avoid potential position changes due to context preprocessing (e.g., different tokenization).[3] When there is no short answer, we add brackets to the whole long answer. We then concatenate the titles with long answers as contexts. For details, see examples from Figure 5 and Figure 6 in Appendix A.

As in (Mishra et al., 2020), we only keep questions with long answers starting from the HTML paragraph tag. After preprocessing (Table 1), we get 110,865 question-context pairs, while Mishra et al. (2020) gets 77,501 pairs since they only keep long answer questions. We split the dataset with a 90/10 ratio for training/validation.

We use the original NQ dev set, which contains 7830 questions, as our test set. We follow the same extraction procedure as with the training and validation data modulo two new steps. First, noting that 79 % of Wikipedia pages appearing in the NQ dev set are also present in the NQ training set, we filter all overlapped contexts from the NQ dev set when creating our test set. Second, the original NQ dev set is 5-way annotated; thus, each question may have up to five different long/short answers. We treat each annotation as an independent context, even though they are associated with the same target question. To separately evaluate the QG performance for long answers and short answers, we split test data into *long-answer* questions (NQ-LA) and *short-answer* questions (NQ-SA). Finally, we get 4859 test data in total, with 1495 of them only have long answers while the remaining 3364 have both long and short answers while Mishra et al. (2020) gets 2136 test data from the original dev set.

5.2 News dataset

We use the 12,744 CNN news articles from the CNN/Daily Mail dataset (Hermann et al., 2015)) for the out-of-domain evaluation. We apply the same preprocessing method as in the NQ dataset to create a long-answer test set — News-LA. We use whole news articles, instead of paragraphs, as contexts, considering to generate questions that lead to entire news articles as answers. For each news article, we first remove *highlights*, which is a human-generated summary, and datelines (e.g., NEW DELHI, India (CNN)). We filter out those news articles that are longer than 490 tokens with the BEP tokenization and those overlapped context-question pairs. Finally, we get 3048 data in the News-LA test set.

6 In-Domain Evaluation with Generation Metrics

6.1 Experiment Setup and Training

We use a BERT-base uncased model (Devlin et al., 2019) that contains 12 hidden layers. The vocabulary contains 30,522 tokens. We create the PGN decoder with another BERT model from the same setting, followed by a 2-layer transformer with 12 heads and 3072 intermediate sizes. The maximum allowed context length is 500, while the maximum question length is 50. We train our model on an Amazon EC2 P3 machine with one Tesla V100 GPU, with the batch size 10, and the learning rate 5×10^{-5} with the Adam optimizer (Kingma and Ba, 2015) on all parameters of the BERTPGN model (both BERT models are trainable). We train 20 epochs of our model and evaluate with the dev set to select the model according to perplexity. Each epoch takes around 20 minutes to finish. Throughout the paper, we use the implementation of BLEU, METEOR, and ROUGE_L by Sharma et al. (2017).

6.2 In-Domain Evaluation

We first evaluate our model using BLEU, METEOR, and ROUGE_L to compare with Mishra et al. (2020) on long answers (first two rows in Table 2). The transformer-based iwslt_de_en is a German to English translation model with 6 encoder and decoder layers, 16 encoder and decoder attention heads, 1024 embedding dimension, and 4096 embedding dimension of feed forward network. The other transformer-based multi-source method, which is based on (Libovický et al., 2018), combines each context with a retrieval-based summary

[3] Using brackets here is an arbitrary but functional choice.

	B1	B4	ME	RL
TX iwslt_de_en	36.8	13.9	17.5	35.6
TX Multi-Source	36.0	13.3	16.8	34.6
BERTPGN LA	**43.9**	**20.1**	**22.6**	**42.2**
BERTPGN SA	**54.7**	**28.1**	**27.9**	**53.2**

Table 2: Comparing our model (BERTPGN) on NQ-LA and NQ-SA with two models in (Mishra et al., 2020)—their best performing Transformer_iwslt_de_en and multi-source transformer combining contexts and automatically generated summaries, with automatic evaluation BLEU-1, BLEU-4, METEOR, and Rouge_L.

B4	NQ-LA	NQ-SA
no-pointer	17.1	23.6
no-BERT-LM (*)	18.9	26.5
* - no-type-id	19.0	20.8
* - no-init	15.3	19.3
* - 2-layer	14.9	19.1

Table 3: Ablation study of the BERTPGN. Removing the pointer network drops BLEU-4 by around 3 points for both test sets. Removing BERT initialization affects both the NQ-LA and NQ-SA substantially but more mildly than removing the pointer. Removing type IDs affects the NQ-SA by 5.7 drop in BLEU-4.

as input. We decode questions from our model using beam search (beam=3).[4] Evaluating on NQ-LA, our BERTPGN model outperforms both existing models substantially with near seven points for all metrics. The performance for short answer questions NQ-SA is even better, with near eight more BLEU-4 points than NQ-LA.

6.3 Ablation Study

We first examine the effect of the pointer network from the BERTPGN. We then run ablation study by first removing BERT-LM in the decoder, and independently

- removing type IDs from BERT encoder
- removing BERT initialization for BERT encoder
- substituting BERT encoder with a 2-layer transformer

We train our BERTPGN models from scratch for each setting and conduct these ablation studies for NQ-LA and NQ-SA separately (Table 3).

Removing the pointer from the BERTPGN makes the BLEU-4 scores drop for both NQ-LA and NQ-SA more than removing the BERT as the LM in the decoder. Type IDs are more helpful for NQ-SA (approximately a 5-point drop in BLEU-4) than NQ-LA since NQ-SA needs to use type IDs to mark answers. Removing BERT initialization causes notable drops for both NQ-LA (3.6 drops in BLEU-4) and NQ-SA (7.2 in BLEU-4), which implies that BERT achieves better generalization when encoding these considerably long contexts. Another interesting finding is that the NQ-LA is more sensitive to the number of layers of the encoder than NQ-SA. When decreasing the layers to two from twelve, NQ-LA drops by 0.4 in BLEU-4 while NQ-SA drops by 0.2.

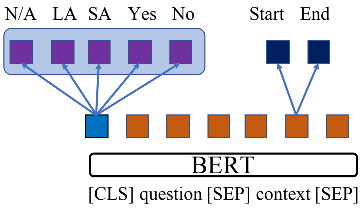

Figure 2: The BERT-joint architecture (Alberti et al., 2019b). Input is the combined question and context, and the outputs are an answer-type classification from the [CLS] token and start/end of answer spans for each token from the context.

7 Out-of-Domain Evaluation with QA Systems

We use a well-trained question answering system as the evaluation method, given that the automated scoring metrics have two notable drawbacks when evaluating long-answer questions: (1) There are usually multiple valid questions for long-answer question generation as contexts are much longer than previous work. However, most datasets only have one gold question for each context; (2) They cannot measure generated questions when there is no gold question, which is the right problem that we encountered for our News-LA dataset.

7.1 The QA Metrics

We use the BERT-joint model (Alberti et al., 2019b) (Figure 2) for NQ question answering to evaluate our long answer question generation. The BERT-joint model takes the combination a question and the corresponding context as an input, outputs the probability of answer spans and the probability of answer types. For a context of size n, it produces p_{start} and p_{end} for each token, indicating whether this token is a start or end token of an answer span. It then chooses the answer span (i, j) where $i < j$

[4] Mishra et al. (2020) have not described the decoding method and possible beam size, but they use models from (Ott et al., 2018) that uses beam=4.

	B1	B4	ME	RL
Du-17 best	43.1	12.3	16.6	39.8
M_{SD}	**46.0**	**14.8**	**19.2**	**42.0**

Table 4: The performance of our answer-free baseline, compared with the best model from (Du et al., 2017).

that maximizes $p_{start}(i) \cdot p_{end}(j)$ as the probability of the answer. It also defines the probability of no answer to be $p_{start}([CLS]) \cdot p_{end}([CLS])$, i.e., an answer span that starts then stops at the `[CLS]` token. Furthermore, the BERT-joint model computes the probability of *types* of the question—*undetermined*, *long answer*, *short answer*, and *YES-or-NO*. This model achieves 66.2 % F1 on NQ long answer test set, which is 10 % better compared to models used in (Kwiatkowski et al., 2019; Parikh et al., 2016). We define the *answerability* score (s_{ans}) as $\log{(p_{ans}/p_{no\_ans})}$, and the *granularity* score (s_{gra}) as $\log{(p_{la}/p_{sa})}$ when evaluating our long answer question generation with the BERT-joint model.

7.2 QG Models to Compare

We construct a baseline model to compare as follows. Using the same BERTPGN architecture, we train a model on the SQuAD sentence-question pairs prepared by Du et al. (2017). When generating questions for news articles, we use the first line of each news article as the context, with the assumption that the first line is a genuine summary produced by humans. Notice that the resulting baseline is the state-of-the-art for answer-free (the model does not know the whereabouts of answer spans) question generation with SQuAD (Table 4). We refer to the model as M_{SD} hereafter. Similarly, we call our BERTPGN model trained on the NQ dataset as M_{NQ}. We use beam search (beam=3) for both models.

7.3 Evaluation Results

We show the QA evaluation results in Figure 3. In the context column, M_{NQ} shows a lower answerability score than the baseline model M_{SD}. While granularity scores show a reverse trend, i.e., higher scores for M_{NQ} than those of M_{SD}. This result implies that M_{NQ} generates more coarse-style questions that have long answers, but these questions are considerably more difficult to answer by the QA model, comparing to short-answer questions.

It is also reasonable to assume that news articles' summaries are proper answer-candidates for long-answer questions. Highlights in news articles are human-generated summaries, so we also combine the same set of questions with their corresponding highlights as input for the BERT-joint QA system with results shown as the highlights column in Figure 3. The answerability scores drop for both models comparing the column highlights to the column of context, which is reasonable as the models never see highlights when generating questions. However, the baseline method M_{SD} drops more significantly than M_{NQ}, suggesting that the baseline model is more context-dependent while our model M_{NQ} generates more self-explanatory questions. From the granularity scores of highlights, we find that confidence to determine answer types is lower for both models than that of the context column. However, the M_{NQ} still shows higher granularity scores than the M_{SD}.

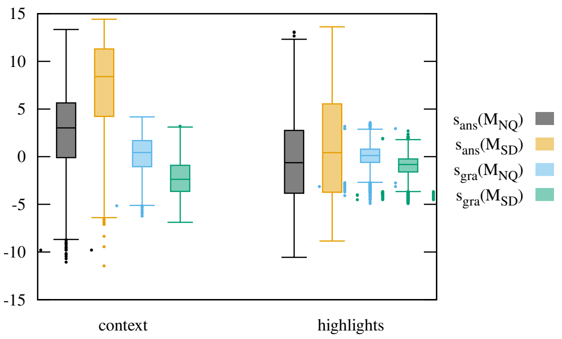

Figure 3: Answerability and granularity scores of generated questions for News-LA with the BERT-joint model (Alberti et al., 2019b) as the evaluation QA model by answering generated questions from either news article *context* or news article *highlights*. We compare two models: (1) NQ: BERTPGN trained with NQ dataset and generate on whole news articles; (2) SD: BERTPGN trained with SQuAD dataset and generate on the first line of each news article.

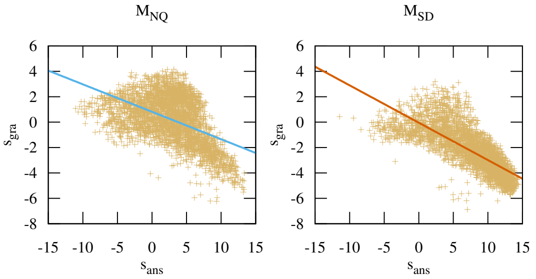

Figure 4: Scatter plots of generated questions of the News-LA from M_{NQ} (left) and M_{SD} (right). s_{ans} and s_{gra} are negatively correlated, but the M_{NQ} model tends to generate more questions with positive anserability and granularity. Straight lines show fitted linear regressions.

We map generated questions for the News-LA on a 2D plot with x-axis the answerability score and y-axis the granularity score for both models in Figure 4. They also confirm the negative correlation between answerability and granularity of generated questions. However, the M_{NQ} generates more questions with both positive s_{ans} and s_{gra} than those from M_{SD}, indicating the effectiveness of our model to generate introductory and self-explanatory questions.

8 Out-of-Domain Human Evaluation

(%)	Context		Span		Entire	
	T	F	T	F	T	F
M_{NQ}	38	62	77	23	49	51
M_{SD}	70	30	89	11	40	60

Table 5: Ratios (shown as a percentage) between *True* and *False* for human evaluation with three statements (*Context*, *Span*, and *Entire*) on generated questions. We count true/false marked by annotators with unanimity amongst all three annotators for each statement.

We further conduct a human evaluation using MTurk for the News-LA test set to verify that we can generate self-explanatory and introductory questions and that the automatic evaluation in Section 7 agrees with human evaluation. We ask annotators to read news articles and mark true or false for seven statements regarding generated questions. For each context-question pair, these statements include (see examples in Appendix B)

- Question is *context* dependent
- Question is *irrelevant* to the article
- Question implies a *contradiction* to facts present in the article
- Question focuses on a *peripheral* topic
- There is a short *span* to answer the question
- The *entire* article can be an answer
- *None* answer in the article

We randomly select 1000 news articles in News-LA to perform our human evaluation with three different annotators per news article. We received three valid annotations for 943 news articles from a set of 224 annotators. We first consider true/false results regarding three metrics – *Context*, *Span*, and *Entire* – considering only when unanimity is reached among annotators (Table 5). M_{NQ} questions are more context-free than M_{SD} ones, with 38 % true and 62 % false towards the *Context* statement. Second, the M_{NQ} questions are more likely to be answered by entire news articles (49 % true of *Entire* vs. 40 %) while less likely to be answered by spans from news articles (77 % true of *Span* vs. 89 %) comparing with M_{SD} questions. These human evaluation results confirm that M_{NQ} questions are more self-explanatory and introductory than M_{SD}.

	s_{ans}		s_{gra}	
	M_{NQ}	M_{SD}	M_{NQ}	M_{SD}
Context	0.1	–0.1	0.1	0.5
Irrelevant	–1.0	–0.6	**0.7**	0.4
Contradiction	–0.5	–0.3	0.4	0.2
Peripheral	–0.3	–0.3	0.2	0.2
Span	**1.5**	**1.1**	**–0.8**	**–0.6**
Entire	0.4	0.3	0.4	0.3
None	**–1.5**	**–1.2**	0.6	**0.6**

Table 6: Pearson correlation (1×10^{-1}) between human (Section 8) and automatic (Section 7) evaluation. For each column, we mark the most positive and negative correlated scores in bold text.

We compute the s_{ans} and s_{gra} for the 943 generated questions (Section 7). We then normalize these two scores and conduct a Pearson correlation analysis (Benesty et al., 2009) with human evaluation results. We use all human evaluation results, regardless of agreements among annotators. From Table 6, we find that *Span* has the strongest positive correlation with the s_{ans}, while *None* shows the strongest negative correlation – aligning with the findings for answerability. *Span* also shows the strongest negative correlation with the s_{gra} for both M_{NQ} and M_{SD}, but the highest positive correlation with granularity varies, with *Irrelevant* for M_{NQ} questions and *None* for M_{SD} questions.

9 Conclusion

We tackle the problem of question generation targeted for human information seeking using automatic question answering technology. We focus on generating questions for news articles that can be answered by longer passages rather than short text spans as suggested questions. We build a BERT-based Pointer-Generator Network as the QG model, trained with the Natural Questions dataset. Our method shows state-of-the-art performance in terms of BLEU, METEOR, and ROUGE_L scores on our NQ question generation dataset. We then apply our model to the out-of-domain news articles without further training. We use a QA system to evaluate our QG models as there are no gold questions for comparison. We also conduct a human evaluation to confirm the QA evaluation results.

Broader Impact

We describe a method for an autonomous agent to suggest questions based on machine-reading and question generation technology. Operationally, this work focuses on newswire-sourced data where the generated questions are answered by the text – and is applicable to multi-turn search settings. Thus, there are several potentially positive social impacts. By presenting questions with known answers in the text, users can more efficiently learn about topics in the source documents. Our focus on *self-explanatory* and *introductory* questions increases the utility of questions for this purpose.

Conversely, there is potential to bias people toward a subset of the news chosen by a purported fair search engine, which may be more difficult to detect as the provided questions remove some of the article contexts. In principle, this is mitigated by selecting content that maintains high journalistic standards – but such a risk remains if the technology is deployed by bad-faith actors.

The data for our experiments was derived from the widely used Natural Questions (Kwiatkowski et al., 2019) and CNN/Daily Mail (Hermann et al., 2015) datasets, which in turn were derived from public news sourced data. Our evaluation annotations were performed on Amazon Mechanical Turk, where three authors completed a sample task and set a wage corresponding to an expected rate of 15 $/h.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019a. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Chris Alberti, Kenton Lee, and Michael Collins. 2019b. A bert baseline for the natural questions.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 37–40. Springer.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2020. Towards question-answering as an automatic metric for evaluating the content quality of a summary.

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.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Hady Elsahar, Christophe Gravier, and Frederique Laforest. 2018. Zero-shot question generation from knowledge graphs for unseen predicates and entity types. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 218–228, New Orleans, Louisiana. Association for Computational Linguistics.

Deepak Gupta, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Reinforced multi-task approach for multi-hop question generation.

Michael Heilman and Noah A Smith. 2010. 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*, pages 609–617.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701. MIT Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase. 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 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. Gooaq: Open question answering with diverse answer types.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Philippe Laban, John Canny, and Marti A. Hearst. 2020. What's the latest? a question-driven news chatbot. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 380–387, Online. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. Association for Computational Linguistics.

Jingjing Li, Yifan Gao, Lidong Bing, Irwin King, and Michael R. Lyu. 2019. Improving question generation with to the point context. 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 3216–3226, Hong Kong, China. Association for Computational Linguistics.

Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu. 2020. Harvesting and refining question-answer pairs for unsupervised qa. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. Association for Computational Linguistics.

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, Mingjun Zhao, Di Niu, Kunfeng Lai, Yancheng He, Haojie Wei, and Yu Xu. 2019. Learning to generate questions by learningwhat not to generate. In *The World Wide Web Conference*, WWW '19, page 1106–1118, New York, NY, USA. Association for Computing Machinery.

Christopher Malon and Bing Bai. 2020. Generating followup questions for interpretable multi-hop question answering.

Shlok Kumar Mishra, Pranav Goel, Abhishek Sharma, Abhyuday Jagannatha, David Jacobs, and Hal Daumé III. 2020. Towards automatic generation of questions from long answers.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *CoRR*, abs/1905.08949.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, 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.

Peng Qi, Yuhao Zhang, and Christopher D. Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, Berlin, Germany. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3930–3939, Brussels, Belgium. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Luu Anh Tuan, Darsh Shah, and Regina Barzilay. 2020. Capturing greater context for question generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9065–9072.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

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.

Jianxing Yu, Xiaojun Quan, Qinliang Su, and Jian Yin. 2020. Generating multi-hop reasoning questions to improve machine reading comprehension. In *Proceedings of The Web Conference 2020*, WWW '20, page 281–291, New York, NY, USA. Association for Computing Machinery.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. 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*, pages 3901–3910, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

We show several generated questions here. Each frame box contains a news article, with two questions generated by M_{NQ} (showing in bold texts) and M_{SD} respectively. News articles are selected from the CNN/Daily Mail dataset with preprocessing described in Section 5.2. We also compare these generated questions in Table 7.

Two Italians, a Dane, a German, a Frenchman and a Brit walk into a space station... or will, in 2013, if all goes according to European Space Agency plans. Europe's six new astronauts hope to join their American counterparts on the Internation Space Station. The six new astronauts named Wednesday were chosen from more than 8,400 candidates, and are the first new ESA astronauts since 1992, the space agency said in a statement. They include two military test pilots, one fighter pilot and one commercial pilot, plus an engineer and a physicist. ”This is a very important day for human spaceflight in Europe,” said Simonetta Di Pippo, Director of Human Spaceflight at ESA. ”These young men and women are the next generation of European space explorers. They have a fantastic career ahead, which will put them right on top of one of the ultimate challenges of our time: going back to the Moon and beyond as part of the global exploration effort.” Humans have not walked on the moon since 1972, just over three years after the first manned mission to Earth's nearest neighbor. The six will begin space training in Germany, with an eye to being ready for future missions to the International Space Station and beyond in four years. They are: Samantha Cristoforetti of Italy, a fighter pilot with degrees in engineering and aeronautical sciences; Alexander Gerst, a German researcher with degrees in physics and earth science; Andreas Mogensen, a Danish engineer with the private space firm HE Space Operations; Luca Parmitano of Italy, an Air Force pilot with a degree in aeronautical sciences; Timothy Peake, an English test pilot with the British military; and Frenchman Thomas Pesquet, an Air France pilot who previously worked as an engineer at the French space agency.

- **who are the new astronauts on the moon**
- how many italians walk into a space station in 2013

After several delays, NASA said Friday that space shuttle Discovery is scheduled for launch in five days. The space shuttle Discovery, seen here in January, is now scheduled to launch Wednesday. Commander Lee Archambault and his six crewmates are now scheduled to lift off to the International Space Station at 9:20 p.m. ET Wednesday. NASA said its managers had completed a readiness review for Discovery, which will be making the 28th shuttle mission to the ISS. The launch date had been delayed to allow ”additional analysis and particle impact testing associated with a flow-control valve in the shuttle's main engines,” the agency said. According to NASA, the readiness review was initiated after damage was found in a valve on the shuttle Endeavour during its November 2008 flight. Three valves have been cleared and installed on Discovery, it said. Discovery is to deliver the fourth and final set of ”solar array wings” to the ISS. With the completed array the station will be able to provide enough electricity when the crew size is doubled to six in May, NASA said. The Discovery also will carry a replacement for a failed unit in a system that converts urine to drinkable water, it said. Discovery's 14-day mission will include four spacewalks, NASA said.

- **when is the space shuttle discovery coming out**
- how many days is the space shuttle discovery scheduled to launch

Unemployment in Spain has reached 20 percent, meaning 4.6 million people are out of work, the Spanish government announced Friday. The figure, from the first quarter, is up from 19 percent and 4.3 million people in the previous quarter. It represents the second-highest unemployment rate in the European Union, after Latvia, according to figures Friday from Eurostat, the EU's statistics service. Spanish Prime Minister Jose Luis Rodriguez Zapatero told Parliament on Wednesday he believes the jobless rate has peaked and will now start to decline. The first quarter of the year is traditionally poor for Spain because of a drop in labor-intensive activity like construction, agriculture and tourism. This week, Standard & Poor's downgraded Spain's long-term credit rating and said the outlook is negative. ”We now believe that the Spanish economy's shift away from credit-fuelled economic growth is likely to result in a more protracted period of sluggish activity than we previously assumed,” Standard & Poor's credit analyst Marko Mrsnik said. Gross domestic product growth in Spain is expected to average 0.7 percent annually through 2016, compared with previous expectations of 1 percent annually, he said. Spain's economic problems are closely tied to the housing bust there, according to The Economist magazine. Many of the newly unemployed worked in construction, it said. The recession revealed how dependent public finances were on housing-related tax revenues, it said. Another problem in Spain is that wages are set centrally and most jobs are protected, making it hard to shift skilled workers from one industry to another, the magazine said. Average unemployment for the 27-member European Union stayed stable in March at 9.6 percent, Eurostat said Friday. That percentage represents 23 million people, it said. The lowest national unemployment rates were in the Netherlands and Austria, which had 4.1 and 4.9 percent respectively, Eurostat said.

- **what is the average unemployment rate in spain**
- what percentage of spain's population is out of work

Atlanta rapper DeAndre Cortez Way, better known by his stage name Soulja Boy Tell 'Em or just Soulja Boy, was charged with obstruction after running from police despite an order to stop, a police spokesman said Friday. Rapper Soulja Boy was arrested in Georgia after allegedly running from police. The 19-year-old singer was among a large group that had gathered at a home in Stockbridge, 20 miles south of Atlanta, said Henry County, Georgia, police Capt. Jason Bolton. Way was arrested Wednesday night along with another man, Bolton said. Police said Way left jail Thursday after posting a $550 bond. Bolton said officers responded to a complaint about a group of youths milling around the house, which appeared to be abandoned. When police arrived, they saw about 40 people. Half of them

President of the United Nations General Assembly [Miroslav Lajčák of Slovakia] has been elected as the United Nations General Assembly President of its 72nd session beginning in September 2017.
who is the current president of un general assembly

Learner 's permit Typically , a driver operating with a learner 's permit must be accompanied by [an adult licensed driver who is at least 21 years of age or older and in the passenger seat of the vehicle at all times] .
who needs to be in the car with a permit driver

Java development Kit [The Java Development Kit (JDK) is an implementation of either one of the Java Platform , Standard Edition , Java Platform , Enterprise Edition , or Java Platform , Micro Edition platforms released by Oracle Corporation in the form of a binary product aimed at Java developers on Solaris , Linux , macOS or Windows . The JDK includes a private JVM and a few other resources to finish the development of a Java Application . Since the introduction of the Java platform , it has been by far the most widely used Software Development Kit (SDK) . On 17 November 2006 , Sun announced that they would release it under the GNU General Public License (GPL) , thus making it free software . This happened in large part on 8 May 2007 , when Sun contributed the source code to the OpenJDK .]
what is the use of jdk in java

Figure 5: Examples of the NQ data preprocessing from the training set. Orange texts are Wikipedia titles that added in the front the each long answers. In first two examples, annotators mark there are short answers represented in cyan; while for the last example, there is no short answer marked by annotators so we mark the whole paragraph as the answer. Cyan texts are tagged with type ID '1' during preprocessing.

Context Question	Therefore sign [**(1)** In logical argument and mathematical proof, [**(2)** [**(3)** the [**(4)** therefore sign **(/4)**] **(/3)**] (∴) is generally used before [**(5)** a logical consequence, such as the conclusion of a syllogism. **(/5)**] **(/2)**] The symbol consists of three dots placed in an upright triangle and is read therefore. It is encoded at U+2234 ∴ therefore (HTML &#8756; &there4;). For common use in Microsoft Office hold the ALT key and type "8756". While it is not generally used in formal writing, it is used in mathematics and shorthand. It is complementary to U+2235 ∵ because (HTML &#8757;). **(/1)**] *what do the 3 dots mean in math*
SA 1 Predicted	whole paragraph *what is the therefore sign in a syllogism*
SA 2 Predicted	[the therefore sign (∴) is generally used before a logical consequence, such as the conclusion of a syllogism.] *what is the meaning of therefore in triangle*
SA 3 Predicted	[the therefore sign] *what is the name of the three dots in a triangle called*
SA 4 Predicted	[therefore sign] *what is the name of the three dots in a triangle called*
SA 5 Predicted	[a logical consequence , such as the conclusion of a syllogism] *when is the therefore sign used in a syllogism*

Figure 6: Example of the question generation from Natural Questions dataset with BERTPGN. We use '[(i)' and '(/i)]' to represent the start and end position of the *i*-th answer span. The context is the long answer for the question *what do the 3 dots mean in math*. Five short answers (SA) marked by five different annotators. Our BERTPGN model with nucleus sampling (Holtzman et al., 2019) with temperature of 0.1 produces different but related questions for each short answers as well as the whole context with brackets over each of them.

ran away, including Way, Bolton said. The ones who remained told officers they were at the home to film a video. Way was arrested when he returned to the house to get his car, Bolton said. He said the house was dark inside and looked abandoned. "He just ran from the police, and then he decided to come back," according to Bolton. The second man who returned for his vehicle was arrested after police found eight $100 counterfeit bills inside, according to the officer. Way broke into the music scene two years ago with his hit "Crank That (Soulja Boy)." The rapper also describes himself as a producer and entrepreneur.

- **what is the meaning of soulja boy tell em**
- what was deandre cortez way known as

The U.S. military is gearing up for a possible influx of Haitians fleeing their earthquake-stricken country at an Army facility not widely known for its humanitarian missions: Guantanamo Bay. Soldiers at the base have set up tents, beds and toilets, awaiting possible orders from the secretary of defense to proceed, according to Maj. Diana Haynie, a spokeswoman for Joint Task Force Guantanamo Bay. "There's no indication of any mass migration from Haiti," Haynie stressed. "We have not been told to conduct migrant operations." But the base is getting ready "as a prudent measure," Haynie said, since "it takes some time to set things up." Guantanamo Bay is about 200 miles from Haiti. Currently, military personnel at the base are helping the earthquake relief effort by shipping bottled water and food from its warehouse. In addition, Gen. Douglas Fraser, commander of U.S. Southern Command, said the Navy has set up a "logistics field," an area to support bigger ships in the region. The military can now use that as a "lily pad" to fly supplies from ships docked at Guantanamo over to Haiti, he said. "Guantanamo Bay proves its value as a strategic hub for the movement of supplies and personnel to the affected areas in Haiti," Haynie said. As part of the precautionary measures to prepare for possible refugees, the Army has

BERTPGN-NQ-whole-article	BERTPGN-SQuAD-first-line
who are the new astronauts on the moon	how many italians walk into a space station in 2013
when is the space shuttle discovery coming out	how many days is the space shuttle discovery scheduled to launch
what is the average unemployment rate in spain	what percentage of spain's population is out of work
what is the meaning of soulja boy tell em	what was deandre cortez way known as
where does the us refugees at guantanamo bay come from	what is the name of the us military facility in the us
what happened to the girl in the texas polygamist ranch	what was the name of the texas polygamist ranch
who scored the first goal in the premier league	which team did everton fc beat to win the premier league's home draw with tottenham on sunday

Table 7: Comparing generated questions side-by-side. Our model uses uncased vocabulary and omits the final question mark.

erected 100 tents, each holding 10 beds, according to Haynie. Toilet facilities are nearby. If needed, hundreds more tents are stored in Guantanamo Bay and can be erected, she said. The refugees would be put on the leeward side of the island, more than 2 miles from some 200 detainees being held on the other side, Haynie said. The refugees would not mix with the detainees. Joint Task Force Guantanamo Bay is responsible for planning for any kind of Caribbean mass immigration, according to Haynie. In the early 1990s, thousands of Haitian refugees took shelter on the island, she said.

- **where does the us refugees at guantanamo bay come from**
- what is the name of the us military facility in the us

A Colorado woman is being pursued as a ”person of interest” in connection with phone calls that triggered the raid of a Texas polygamist ranch, authorities said Friday. Rozita Swinton, 33, has been arrested in a case that is not directly related to the Texas raid. Texas Rangers are seeking Rozita Swinton of Colorado Springs, Colorado, ”regarding telephone calls placed to a crisis center hot line in San Angelo, Texas, in late March 2008,” the Rangers said in a written statement. The raid of the YFZ (Yearning for Zion) Ranch in Eldorado, Texas, came after a caller – who identified herself as a 16-year-old girl – said she had been physically and sexually abused by an adult man with whom she was forced into a ”spiritual marriage.” The release said a search of Swinton's home in Colorado uncovered evidence that possibly links her to phone calls made about the ranch, run by the Fundamentalist Church of Jesus Christ of Latter-day Saints. ”The possibility exists that Rozita Swinton, who has nothing to do with the FLDS church, may have been a woman who made calls and pretended she was the 16-year-old girl named Sarah,” CNN's Gary Tuchman reported. Swinton, 33, has been charged in Colorado with false reporting to authorities and is in police custody. Police said that arrest was not directly related to the Texas case. Authorities raided the Texas ranch April 4 and removed 416 children. Officials have been trying to identify the 16-year-old girl, referred to as Sarah, who claimed she had been abused in the phone calls. FLDS members have denied the girl, supposedly named Sarah Jessop Barlow, exists. Some of the FLDS women who spoke with CNN on Monday said they believed the calls were a hoax. While the phone calls initially prompted the raid, officers received a second search warrant based on what they said was evidence of sexual abuse found at the compound. In court documents, investigators described seeing teen girls who appeared pregnant, records that showed men marrying multiple women and accounts of girls being married to adult men when they were as young as 13. A court hearing began Thursday to determine custody of children who were removed from the ranch.

- **what happened to the girl in the texas polygamist ranch**
- what was the name of the texas polygamist ranch

Everton scored twice late on and goalkeeper Tim Howard saved an injury-time penalty as they fought back to secure a 2-2 Premier League home draw with Tottenham on Sunday. Jermain Defoe gave the visitors the lead soon after the interval when nipping in front of Tony Hibbert to convert Aaron Lennon's cross at the near post for his 13th goal of the season. And they doubled their advantage soon after when defender Michael Dawson headed home a Niko Kranjcar corner. But Everton got a foothold back in the game when Seamus Coleman's run and cross was converted by fellow-substitute Louis Saha in the 78th minute. And Tim Cahill rescued a point for the home side with four minutes remaining when he stooped low to head home Leighton Baines' bouncing cross. However, there was still further drama to come when Hibbert was penalized for crashing into Wilson Palacios in the area. However, England striker Defoe smashed his penalty too close to Howard and the keeper pulled off a fine save to give out-of-form Everton a morale-boosting point. The result means Tottenham remain in fourth place, behind north London rivals Arsenal, while Everton have now won just one of their last nine league games. In the day's other match, Bobby Zamora scored the only goal of the game as Fulham beat Sunderland 1-0 to move up to eighth place in the table.

- **who scored the first goal in the premier league**
- which team did everton fc beat to win the premier league's home draw with tottenham on sunday

B Human Evaluation Criteria

Question is context dependent

Some questions are context-dependent, e.g.,

- “who intends to boycott the election” - which election?

- “where did the hijackers go to” - what hijackers?
- “what type of hats did they use” - who are they?
- “how many people were killed in the quake” - which quake?

Compared to these context-independent, self-contained questions:

- “what was toyota’s first-ever net loss”
- “who is hillary’s secretary of state”
- “what is the name of the motto of the new york times ”

Question is irrelevant to the article

Given a news article:

> ”Usually when I mention suspended animation people will flash me the Vulcan sign and laugh,” says scientist Mark Roth. But he’s not referring to the plot of a ”Star Trek” episode. Roth is completely serious about using lessons he’s learned from putting some organisms into suspended animation to help people survive medical trauma. He spoke at the TED2010 conference in Long Beach, California, in February. The winner of a MacArthur genius fellowship in 2007, Roth described the thought process that led him and fellow researchers to explore ways to lower animals’ metabolism to the point where they showed no signs of life – and yet were not dead. More remarkably, they were able to restore the animals to normal life, with no apparent damage. Read more about Roth on TED.com The Web site of Roth’s laboratory at the Fred Hutchinson Cancer Research Center in Seattle, Washington, describes the research this way: ”We use the term suspended animation to refer to a state where all observable life processes (using high resolution light microscopy) are stopped: The animals do not move nor breathe and the heart does not beat. We have found that we are able to put a number of animals (yeast, nematodes, drosophila, frogs and zebrafish) into a state of suspended animation for up to 24 hours through one basic technique: reducing the concentration of oxygen.” Visit Mark Roth’s laboratory Roth is investigating the use of small amounts of hydrogen sulfide, a gas that is toxic in larger quantities, to lower metabolism. In his talk, he imagined that ”in the not too distant future, an EMT might give an injection of hydrogen sulfide, or some related compound, to a person suffering severe injuries, and that person might de-animate a bit ... their metabolism will fall as though you were dimming a switch on a lamp at home. ”That will buy them the time to be transported to the hospital to get the care they need. And then, after they get that care ... they’ll wake up. A miracle? We hope not, or maybe we just hope to make miracles a little more common.”

The question: “what is the meaning of suspended animation in star trek” is irrelevant to the news since the news is not talking about Star Trek.

However, the question “what is the meaning of suspended animation” is related.

Question implies a contradiction to facts present in the article

Given a news article:

> At least 6,000 Christians have fled the northern Iraqi city of Mosul in the past week because of killings and death threats, Iraq’s Ministry of Immigration and Displaced Persons said Thursday. A Christian family that fled Mosul found refuge in the Al-Sayida monastery about 30 miles north of the city. The number represents 1,424 families, at least 70 more families than were reported to be displaced on Wednesday. The ministry said it had set up an operation room to follow up sending urgent aid to the displaced Christian families as a result of attacks by what it called ”terrorist groups.” Iraqi officials have said the families were frightened by a series of killings and threats by Muslim extremists ordering them to convert to Islam or face death. Fourteen Christians have been slain in the past two weeks in the city, which is about 260 miles (420 kilometers) north of Baghdad. Mosul is one of the last Iraqi cities where al Qaeda in Iraq has a significant presence and routinely carries out attacks. The U.S. military said it killed the Sunni militant group’s No. 2 leader, Abu Qaswarah, in a raid in the northern city earlier this month. In response to the recent attacks on Christians, authorities have ordered more checkpoints in several of the city’s Christian neighborhoods. The attacks may have been prompted by Christian demonstrations ahead of provincial elections, which are to be held by January 31, authorities said. Hundreds of Christians took to the streets in Mosul and surrounding villages and towns, demanding adequate representation on provincial councils, whose members will be chosen in the local elections. Thursday, Iraq’s minister of immigration and displaced persons discussed building housing complexes for Christian families in northern Iraq and allocating land to build the complexes. Abdel Samad Rahman Sultan brought up the issue when he met with a representative of Iraq’s Hammurabi Organization for Human Rights and with the head of the Kojina Organization for helping displaced persons. A curfew was declared Wednesday in several neighborhoods of eastern Mosul as authorities searched for militants behind the attacks.

The question “how many christians fled to mosul in the past” is contradicted to the fact — 6000 christians fled from Mosul — in the news.

Question focuses on a peripheral topic

Given a news article:

> One of the Marines shown in a famous World War II photograph raising the U.S. flag on Iwo Jima was posthumously awarded a certificate of U.S. citizenship on Tuesday. The Marine Corps War Memorial in Virginia depicts Strank and five others raising a flag on Iwo Jima. Sgt. Michael Strank, who was born in Czechoslovakia and came to the United States when he was 3, derived U.S. citizenship when his father was naturalized in 1935. However, U.S. Citizenship and Immigration Services recently discovered that Strank never was given citizenship papers. At a ceremony Tuesday at the Marine Corps Memorial – which depicts the flag-raising – in Arlington, Virginia, a certificate of citizenship was presented to Strank’s younger sister, Mary Pero. Strank and five other men became national icons when an Associated Press photographer captured the image of them planting an American flag on top of Mount Suribachi on February 23, 1945. Strank was killed in action on the island on March 1, 1945, less than a month before the battle between Japanese and U.S. forces there ended.

> Jonathan Scharfen, the acting director of CIS, presented the citizenship certificate Tuesday. He hailed Strank as ”a true American hero and a wonderful example of the remarkable contribution and sacrifices that immigrants have made to our great republic throughout its history.”

The question “who presented the american flag raising on iwo jima” focuses on a peripheral topic — the name of the one raising the flag.

While the question “who was awarded a certificate of citizenship raising the u.s. flag” focuses on the main topic - getting a citizenship.

There is a short span to answer the question

Given a news:

> Los Angeles police have launched an internal investigation to determine who leaked a picture that appears to show a bruised and battered Rihanna. Rihanna was allegedly attacked by her boyfriend, singer Chris Brown, before the Grammys on February 8. The close-up photo – showing a woman with contusions on her forehead and below her eyes, and cuts on her lip – was published on the entertainment Web site TMZ Thursday. TMZ said it was a photo of Rihanna. Twenty-one-year-old Rihanna was allegedly attacked by her boyfriend, singer Chris Brown, on a Los Angeles street before the two were to perform at the Grammys on February 8. ”The unauthorized release of a domestic violence photograph immediately generated an internal investigation,” an L.A. police spokesman said in a statement. ”The Los Angeles Police Department takes seriously its duty to maintain the confidentiality of victims of domestic violence. A violation of this type is considered serious misconduct, with penalties up to and including termination.” A spokeswoman for Rihanna declined to comment. The chief investigator in the case had told CNN earlier that authorities had tried to guard against leaks. Detective Deshon Andrews said he had kept the case file closely guarded and that no copies had been made of the original photos and documents. Brown was arrested on February 8 in connection with the case and and booked on suspicion of making criminal threats. Authorities are trying to determine whether Brown should face domestic violence-related charges. Brown apologized for the incident this week. ”Words cannot begin to express how sorry and saddened I am over what transpired,” the 19-year-old said in a statement released by his spokesman. ”I am seeking the counseling of my pastor, my mother and other loved ones and I am committed, with God’s help, to emerging a better person.”

The question “who have launched an internal investigation of the leaked rihanna’s picture” can be answered by “Los Angeles police”.

The entire article can be an answer

Given a news:

> A high court in northern India on Friday acquitted a wealthy businessman facing the death sentence for the killing of a teen in a case dubbed ”the house of horrors.” Moninder Singh Pandher was sentenced to death by a lower court in February. The teen was one of 19 victims – children and young women – in one of the most gruesome serial killings in India in recent years. The Allahabad high court has acquitted Moninder Singh Pandher, his lawyer Sikandar B. Kochar told CNN. Pandher and his domestic employee Surinder Koli were sentenced to death in February by a lower court for the rape and murder of the 14-year-old. The high court upheld Koli’s death sentence, Kochar said. The two were arrested two years ago after body parts packed in plastic bags were found near their home in Noida, a New Delhi suburb. Their home was later dubbed a ”house of horrors” by the Indian media. Pandher was not named a main suspect by investigators initially, but was summoned as co-accused during the trial, Kochar said. Kochar said his client was in Australia when the teen was raped and killed. Pandher faces trial in the remaining 18 killings and could remain in custody, the attorney said.

The question “what was the case of the house of horrors in northern india” can be answered by the whole news article. There is no short span can be extracted as an answer.

None answer in the article

Given a news:

> Buy a $175,000 package to attend the Oscars and you might buy yourself trouble, lawyers for the Academy Awards warn. The 81st annual Academy Awards will be held on February 22 from Hollywood’s Kodak Theatre. The advertising of such packages – including four tickets to the upcoming 81st annual Academy Awards and a hotel stay in Los Angeles, California – has prompted the Academy of Motion Picture Arts and Sciences to sue an Arizona-based company. The Academy accused the company Experience 6 of selling ”black-market” tickets, because tickets to the lavish movie awards show cannot be transferred or sold. Selling tickets could become a security issue that could bring celebrity stalkers or terrorists to the star-studded event, says the lawsuit, which was filed Monday in federal court in the Central District of California. ”Security experts have advised the Academy that it must not offer tickets to members of the public and must know identities of the event attendees,” the lawsuit says. ”In offering such black-market tickets, defendants are misleading the public and the ticket buyers into thinking that purchasers will be welcomed guests, rather than as trespassers, when they arrive for the ceremony.” Experience 6 did not return calls from CNN for comment. On Tuesday morning, tickets to the event were still being advertised on the company’s Web site. The Oscars will be presented February 22 from Hollywood’s Kodak Theatre. The Academy Awards broadcast will air on ABC. Hugh Jackman is scheduled to host.

The questions “where does the 81st annual academy awards come from” and “how much did the academy pay to attend the oscars” cannot be answered from the news.