"What makes a question inquisitive?"
A Study on Type-Controlled Inquisitive Question Generation

Lingyu Gao[*1], Debanjan Ghosh[2], and Kevin Gimpel[1]

[1]Toyota Technological Institute at Chicago
[2]Educational Testing Service
{lygao, kgimpel}@ttic.edu,
dghosh@ets.org

Abstract

We propose a type-controlled framework for inquisitive question generation. We annotate an inquisitive question dataset with question types, train question type classifiers, and fine-tune models for type-controlled question generation. Empirical results demonstrate that we can generate a variety of questions that adhere to specific types while drawing from the source texts. We also investigate strategies for selecting a single question from a generated set, considering both an informative vs. inquisitive question classifier and a pairwise ranker trained from a small set of expert annotations. Question selection using the pairwise ranker yields strong results in automatic and manual evaluation. Our human evaluation assesses multiple aspects of the generated questions, finding that the ranker chooses questions with the best syntax (4.59), semantics (4.37), and inquisitiveness (3.92) on a scale of 1-5, even rivaling the performance of human-written questions.

1 Introduction

Recently, interest has grown in the task of automatic question generation (AQG) from text (Sun et al., 2018; Kumar and Black, 2020). AQG is useful in building conversational AI systems (Bordes et al., 2017; Gao et al., 2019), generating synthetic examples for QA (Alberti et al., 2019; Dong et al., 2019; Sultan et al., 2020), and educational applications, such as intelligent tutoring and instructional games (Chen et al., 2018; Flor and Riordan, 2018). In the majority of such studies, AQG focuses on generating factual questions that tend to ask about specific information in the text (i.e., "who did what to whom") (Du et al., 2017).

Instead of asking factual questions with answers already present in the text, Ko et al. (2020) argued that human readers instinctively ask questions that are curiosity-driven, answer-agnostic, and seek a

Context	. . . The plan places an indicated value on the real estate operation, Santa Fe Pacific Realty Corp., of $ 2 billion.
Source sentence	Santa Fe Pacific directors are expected to **review** the plan at a meeting today, according to people familiar with the transaction.
BASE	What kind of meeting?
SPAN	How will the directors review the plan?
Explanation	Why are they reviewing the plan?
Background	Is it expected to review the plan today?
Elaboration	What will the review entail?
Instantiation	Which directors are expected to review the plan?
Definition	what is that?
Forward	What are the directors expected to review?
Informative	Who are Santa Fe Pacific directors expected to review?

Table 1: Examples of generated questions given the article context and source sentence (with span in **bold**).

high-level understanding of the document being read. They released a dataset of such curiosity-driven questions (henceforth INQUISITIVE; for details, see Section 2). The objective of our work is to generate deeper, inquisitive questions based on the INQUISITIVE dataset.

Our motivations for generating inquisitive questions are two-fold. Educators can obtain *diverse* questions for a specific source text when designing quizzes or choosing questions to test students' reading comprehension ability. They can focus into different aspects of the context (e.g., questioning the background information or asking to elaborate a fact) for diverse question generation (Cho et al., 2019; Wang et al., 2020; Sultan et al., 2020). Likewise, students can also be assisted in knowledge acquisition and building reasoning skills by practising over a large number of diverse questions (Cao and Wang, 2021).

Though our initial efforts are similar to Ko et al. (2020), we found this to be insufficient as it does not leverage the inherent diversity of question types

---

240

in the dataset. Ko et al. (2020) concatenated the context, source sentence, and the question to learn a language model for question generation using GPT-2 (Radford et al., 2019). On the contrary, we first annotated 1550 questions from the training partition of the INQUISITIVE dataset to identify the question types, such as questions requesting background information, asking about the cause of an event, asking for details on underspecified facts, etc. (see Section 2.2 for details). We finetune a RoBERTa-large model (Liu et al., 2019) to predict the question types on the rest of the dataset. We then use the question types in a controlled generation framework based on BART (Lewis et al., 2020) to generate type-specific inquisitive questions.

Consider the example in Table 1. The BASE model is BART finetuned on INQUISITIVE to generate questions from the context and source sentence. The SPAN model additionally uses the span, a part of the source sentence the annotators are curious about. We then show six questions of specific types (e.g., Explanation, ..., Forward) generated by our type-controlled finetuned BART model. In comparison, the informative question is generated by finetuning on SQuAD (Rajpurkar et al., 2016), a popular dataset for generating factual questions. The informative question asks for surface-level information ("who are Santa Fe Pacific directors expected to review?") whereas the inquisitive questions ask for deeper information (e.g., "why are they reviewing the plan?"), such as the reason for the directors' actions.

As mentioned earlier, our motivations for generating diverse inquisitive questions are to provide educational tools and resources. However, there are also cases where an educator or student may prefer only a single high-quality question for a span or a ranked list of questions. We investigate two strategies for automatic question selection/ranking for this latter scenario. The first strategy ranks questions using an inquisitive vs. informative question classifier, where questions from SQuAD are used as informative questions. In the second strategy, we collect expert annotations of partial rankings for a subset of generated questions, and then train a pairwise ranker to select the best question (denoted as TYPE_r). In automatic evaluation, we find that TYPE_r yields questions that have reasonably strong match to references while also being novel relative to the training set (Section 4.1). We report a large-scale human evaluation via Mechanical Turk

and demonstrate that questions generated from the same TYPE_r model have the best syntax (4.59), semantics (4.37), and inquisitiveness (3.92) on a scale of 1-5 (Section 4.2). We make the annotations, code, and the MTurk judgements from our research publicly available.[1]

2 Data

We will now describe the annotation of questions with question types, which is one of the main contributions of our work. We describe the annotation process in detail in Section 2.2. But first, we briefly introduce the INQUISITIVE dataset.

Human annotators created the inquisitive questions while reading the initial part (i.e., five sentences) of news articles from the WSJ portion of the Penn Treebank (Marcus et al., 1993) or Associated Press articles from the TIPSTER corpus (Harman and Liberman, 1993).[2] Annotators first highlighted a span within the sentence that they were curious about and then wrote a maximum of three questions. Next, a separate set of annotators validated the questions and excluded unqualified questions (around 5%).

An instance in INQUISITIVE has the following components: a **source sentence**, the sentence the annotator read when asking the question, **context** that includes all the sentences before the source sentence in the same article, a **span** within the source sentence the annotators were most curious about, and finally, the **question** the annotator wrote. INQUISITIVE is split into *training* (15,897 instances), *test* (1,885 instances), and *dev* (1,984 instances).

2.1 Question Type Annotation

In the USA, K-12 standards describe what students should understand and be able to do by the end of each grade.[3] The guidelines state that even in very early grades students should understand how individuals and events evolve and interact in a text. The *hows* and *whys* of the text (i.e., inquisitive questions) come naturally to us (Ko et al., 2020).

Ko et al. (2020) evaluated the question types over a small set of 120 questions and identified a few question types that appear frequently and address various *how* and *why* questions.[4] Although they

---

[1] https://github.com/EducationalTestingService/inquisitive-questions
[2] They also use Newsela (Xu et al., 2015) but it was not publicly released.
[3] http://www.corestandards.org
[4] The evaluation is not available in the released dataset.

Question Type (# samples)	Example	
	[*context*] [*source sentence* with span in **bold**]	Question
Explanation (443)	[. . . unraveling of the on-again, off-again UAL buy-out slammed the stock market.][Now, stock prices seem to be in a general **retreat**.]	Why are the stock prices retreating?
Elaboration (364)	[. . . Beth Capper has gone without food . . .][It's not drugs or alcohol or even baby formula that has **put her in such a bind**.]	What has put her in this bind?
Background (407)	[. . . John R. Stevens, . . . , was named senior executive vice president. . .][He **will continue** to report to Donald Pardus, . . .]	How long has he been reporting to Donald Pardus?
Definition (114)	[Oh, that terrible Mr. Ortega.][Just when American liberalism had pulled the **arms plug** on the Contras . . .]	What is the arms plug?
Instantiation (159)	[. . . in their office, Rajiv Maheswaran and Yu-Han Chang can catch a glimpse of Staples Center . . .][Whiteboards inside their office are filled with **algorithms** in shades of red, blue and green.]	what kind of algorithms?
Forward-looking (31)	[The federal government would not actually shut down. Agents would still patrol . . .][Mail carriers would **still deliver mail**.]	Would it arrive on time?
Other (32)	[. . . the entire neighborhood can fall victim.] [At this stage some people just **"walk away"** from homes. . .]	Why is it quoted?

Table 2: Annotated question type distributions and salient examples of each question type. Context and source sentences are presented where the spans in source sentences are bold. More examples are in the Appendix.

presented a fully data-driven approach without any theoretical underpinnings we notice such curiosity driven questions – such as asking for background information, elaborating details, and why one action led to another – are linked to Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In RST, relations such as *background*, *elaboration*, and *cause* provide a systematic way to analyze the text and understand the discourse relations among segments of the text. Likewise, the questions generated in this work inquire about the background or causal information and those are close to the rhetorical relations in the text. For our annotation, we use the same set of question types as Ko et al. (2020), which are described below:

- **Explanation**: Questions signaled by the interrogative "why" as well as its paraphrases such as "what is the reason". These questions are often asked to explain why something happened or identify its cause ("why did he choose to speak to the press?").
- **Elaboration**: Questions that seek more details about concepts, entities, relations, or events expressed in the text, e.g., "what are some details about this performance?"
- **Background**: Questions that seek more information about the context of the story or seek clarification about something described in the text ("how much loan was guaranteed?").

- **Definition**: Questions that ask for the meaning of a specific term ("what does hubris mean?").
- **Instantiation**: Questions that ask about a specific instance (e.g., "what is the name of the newspaper?") or a set of instances (e.g., "who are these other cable partners?").
- **Forward-looking**: Questions that ask about future events, e.g., "would it arrive on time?"
- **Other**: Other types of questions, e.g., inference questions ("how many women were found?") that ask to deduce information from the source, or that ask something irrelevant ("Does seaweed look like cotton candy?")

Three expert annotators who are experienced at annotation tasks initially annotated 50 questions with the types above. Pairwise κ's between annotators were 0.570, 0.572, and 0.872 (moderate and substantial agreement). The annotators exchanged notes and decided on final annotation guidelines. In the next round, each annotator independently annotated 500 random questions from the training partition of INQUISITIVE, thus producing a total set of 1,550 annotated questions. We used majority vote for the first 50 questions. Table 2 presents the question type distribution with salient examples.

Table 3 shows the most common leading unigrams for each question type in our annotated

Explanation	Elaboration	Background	Definition	Instantiation	Forward-looking	Other
why (396)	what (164)	what (108)	what (95)	what (62)	what (9)	why (5)
what (28)	how (135)	how (91)	does (5)	which (50)	how (8)	does (5)
is (5)	is (11)	is (40)	how (3)	who (36)	will (3)	is (4)
how (4)	where (6)	who (34)	who (2)	in (3)	would (2)	what (3)
if (3)	in (5)	where (18)	definition (2)	at (2)	did (2)	of (2)

Table 3: Most common leading unigrams in annotated questions (lowercased) for each type (counts in parentheses).

data.[5] Although WH question words such as "why", "when", "who", etc. have been used to generate a variety of question types before (Zhou et al., 2019), they cannot fully express the semantic content of questions (Cao and Wang, 2021). Likewise, we observe there is no one-to-one relationship between WH words and question types. Each type encompasses multiple question words. Some types, like Explanation and Definition, have a single dominant leading unigram, while others have two or three. The word "what" is the most common leading unigram for five question types.

2.2 Question Type Prediction

We aim to generate a question that follows a particular question type as control code. However, to do so, we must first determine the question types in the entire INQUISITIVE dataset. To this end, we finetune RoBERTa-large as a multi-class classifier on the annotated set of 1,550 questions and use the classifier to predict the question types of the remaining questions in INQUISITIVE. As input, we concatenate the context, source sentence, span, and question, using the "[SEP]" token as delimiter.[6] We use 1,400 examples for training and the remaining 150 as the validation set (also used for early stopping)[7], on which we reach an accuracy of 73.3%.

3 Methods

In this section, we present our computational approaches for question generation. The input x is a sequence of tokens $x = \langle x_1, ..., x_n \rangle$, which may consist of one or more sentences. The output is a question q that consists of sequence of tokens, i.e., $q = \langle q_1, ..., q_m \rangle$. Using the standard autoregressive sequence-to-sequence architecture (Sutskever et al.,

2014) we model $P_\theta(q \mid x)$ as follows:

$$P_\theta(q \mid x) = \prod_i P_\theta(q_i \mid q_1, \ldots, q_{i-1}, x) \quad (1)$$

We use the pretrained BART model (Lewis et al., 2020), a transformer (Vaswani et al., 2017) composed of a bidirectional encoder and an autoregressive decoder. In our simplest setup (called BASE), we concatenate the source sentence and the context. The next setting also concatenates the span; we refer to it as SPAN. Each element (e.g., context, span) is separated with the special token "[SEP]".

3.1 Controlled Generation

Our next set of models use the question types as control codes to guide question generation. Controlled generation models (Kikuchi et al., 2016; Hu et al., 2017; Ficler and Goldberg, 2017; Tsai et al., 2021) condition on a control code c in addition to the input x to model the distribution of $P_\theta(q \mid x, c)$. Similar to Eq. (1), we can write,

$$P_\theta(q \mid x, c) = \prod_i P_\theta(q_i \mid q_1, \ldots, q_{i-1}, x, c) \quad (2)$$

Text generation conditioned on such control codes, such as sentiment control of movie reviews, style for chatbots, diverse story continuations, etc., have been used effectively in recent research (Tu et al., 2019; Krause et al., 2021; Roller et al., 2021). We use the same idea for question generation by conditioning on the question type c as identified in Section 2.2. We simply concatenate the question type as an additional token and finetune BART. Using the example from Table 1, the input to BART with the question type Explanation would be:
The plan places ...2 billion [SEP] Santa Fe ...transaction [SEP] review [SEP] Explanation

Inference. We specify the question type to generate specific questions. Top-k sampling with $k = 5$ is used to generate questions, where the questions are constrained to be from 5 to 30 tokens, with a length penalty 2.0 (Ott et al., 2019). The length

---

[5]See appendix for bigrams.

[6]We use a special token "NO_CONTEXT" if the source sentence is the first sentence in the article.

[7]We keep the distribution of question types in train and dev set roughly the same, and the majority question type (Explanation) is about 29% of the total data.

penalty is an exponential penalty on the length, where a penalty > 1 favors longer generations.

For each test instance, we generate a question for all question types except "Other".[8] Table 1 shows examples of generated questions.

3.2 Automatic Question Type Selection

As stated in the Introduction section, besides being able to generate a variety of questions based on a single span, another motivation of this work is to identify a single high-quality question or to rank the list of the questions. In case of controlled generation, one challenge is determining what control code to use at inference time when a single output is desired. We explore two ways to choose a single question from the six generated for each input.

Informative vs. inquisitive question classifier. We consider using a binary question classifier (RoBERTa-large with default parameters) to classify whether a question is from INQUISITIVE or SQuAD. We view SQuAD questions as more "informative" than inquisitive so we hope for this classifier to capture what it means for a question to be inquisitive. We train on the training questions in INQUISITIVE and an equal number of questions drawn from SQuAD.[9] At inference time, given one generated question for each type, we choose the one that maximizes the classifier's probability of being inquisitive. Our hypothesis is that an inquisitive/informative classifier can serve as a scoring function for selecting the best candidate from a set of inquisitive questions. For the example in Table 1 the classifier chose the Definition question "what is that?" with the highest inquisitiveness probability. Below we refer to this method as TYPE$_s$, where the "s" indicates that the SQuAD dataset is used.

Pairwise ranking classifier with expert annotations. In this setup, we collect a small set of question ranking annotations and train a pairwise ranking classifier (Liu et al., 2009) to select the best question. First, we randomly select 300 instances from the 1,885-instance test set from INQUISITIVE. Next, two expert annotators (each with extensive annotation experience) independently ranked each

of the six generated questions per instance. The annotators' task was to rank the questions according to their inquisitiveness and relevance to the context, source, and span. The annotators judged all six questions for each instance and identified at least three questions (rank 1-3) as the best where the rest of the questions were deemed to be of lower quality. In some cases, the annotators even ranked top-five questions (rank 1-5). Precision@1, 2, 3 ranks are 0.70, 0.88, and 0.95 respectively (i.e., in 70% cases one annotator's top-1 selection was found in the other annotator's top-3 selection).[10]

We then approximate the learning-to-rank problem (Joachims et al., 2007; Liu et al., 2009) with a classification problem, i.e., by training a binary classifier to determine whether one question is better than another. For a single input, let Q, q_{rel}, and q_{nrel} represent the total set of generated questions, relevant questions, and irrelevant questions, respectively. In our pairwise ranking setup, the training instances are the combination of (a) a question q_i from q_{rel} and a question q_j from q_{nrel}, and (b) two questions q_i and q_j from q_{rel} if and only if the two questions are separated by ≥ 2 ranks. Algorithm 1 in the appendix details the procedure.

In addition to the two questions q_i and q_j, we also use the source sentence as another input. During training, for each instance from (a) and (b) above, we create two training examples of the form *source* + "[SEP]" + q_i + "[SEP]" + q_j and *source* + "[SEP]" + q_j + "[SEP]" + q_i. If the first question in the sequence has a better rank we label the instance as positive, otherwise negative. This way we have 2,867 examples; we use 2,581 for training and the rest for validation. We finetune a RoBERTa-large model as a binary classifier with default hyperparameters, attaining a validation accuracy of 76.2%.

For each test instance, similar to the training setup, for each generated question pair q_i, q_j we form a pair of examples. Given that we have six question types, we create altogether thirty examples and classify them using the RoBERTa-large classifier. We return the question that is preferred the largest number of times.[11] Given the example in Table 1 this model selects the Explanation question, i.e., "Why are they reviewing the plan?" Below we refer to this method as TYPE$_r$, where the "r" represents the use of the ranker described above.

---

[8]We made this choice because "Other" includes many subtypes, e.g., inference questions and comparisons, giving us only a few examples per type. We leave this to future work.

[9]We also attempted to include the source sentences. However, given the differences between the two datasets (WSJ/AP for INQUISITIVE vs. Wikipedia for SQuAD), this caused the classifier to focus more on the source sentences than the questions.

[10]Please refer to Table 9 in the Appendix section for examples.

[11]In case of ties, we use the classifier scores as the tie breaker.

4 Experiments

For all models, we use BART-large with the same settings. For training, we use the Adam optimizer (Kingma and Ba, 2015) with learning rate 3e-5, weight decay 0.01, clip norm 0.1, dropout 0.1, 15 epochs in total, warm-up updates 500, use cross entropy loss with label smoothing ($\alpha = 0.1$), and set the maximum number of tokens per batch to 1024. More details of the experimental setup are given in the Appendix (Section A.1).

We evaluate the following five settings:

- BASE: uncontrolled generation using the context and source sentence as input
- SPAN: uncontrolled generation using the context, source sentence, and span as input
- TYPE_s: type-controlled generation with type selection via informative vs. inquisitive classifier
- TYPE_r: type-controlled generation with type selection via pairwise ranking classifier
- TYPE_o: type-controlled generation with question type of reference question

Since the TYPE methods use question types, in order to compare those methods to others, we need a way to automatically select a single generated question. For TYPE_o, we run our question type classifier on a human-written reference question and use the predicted type. Thus, TYPE_o is an oracle method (hence the mnemonic "o" in its name) since it assumes access to a reference question. For TYPE_s and TYPE_r we use the classifiers described in Section 3.2. All TYPE methods use the context, source sentence, and span as input, like SPAN.

4.1 Automatic Evaluation

Since inquisitive question generation is an open-ended task, a high-quality generated question may not overlap with the gold question. However, automatic metrics that measure the overlap between generations and gold questions could still be useful diagnostics for characterizing models.

Table 4 presents several automatic metrics: BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), perplexity under GPT2-XL (Radford et al., 2019), and the entropy (averaged over questions) of the RoBERTa-large question type classifier applied to the generated question.[12] Although INQUISITIVE contains

---

[12] We reported the average scores of 5 runs with different random seeds.

a test set of 1,885 instances (see Section 2), we used only 1,585 instances as our test set because we chose the remaining (random) 300 instances to build our pairwise ranking classifier (Section 3.2).

For BLEU, METEOR, ROUGE-L, and BERTScore, the oracle model TYPE_o achieves the highest scores, presumably because this model generates questions that are similar to the reference types. We notice, TYPE_r, and SPAN have similar scores, with TYPE_r being slightly ahead for BLEU and METEOR. In the case of TYPE_s, the low scores across metrics can be attributed to the fact that the inquisitive vs. informative classifier prefers question types that are unique to the INQUISITIVE dataset, such as Definition and Instantiation questions. These types are not appropriate for all spans and in many cases are quite different from the reference questions.

We also find that TYPE_r has the lowest GPT2 perplexity, indicating that the ranker is favoring highly probable questions according to a general-purpose language model. A lower perplexity is likely indicative of greater fluency, a point we will return to in our human evaluations. Likewise, the lowest entropy of TYPE_r implies that its questions can be classified with high confidence by our question type classifier. In contrast, TYPE_s shows higher entropy, i.e., its questions are more difficult to classify. The entropy of the human-generated questions is higher than nearly all of our models, indicating that the human questions are also harder to classify than model outputs.

The last three columns of Table 4 show the metrics designed by Ko et al. (2020), namely *Train-n*, *Article-n*, and *Span*. These metrics measure the extent of copying from the source materials into the generated questions, i.e., % of n-grams in the generated questions that appear in the training questions (*Train-n*) and the context/source sentence (*Article-n*). For brevity, we only report *Train-2* and *Article-2*. *Span* measures the % of words in the annotated span present in the generated questions.

Among our models, TYPE_r attains the lowest value of the *Train-2* metric, which is also closest to the HUMAN value. Aside from TYPE_s, the other models have higher *Article-2* than HUMAN, meaning that the generated questions have a higher % of n-grams that appear in the source sentence or the context. TYPE_r has the highest value for the *Span* metric, indicating that the ranker prefers questions that use words from the span. SPAN is second high-

Model	%BLEU	%METEOR	%ROUGE-L	%F$_{\text{BERT}}$	GPT2 ppl	Entropy	Train-2	Article-2	Span
HUMAN	-	-	-	-	272	0.777	0.467	0.126	0.354
BASE	4.3	11.8	27.4	39.6	119	0.699	0.518	0.186	0.184
SPAN	8.5	17.5	36.1	47.6	148	0.726	0.505	0.182	0.452
TYPE$_s$	5.7	13.6	30.9	41.6	219	0.823	0.530	0.090	0.346
TYPE$_r$	8.6	18.3	35.3	47.4	89	0.612	0.473	0.195	0.542
TYPE$_o$	9.7	19.5	39.1	50.1	154	0.751	0.488	0.149	0.475

Table 4: Automatic metrics on our test set for our models as well as the reference questions (HUMAN).

Model	Syntax	Semantics	Relevancy	Inquisitive
BASE	4.30	4.11	4.16	3.71
SPAN	4.30	4.17	4.32	3.75
TYPE$_s$	4.02	3.50	3.51	3.14
TYPE$_r$	4.59	4.37	4.27	3.92
TYPE$_o$	4.33	4.10	4.09	3.78
HUMAN	4.36	4.41	4.33	3.98

Table 5: Results of human evaluation. The HUMAN row shows judgments for reference questions from the INQUISITIVE dataset.

est and BASE, which does not use the annotated span, has the lowest value.

In the Appendix, we also report an automatic evaluation of controllability, finding that certain question types (Explanation, Definition, and Instantiation) can be generated with high precision, while others (Elaboration, Background, and Forward-looking) are more easily confused.

4.2 Human Evaluation

In this section, we report the results from a human evaluation we have conducted to assess a variety of subjective aspects of the generated questions, namely the *syntax*, *semantics*, *relevancy*, and the degree of *inquisitiveness*.

We collected annotations using the crowdsourcing platform Amazon Mechanical Turk (MTurk). We randomly selected 500 test instances. For each, we asked three annotators the following four questions to measure quality along four aspects:

1. Does the question seem syntactically correct?

2. Does the question make sense (semantically)?

3. Does the question seem relevant to the source?

4. Does the question show inquisitiveness to learn more about the topic?

The annotators were given the following three options to choose from: *yes*, *somewhat*, and *no*. Each human intelligence task (HIT) contained five instances to judge and we paid $2 per HIT.

1	is it the aha?
2	how much has inflation?
3	nativity happens for buddha?
4	When he decide?
5	how much has inflation?

Table 6: Examples of gold questions from INQUISITIVE dataset that are judged as ungrammatical by the Turkers.

Table 5 presents the average of the human judgments, where the answers *yes*, *somewhat*, and *no* are converted to scores 5, 3, and 1, respectively. In all four aspects, we notice several scores are over 4. For the *inquisitiveness* aspect, the TYPE$_r$ model achieves the highest score among all models. This score is higher than the oracle model (TYPE$_o$) showing the usefulness of the ranker to generate inquisitive questions. Likewise, TYPE$_r$ achieves the highest average score for *semantics*, showing that its questions are semantically meaningful almost all the time. We also note that both TYPE$_r$ and SPAN are competitive in *relevancy*. Finally, for *syntax*, each model (aside from TYPE$_s$) was rated close to 4.5. Although transformers usually produce fluent output (Yates et al., 2021), TYPE$_r$ scored higher than the human generated gold questions on *syntax*, which warrants further investigation.

We manually analyzed all the questions from TYPE$_r$ and HUMAN where the majority of annotators rated 1 for *syntax*, and we found out there are 14 and 33 such questions, respectively, explaining why TYPE$_r$ scored higher. This also explains why the perplexity of HUMAN is high in Table 4. In Table 6 we provide a few examples from the HUMAN set that were deemed grammatically incorrect. Table 14 in the Appendix contains examples of grammatical errors from all models.

If we compute a simple average over all the aspects for each model, TYPE$_r$ scores the highest, slightly better than HUMAN, and the SPAN model is second. The SPAN model is trained on the IN-QUISITIVE dataset so it has the freedom of generating inquisitive questions. However, what separates

Source	Model	Question	Syn.	Sem.	Rel.	Inq.
. . . The State Security court said it was ordering the seizure of Kemal's book, . . . , because it **provokes** **"hatred and enmity"** on the basis of differences . . .	BASE	What is the reason for the seizure?	4.3	3.7	4.3	4.3
	SPAN	How does the author's book provoke hatred and enmity?	5.0	5.0	5.0	5.0
	TYPE$_s$	what is hatred?	4.3	4.3	2.3	1.7
	TYPE$_r$	How can a book provoke hatred and enmity?	5.0	5.0	5.0	3.0
	TYPE$_o$	How did it provocate the book and what did the author write in the book?	5.0	2.3	3.0	2.3
	HUMAN	How does the book provoke "hatred and enmity"?	4.3	5.0	4.3	3.7
Olin Corp. said **third-quarter** net income rose 26 % on the strength of its chemical business.	BASE	Why were they on the strength of their chemical business?	1.0	1.0	3.0	3.0
	SPAN	what happened to the second quarter?	3.7	3.7	4.3	3.0
	TYPE$_s$	What was the quarter?	4.3	4.3	4.3	4.3
	TYPE$_r$	Why was third-quarter net income up 26 %?	5.0	5.0	4.3	4.3
	TYPE$_o$	What year did the third-quarter consist of?	5.0	5.0	4.3	5.0
	HUMAN	What happened to the net income in the first and second quarter?	3.7	5.0	5.0	5.0
. . . most significant change in surgical training since the early 1900s, they are working with **local medical device companies** to develop new generations of software . . .	BASE	How are medical device companies working with the University of Minnesota??	5.0	5.0	5.0	4.3
	SPAN	Which local medical device companies?	2.3	3.0	4.3	3.0
	TYPE$_s$	who are the local medical device companies?	4.3	3.7	2.3	2.3
	TYPE$_r$	Why are they working with local medical device companies?	5.0	5.0	5.0	5.0
	TYPE$_o$	Who are the local medical device companies?	5.0	3.7	4.3	5.0
	HUMAN	Which medical device companies are being worked with?	2.3	3.7	5.0	5.0

Table 7: Examples of generated questions from different models. Syn., Sem., Rel., Inq. represent Syntax, Semantics, Relevancy and Inquisitiveness, respectively. For brevity the context is not shown. Spans are bold.

SPAN from TYPE$_r$ is, for the latter, we have the ability to control the generation with specific question types and also select the *best* question for the same source sentence. We also notice that the generations from TYPE$_s$ scored lowest across all four aspects. The TYPE$_s$ model often selects Definition/Instantiation question types that are unsuitable for the source sentence and the span, which is why the annotations score low for this type of question.

Table 7 shows several examples from our models along with average human ratings for all four aspects. We highlight three salient observations here. First, in general, TYPE$_r$ has high scores across all aspects for all examples. Second, the Turkers have treated the aspects independently as we have requested. Even if they rated the HUMAN annotations 2.3 and 3.7 for *syntax* and *semantics* for the last example, they have given high ratings for the other two aspects. Third, interestingly, "what is hatred?", a very generic question, scored high on *syntax* and *semantics* (TYPE$_s$ model for the first example) but low on the other two aspects due to its lack of *relevancy* and *inquisitiveness*.

Finally, we note that for the first example in Table 7, the SPAN and HUMAN questions are extremely similar, but their ratings differ for three out of the four attributes. This example illustrates the variability of human judgments for this task, which suggests that more annotations may be needed to increase confidence in our results.

5 Related Work

In recent years, automatic question generation has attracted many NLP researchers, perhaps due to its versatility, e.g., question generation for conversational AI (Bordes et al., 2017; Gao et al., 2019), synthetic examples for QA tasks (Alberti et al., 2019; Olney et al., 2012; Sultan et al., 2020), clarifications on information-seeking conversation (Aliannejadi et al., 2019), and knowledge evaluation and educational application areas (Mitkov and Ha, 2003; Brown et al., 2005; Chen et al., 2009; Stasaski et al., 2021), which is specifically related to our use cases.

In earlier work, methods such as transforming declarative sentences into questions (Heilman and Smith, 2010) or using semantic roles (Flor and Riordan, 2018) were popular. However, recently sequence-to-sequence architectures (Du et al., 2017; FitzGerald et al., 2018) and pretrained models (Cao and Wang, 2021) are more often used. Similar to Ko et al. (2020), our work is related to answer-agnostic question generation. We focus on exploiting question type information for generating deeper questions. Although related work in the answer-unaware setting exists (Nakanishi et al., 2019), they mostly focus on identifying question-worthy text for generation (Scialom and Staiano,

2020; Wang et al., 2019) from factual (Rajpurkar et al., 2016), conversational (Choi et al., 2018), or social media platforms (Fan et al., 2019), different from the WSJ/AP news dataset used in our work.

We are building on past work on controllable generation, generating text that reflects specific characteristics of control variables. In some earlier work, embedding vectors of the control variables were fed into the model for controlling the output (Kikuchi et al., 2016; Fan et al., 2018; Tu et al., 2019). However, our approach resembles recent efforts where the control variable is concatenated to the main input using some separator (Keskar et al., 2019; Schiller et al., 2021). Methods such as PPLM are useful for similar guided controllable generations (Dathathri et al., 2020); however, PPLM requires gradient descent at inference time, while our question type selection approach is highly scalable and efficient.

We consider controllable question generation based on specific question types, noting that different question templates or ontologies have been studied for question generation. For example, a Wikipedia-driven ontology is used for generation (Labutov et al., 2015), or contextualized questions are generated for any semantic role (Pyatkin et al., 2021). Likewise, Pascual et al. (2021) proposed guided generation focusing on including specific keywords (e.g., "wh" words for questions), while we showed in Table 3 that "wh" words do not have a 1-to-1 relationship with question types.

Our work is closer to that of Cao and Wang (2021), who proposed a question type ontology (based on cognitive science) inspired by manually constructed templates (Olney et al., 2012). On the contrary, we chose a dataset that focuses on inquisitive questions only and chose our question types accordingly, while they used a dataset with a broader set of questions. In addition, instead of predicting the text span ("focus" in (Cao and Wang, 2021)) we directly use the annotated span in our research. Finally, we focused on post-processing the generations to identify the best question (or rank them) related to the source content.

6 Conclusions and Future Work

We proposed a type-controlled framework that generates inquisitive questions given a source sentence, annotated span, and a longer context. We annotated a set of question types related to curiosity driven questions and demonstrated that our framework can generate a variety of questions from a single input. We also developed an effective method (TYPE_r) to select a single question using a pairwise ranker trained on a small set of ranking annotations. Our generations, especially from TYPE_r, show high novelty. The human evaluation demonstrates that questions generated from TYPE_r rival human-written questions on all four aspects of quality.

Future work could include annotating a larger partition of the INQUISITIVE dataset while exploring finer-grained analysis of question types (e.g., sub-categories of elaboration questions). We are also interested in employing a framework to generate questions and identify the span jointly.

Acknowledgments

7 Ethical Considerations

We leverage the freely available open access question dataset INQUISITIVE for annotation and model training. Though we have not exhaustively checked the source dataset manually, given they are sourced from the WSJ partition of the Penn Treebank and Associated Press articles from the TIPSTER corpus, we consider them relatively safe and do not find any objectionable content.

Training is done using large pretrained models that have been shown to have bias. Although the generated questions do not appear biased, they may hallucinate content, which is a common problem for neural generation models.

Finally, we obtained institutional review board permission to conduct MTurk based data collection.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. 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.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking

conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 475–484. ACM.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. 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 6424–6439, Online. Association for Computational Linguistics.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. LearningQ: A large-scale dataset for educational question generation. In *Proceedings of International Conference on Web and Social Media (ICWSM)*.

Wei Chen, Gregory Aist, and Jack Mostow. 2009. Generating questions automatically from informational text.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. 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 3121–3131, Hong Kong, China. Association for Computational Linguistics.

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.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.

Michael Flor and Brian Riordan. 2018. A semantic role-based approach to open-domain automatic question generation. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 254–263, New Orleans, Louisiana. Association for Computational Linguistics.

Yifan Gao, Piji Li, Irwin King, and Michael R. Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4853–4862, Florence, Italy. Association for Computational Linguistics.

Donna Harman and Mark Liberman. 1993. TIPSTER complete LDC93T3A. web download. Philadelphia: Linguistic Data Consortium.

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, Los Angeles, California. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Thorsten Joachims, Hang Li, Tie-Yan Liu, and ChengXiang Zhai. 2007. Learning to rank for information retrieval (LR4IR 2007). *SIGIR Forum*, 41(2):58–62.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

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*.

Wei-Jen Ko, Te-Yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6544–6555. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vaibhav Kumar and Alan W. Black. 2020. Clarq: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7296–7301. Association for Computational Linguistics.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 889–898, Beijing, China. Association for Computational Linguistics.

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.

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.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

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

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ruslan Mitkov and Le An Ha. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 17–22.

Mao Nakanishi, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2019. Towards answer-unaware conversational question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 63–71, Hong Kong, China. Association for Computational Linguistics.

Andrew McGregor Olney, Arthur C. Graesser, and Natalie K. Person. 2012. Question generation from concept maps. *Dialogue Discourse*, 3(2):75–99.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Valentina Pyatkin, Paul Roit, Julian Michael, Yoav Goldberg, Reut Tsarfaty, and Ido Dagan. 2021. Asking it all: Generating contextualized questions for any semantic role. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1429–1441, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Thomas Scialom and Jacopo Staiano. 2020. Ask to learn: A study on curiosity-driven question genera-tion. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2224–2235, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A. Hearst. 2021. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170, Online. Association for Computational Linguistics.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Alicia Y. Tsai, Shereen Oraby, Vittorio Perera, Jiun-Yu Kao, Yuheng Du, Anjali Narayan-Chen, Tagyoung Chung, and Dilek Hakkani-Tür. 2021. Style control for schema-guided natural language generation. *CoRR*, abs/2109.12211.

Lifu Tu, Xiaoan Ding, Dong Yu, and Kevin Gimpel. 2019. Generating diverse story continuations with controllable semantics. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 44–58, Hong Kong. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Siyuan Wang, Zhongyu Wei, Zhihao Fan, Yang Liu, and Xuanjing Huang. 2019. A multi-agent communication framework for question-worthy phrase extraction and question generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial*

Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 7168–7175. AAAI Press.

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020. Diversify question generation with continuous content selectors and question type modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2134–2143, Online. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Trans. Assoc. Comput. Linguistics*, 3:283–297.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Question-type driven question generation. 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 6032–6037, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Experimental Setup

For BASE, SPAN, TYPE$_s$, TYPE$_r$, and TYPE$_o$, we use BART-large with the same settings. We train 15 epochs in total, using cross entropy loss (label smoothing with $\alpha = 0.1$), and set the maximum number of tokens per batch as 1024. There's a normalization layer after the embedding layer, and the embedding matrices for encoder input, decoder input, and decoder output are tied. For training, we use the Adam optimizer (Kingma and Ba, 2015) with learning rate 3e-5, weight decay 0.01, clip norm 0.1, dropout 0.1, and warm-up updates 500.

For the question type classifier, we finetune RoBERTa-large for 15 epochs with batch size 8. We use Adam with learning rate 1e-5, weight decay 0.1, dropout 0.1, and warm-up updates 157. We use the same settings for the inquisitive vs. informative classifier and pairwise ranking classifier except some hyperparameters. For the inquisitive vs. informative classifier, we train for 10 epochs with batch size 32 and warm-up updates set to 300. For the pairwise ranking classifier, we train for 20 epochs with warm-up updates set to 387. Under this setting, we compute all warm-up updates with $6\% N_{tr} N_{epo}/N_{bsz}$, where N_{tr} is the training set size, N_{epo} is the number of training epochs, and N_{bsz} is the batch size.

A.2 Leading Bigrams for Question Types

Table 8 shows the most common leading bigrams for each question type in our annotated data. We observe that for Background questions that start with "what", the bigrams are more scattered with multiple combinations, and "how is/are/was/were/do" etc. appear more often in Elaboration than in Background questions.

A.3 Data Selection for Pairwise Ranking Classifier

Annotators may make the same or completely different choices, and two examples of annotator's ranking choices are shown in Table 9.

Algorithm 1 shows how training data is produced for the pairwise ranking classifier. The training instances are the combination of (a) a question q_i from q_{rel} and a question q_j from q_{nrel} and (line 2-6 in Algorithm 1) (b) two questions q_i and q_j from q_{rel} if and only if the two questions are separated by ≥ 2 ranks (line 8-16 in Algorithm 1).

Algorithm 1 Data selection for pairwise ranker

Input: $Q = \{q_{rel},\ q_{nrel}\}$, where Q is the total set of generated questions for an instance, q_{rel} is the set of relevant questions where $q_{rel} = \{(r_1, q_1), \cdots, (r_n, q_n)\}$, q_{nrel} is the set of non-relevant questions, and r_j is the rank for question q_j.

    ▷ Find relevant vs. non-relevant
1: **for** $q_j \in q_{rel}$ **do**
2:     **for** $q_k \in q_{nrel}$ **do**
3:         **yield** (q_j, q_k)
4:     **end for**
5: **end for**
6:
    ▷ Find questions with rank difference ≥ 2
7: **for** $j = 1, \cdots, n$ **do**
8:     $k \leftarrow j + 2$
9:     **while** $k \leq n$ **do**
10:         **if** $r_k - r_j \geq 2$ **then**
11:             **yield** (q_j, q_k)
12:         **end if**
13:         $k \leftarrow k + 1$
14:     **end while**
15: **end for**

A.4 Controllability Evaluation

We generate test set questions with six question types except "Other", and then classify the generations with our question type classifier. The test accuracy is shown in Table 10, with confusion matrix shown in Figure 1. As the largest number in each row/column is along the diagonal (aside from forward-looking questions, which the classifier never predicts in this set), the model and classifier are in alignment a significant fraction of the time. We also observe that Explanation is doing well in both precision and recall, Elaboration and Background are tricky to discriminate from each other, and Definition and Instantiation are being classified with high precision though not with very high recall. When the model is asked to generate a forward-looking question, the classifier labels it as Elaboration or Background in most cases. This is likely because there are very few forward-looking questions in the training data.

A.5 Additional Results

All results in Table 11 and Table 12 are averaged over 5 different runs with standard deviations.

Table 11 reports BLEU scores for 1/2/3/4-grams.

Explanation	Elaboration	Background	Definition	Instantiation	Forward-looking	Other
why is(87)	what is(39)	who is(20)	what is(53)	what are(24)	how will(5)	why is(4)
why did(75)	what are(31)	how long(19)	what are(17)	who are(16)	what will(2)	does this(3)
why was(51)	how did(27)	what was(17)	what does(15)	who is(8)	when will(2)	is it(2)
why are(50)	what does(17)	how did(16)	what do(6)	what other(5)	will the(2)	of which(2)
why were(24)	how is(12)	what is(14)	how is(2)	what kind(5)	what would(2)	yes what(1)
why would(21)	how do(12)	is that(10)	definition of(2)	what types(4)	what is(2)	what year(1)
why do(19)	how would(9)	how much(10)	is that(1)	which other(4)	what are(1)	does seaweed(1)
why does(18)	how are(9)	what are(9)	what 's(1)	what were(3)	how is(1)	does taxing(1)
why has(9)	how does(9)	is this(9)	does n't(1)	what sort(3)	were they(1)	what was(1)
what caused(7)	how many(8)	how many(8)	does note(1)	which companies(3)	is the(1)	this sounds(1)
what is(7)	what was(7)	is it(7)	does opportunities(1)	which year(3)	did they(1)	they must(1)
why will(7)	what kind(7)	are they(7)	i would(1)	what is(3)	how does(1)	there is(1)
is there(4)	how was(7)	where is(7)	does this(1)	which monday(2)	was their(1)	who is(1)
what makes(4)	what would(6)	how was(6)	what was(1)	what type(2)	did it(1)	what brass(1)
why have(3)	how much(6)	what did(6)	it means(1)	which officials(2)	so that(1)	should they(1)
what was(3)	how will(5)	how does(5)	who were(1)	in which(2)	how would(1)	which year(1)
why should(3)	how long(5)	where did(5)	what comprises(1)	who were(2)	what happened(1)	how many(1)
what were(2)	what makes(5)	what do(5)	thrift industry(1)	which countries(2)	would there(1)	is this(1)
why only(2)	how were(4)	why did(5)	how do(1)	which scientists(2)	would it(1)	is he(1)
why could(2)	in what(4)	when did(5)	who are(1)	which states(2)	will it(1)	which dollar(1)

Table 8: Most common leading bigrams in annotated questions (lowercased) for each type (counts shown in parentheses).

[*context*][*source* with span in **bold**]	[NO_CONTEXT][MILWAUKEE-The electric barrier on the **Chicago Sanitary and Ship Canal** that is considered the last line of defense to stop an Asian carp invasion of Lake Michigan has a problem : Fish can swim through it.]	[LOS ANGELES-Little-known fact : When it comes to extracting oxygen from the air we breathe, we humans are just OK. Birds are more efficient breathers than we are. So are alligators and, according to a new study, monitor lizards, and probably most dinosaurs were as well.][Humans are what are called **tidal breathers**.]
Definition	What is Chicago Sanitary and Ship Canal?	what is a tidal breather?
Background	where is that?	Are they considered \"tidal breathers\"?
Instantiation	Which section of the canal?	Who are these people?
Explanation	Why is this a problem?	Why are humans tidal breathers?
Forward	where is this?	How did they come up with this term?
Elaboration	What is the name of the canal?	Are they not?
Annotator A	1. Forward 2. Explanation 3. Background	1. Definition 2. Explanation 3. Forward
Annotator B	1. Definition 2. Instantiation 3. Elaboration	1. Definition 2. Explanation 3. Forward

Table 9: Examples of different ranking choices of expert annotators.

Question Type	% Acc
Explanation	97.82
Elaboration	65.84
Background	48.91
Definition	54.85
Instantiation	50.23
Forward-looking	0.

Table 10: Test accuracy for question type prediction for model generation of different question types.

While BASE always scores lowest and TYPE$_o$ is always highest, SPAN is second-highest for BLEU-1, BLUE-2 and BLEU-3[13], and beat by TYPE$_r$ for BLEU-4.

Table 12 reports all the metric scores that are specifically implemented by Ko et al. (2020). We see that TYPE$_r$ has lowest scores for *Train-n*. For *Article-n*, the model order is changed when n is varied, e.g., BASE is higher than TYPE$_r$ on *Article-1* but lower on *Article-2* and *Article-3*. Nevertheless, TYPE$_s$ is always lower than HUMAN on *Article-n*, and other models are always higher than scores of HUMAN.

A.6 Additional Examples

Table 13 lists more annotated examples for each question type, and Table 14 includes examples (gold and generated questions by our models) that are judged ungrammatical by annotators.

---

[13]The difference between SPAN and TYPE$_r$ is too small in BLEU-3 to be shown in the table.

Model	%BLEU-1	%BLEU-2	%BLEU-3	%BLEU-4	%METEOR	%ROUGE-L	%F$_{\text{BERT}}$	GPT2 ppl	Entropy
BASE	$26.9_{\pm 0.2}$	$12.0_{\pm 0.3}$	$6.8_{\pm 0.2}$	$4.3_{\pm 0.2}$	$11.8_{\pm 0.3}$	$27.4_{\pm 0.3}$	$39.6_{\pm 0.5}$	$119_{\pm 25}$	$0.699_{\pm 0.015}$
SPAN	$35.1_{\pm 0.9}$	$19.4_{\pm 0.5}$	$12.4_{\pm 0.4}$	$8.5_{\pm 0.4}$	$17.5_{\pm 0.7}$	$36.1_{\pm 0.5}$	$47.6_{\pm 0.4}$	$148_{\pm 10}$	$0.726_{\pm 0.062}$
TYPE$_s$	$28.9_{\pm 1.1}$	$14.6_{\pm 0.7}$	$8.7_{\pm 0.6}$	$5.7_{\pm 0.5}$	$13.6_{\pm 0.5}$	$30.9_{\pm 0.3}$	$41.6_{\pm 0.5}$	$219_{\pm 18}$	$0.823_{\pm 0.024}$
TYPE$_r$	$33.4_{\pm 1.4}$	$18.9_{\pm 1.0}$	$12.4_{\pm 0.8}$	$8.6_{\pm 0.6}$	$18.3_{\pm 0.4}$	$35.3_{\pm 0.7}$	$47.4_{\pm 0.8}$	$89_{\pm 7}$	$0.612_{\pm 0.025}$
TYPE$_o$	$37.7_{\pm 1.0}$	$21.6_{\pm 0.8}$	$14.0_{\pm 0.8}$	$9.7_{\pm 0.8}$	$19.5_{\pm 0.4}$	$39.1_{\pm 0.4}$	$50.1_{\pm 0.5}$	$154_{\pm 18}$	$0.751_{\pm 0.008}$

Table 11: Automatic metrics on our test set for our models.

	Train-2	Train-3	Train-4	Article-1	Article-2	Article-3	Span
Human	0.467	0.203	0.059	0.386	0.126	0.064	0.354
BASE	$0.518_{\pm 0.018}$	$0.267_{\pm 0.019}$	$0.097_{\pm 0.009}$	$0.469_{\pm 0.020}$	$0.186_{\pm 0.020}$	$0.104_{\pm 0.018}$	$0.184_{\pm 0.007}$
SPAN	$0.505_{\pm 0.015}$	$0.246_{\pm 0.020}$	$0.079_{\pm 0.012}$	$0.455_{\pm 0.025}$	$0.182_{\pm 0.022}$	$0.101_{\pm 0.019}$	$0.452_{\pm 0.029}$
TYPE$_s$	$0.530_{\pm 0.006}$	$0.288_{\pm 0.012}$	$0.102_{\pm 0.012}$	$0.315_{\pm 0.015}$	$0.090_{\pm 0.010}$	$0.041_{\pm 0.006}$	$0.346_{\pm 0.023}$
TYPE$_r$	$0.473_{\pm 0.013}$	$0.218_{\pm 0.015}$	$0.068_{\pm 0.010}$	$0.445_{\pm 0.018}$	$0.195_{\pm 0.016}$	$0.112_{\pm 0.013}$	$0.542_{\pm 0.030}$
TYPE$_o$	$0.488_{\pm 0.011}$	$0.233_{\pm 0.012}$	$0.073_{\pm 0.004}$	$0.401_{\pm 0.020}$	$0.149_{\pm 0.016}$	$0.078_{\pm 0.012}$	$0.475_{\pm 0.024}$

Table 12: Metric scores from Ko et al. (2020) that measure the extent of copying content from the training partition, articles, and spans in the source sentences to the generated questions. All scores are reported on our test set.
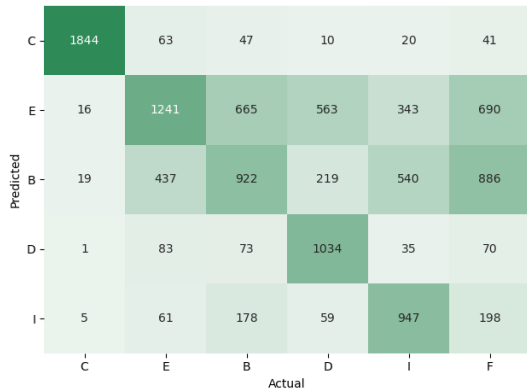

Figure 1: Heatmap showing confusion matrix for type controllability evaluation. The "Actual" type is the desired type passed as control code to the model, and the "Predicted" type is the output of running the question type classifier on the generated question. C: Explanation (causal), E: Elaboration, B: Background, D: Definition, I: Instantiation, F: Forward-looking.

Question Type (# samples)	Example	
	[*context*] [*source* with span in **bold**]	Question
Explanation (443)	[. . . Osip Nikiforov is recording Chopin's Etude Op. 10, No. 1, without capturing any of its sound.] [**Instead,** a sensor-equipped piano is recording the "data" of his performance . . . .]	Why is there a sensor-equipped piano recording data of his performance?
Elaboration (364)	[NO_CONTEXT][Miami Shores, Fla., tech **consultant** Rudo Boothe, age 33, attributes his professional success . . . .]	For what company?
	[NO_CONTEXT][The Agriculture Department says Americans seem to be eating a bit more each year but are **choosier** about what's on the menu.]	what are they choosing?
	[One of Ronald Reagan's attributes as President was that he rarely gave his blessing to the claptrap . . . .] [In fact, he liberated the U.S. from one of the world's most **corrupt** organizations – UNESCO.]	How is UNESCO corrupt?
Background (407)	[NO_CONTEXT][. . . a young man and his mentor practice bull-fighting techniques under the **light** of an atrium.]	Are they practicing at night?
Definition (114)	[NO_CONTEXT][LOS ANGELES - The booming illegal international wildlife trade forced **conservationists** to do the unthinkable Tuesday . . . .]	Who were the conservationists?
	[People start their own businesses for many reasons. But a chance to fill out sales - tax records is rarely one of them.] [Red tape is the **bugaboo** of small business.]	what is a bugaboo?
Instantiation (159)	[The Bush administration's nomination of Clarence Thomas to a seat on the federal appeals court here received a blow this week . . .] [People familiar with the Senate Judiciary Committee, . . . , said some **liberal members** of the panel are likely to question the ABA rating in hearings on the matter.]	Which liberal members are likely to question the ABA ratings?
Forward-looking (31)	[Bethlehem Steel Corp. has agreed in principle to form a joint venture with the world's second-largest steelmaker . . . .] [The entire division employs about **850 workers**.]	How will they need to increase or decrease staff?
Other (32)	[. . . there's one easy way to make a July beach vacation even better than expected: Add seaweed . . .] [. . . his back covered in what looked like strands of **chartreuse cotton candy,** the 7-year-old Beijing boy was having the time of his life Sunday . . .]	Does seaweed look like cotton candy?

Table 13: Annotated question type distributions and salient examples of each question type. Context and source sentences are presented where the spans in source sentences are bold.

HUMAN	why would it do that? is it the aha? in which year? WHAT COUTRIES RECIEVED LOANS? What specifically are the unhappy about with the direction?
BASE	What goal does everyone have? What happened that they didn't agree? What kind of violence? what are these signs? What was Andrew Coltart doing at 69?
SPAN	Why weren't the details unavailable? Why is there a hard time posting an upset over Germany? What is their goal in common? Which lawmakers and others arguing? How did they inflating the stock price?
TYPE$_s$	which meetings? What meetings? What are the details about this other than that? What details? What goal? What goal? what were they? what prefecture?
TYPE$_r$	Who are the Serbs from Croatia and Bosnian Muslims opposed to the Bosnian government? Why would NATO take in Poland, Hungary and others asMembers? How does Dominican authorities know the whereabouts of the banker and two Dominicans? How does a report about AIDS come to a conclusion? Why is this symbol of America?
TYPE$_o$	How many peacekeepers? How was anreement to conceal the agreement made? Did these talks involve a lot of talks? How long has the explosion been taking place? What are terms and syndicate manager?

Table 14: Examples of gold questions from INQUISITIVE and questions generated by models that are judged as ungrammatical by annotators.