

# Automatically Generating Cause-and-Effect Questions from Passages

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, Marti A. Hearst

UC Berkeley

{katie.stasaski, manav.rathod, tonytu16,  
emily.xiao, hearst}@berkeley.edu

## Abstract

Automated question generation has the potential to greatly aid in education applications, such as online study aids to check understanding of readings. The state-of-the-art in neural question generation has advanced greatly, due in part to the availability of large datasets of question-answer pairs. However, the questions generated are often surface-level and not challenging for a human to answer. To develop more challenging questions, we propose the novel task of cause-and-effect question generation. We build a pipeline that extracts causal relations from passages of input text, and feeds these as input to a state-of-the-art neural question generator. The extractor is based on prior work that classifies causal relations by linguistic category (Cao et al., 2016; Altenberg, 1984). This work results in a new, publicly available collection of cause-and-effect questions. We evaluate via both automatic and manual metrics and find performance improves for both question generation and question answering when we utilize a small auxiliary data source of cause-and-effect questions for fine-tuning. Our approach can be easily applied to generate cause-and-effect questions from other text collections and educational material, allowing for adaptable large-scale generation of cause-and-effect questions.

## 1 Introduction

Automated question generation (QG) can have a large educational impact since questions can be generated to check learners’ comprehension and understanding of textbooks or other reading materials (Thalheimer, 2003; Kurdi et al., 2020). A high-quality QG system could reduce the costly human effort required to generate questions as well as free up teachers’ time to focus on other instructional activities (Kurdi et al., 2020). Furthermore, a robust QG system could also expand the variety of

<b>Original Passage</b>	Injuries can also be prevented by proper rest and recovery. If you do not get enough rest, your body will become injured and will not react well to exercise, or improve. You can also rest by doing a different activity.
<b>Extracted Cause</b>	you do not get enough rest
<b>Extracted Effect</b>	your body will become injured and will not react well to exercise, or improve
<b>Generated Cause Q</b>	why will your body become injured and not react well to exercise?
<b>Generated Effect Q</b>	what happens if you don’t get enough rest?

Table 1: Example passage (taken from the TQA dataset), extracted cause/effect, and generated questions. The Extracted Cause is the intended answer for the Generated Cause Question; similarly for Effect.

educational material used for formative assessment, allowing students more opportunities to cement their understanding of concepts.

To be truly effective, automated question generation for education must have the ability to ask questions for students at different levels of development. A frequently used measure of question difficulty is Bloom’s taxonomy (Bloom et al., 1956; Anderson et al., 2001), which defines a framework of how to assess types of questions across different levels of mastery, progressing from simplest to most complex. Factual questions, which involve recalling information, fall on the lowest level (Recall) (Beatty Jr, 1975; Anderson et al., 2001). By contrast, cause and effect questions are categorized at level 6 – Analysis – according to the original Bloom’s taxonomy (Beatty Jr, 1975) or level 2 –

Understanding – in the commonly used revised model (Anderson et al., 2001).

The rise of large question answering datasets, such as SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), and HotPotQA (Yang et al., 2018), have created the ability to train well-performing neural QG systems. However, due to the nature of their training data, current question generation systems mainly generate factual questions characteristic of Bloom’s level 1.

Questions which test knowledge of a causal relation can assess a deeper level of mastery beyond surface-level factual questions. Altenberg (1984) states that “[a] causal relation can be said to exist between two events or states of affairs if one is understood as the cause of or reason for the other.” We aim to generate cause-and-effect questions, which test knowledge of the relationship between these two events or states. An example of our generated questions along with corresponding input, cause, and effect can be seen in Table 1.

To address this task, we propose a novel pipeline to generate and evaluate cause-and-effect questions directly from text. We improve upon a pre-existing causal extraction system (Cao et al., 2016), which uses a series of syntactic rules to extract causes and effects from unstructured text. We utilize the resulting *cause* and *effect* as *intended answers* for a neural question generation system. For each cause and effect, we generate one question, to test each direction of the causal relationship.

Our work sets the stage for *scalable* generation of cause-and-effect questions because it automatically generates causal questions and answers from freeform text. In this paper, we evaluate our approach on two English datasets: SQuAD Wikipedia articles (Rajpurkar et al., 2018) and middle school science textbooks in the Textbook Question Answering (TQA) dataset (Kembhavi et al., 2017).

Our research contributions include:

- A novel cause-and-effect question generation pipeline, including an improved causal extraction system based on a linguistic typology (Cao et al., 2016; Altenberg, 1984),
- An evaluation framework, accompanied by preliminary experimental results showing that fine-tuning on a small, auxiliary dataset of cause-and-effect questions substantially improves both question generation and question answering models,
- A novel collection of 8,808 cause-and-effect

questions, with open source code to apply the pipeline to other text collections, allowing for future work to examine the educational impact of automatically-generated cause-and-effect questions.<sup>1</sup>

## 2 Related Work

In this section, we discuss past work in causal extraction, question answering, question generation, and applications of generated questions.

### 2.1 Causal Extraction

Causal extraction systems aim to identify whether a causal relation is expressed in text and to identify the cause and effect if so. However, the use of neural techniques in causal extraction is sparse and thus most of the work is still tied to a focus on extracting relations based on specific linguistic features (Asghar, 2016). We utilize Cao et al. (2016), which is aimed at extracting causal relations from academic papers via a series of structured syntactic patterns tied to a linguistic typology (Altenberg, 1984).

Causal relation extraction has been applied to inform question answering models (Girju, 2003; Breja and Jain, 2020). Some work has used neural networks to generate explanations from open-domain “why” questions without using external knowledge sources (Nie et al., 2019).

### 2.2 Question Answering

Neural question answering (QA) is a widely-explored area with many large datasets of crowdworker-created questions. SQuAD (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2017) each include over 100,000 crowdworker-created questions from Wikipedia and CNN articles, respectively. NarrativeQA includes questions aimed at larger narrative events, which require reading an entire novel or movie script in order to answer (Kočíský et al., 2018). However, cause-and-effect questions are infrequent (“why” questions compose 9.78% of the dataset), and questions are paired with long documents.

HotPotQA includes over 100,000 questions spanning multiple passages, where questions require combining multiple facts in order to correctly answer the question (Yang et al., 2018). However, HotPotQA is still primarily a factual recall task. For instance, our inspection of the dataset finds

<sup>1</sup><https://github.com/kstats/CausalQG>

that 30% of the questions expect a Person entity as an answer. The three most common question types are “what,” “which,” and “who,” which suggest a factual, information lookup style answer. While HotPotQA questions are potentially more difficult for machines to answer than questions from other QA datasets, the resulting questions are overwhelmingly factual and include entities as the intended answer.

Question answering approaches trained on these datasets include gated attention-based methods (Wang et al., 2017b) as well as transformer-based methods (Yang et al., 2019; Lan et al., 2020). Models which augment QA data with automatically-generated questions have also been explored (Duan et al., 2017).

### 2.3 Question Generation

Past work has utilized purely syntactic cues to generate questions (Heilman and Smith, 2010). However, these systems rely on rules which may be brittle. More recent work has combined a syntactic question generator with backtranslation, to improve robustness and reduce grammatical errors of generated questions (Dhole and Manning, 2020). While Syn-QG is able to reliably generate types of causal questions using two specific patterns, the system is limited in diversity of question wording by syntactic rules.

Question answering datasets have been used to train neural models to generate questions directly from input text. Question generation approaches include specialized attention mechanisms (Zhao et al., 2018) as well as generating via large transformer models (Qi et al., 2020; Chan and Fan, 2019). Jointly training QA and QG models have also been explored (Wang et al., 2017a). Past work has also trained neural QG systems on questions generated via a rule-based system (De Kuthy et al., 2020).

### 2.4 Question Generation Applications

Past work has explored educational applications of question generation (Kurdi et al., 2020). QG-Net utilizes a pointer-generator model trained on SQuAD to automatically generate questions from textbooks (Wang et al., 2018). Additional work has aimed at generating educational questions from a structured ontology (Stasaski and Hearst, 2017). QuizBot exposes students to questions via a dialogue interface, where pre-set questions are chosen in an intelligent order (Ruan et al., 2019).

## 3 Cause-and-Effect Question Generation Pipeline

We propose a novel pipeline which combines a causal extractor with a neural question generation system to produce cause-and-effect questions. The entire pipeline can be seen in Figure 1.

### 3.1 Datasets

Our pipeline can take in freeform text and automatically extract a cause and effect to generate questions. We feed passages from the SQuAD 2.0 development set (Rajpurkar et al., 2018) and the Textbook Question Answering (Kembhavi et al., 2017) datasets into our causal extraction pipeline for evaluation experiments.

SQuAD 2.0 consists of over 100,000 questions from over 500 Wikipedia articles. This dataset is standard to train both QA and QG systems; we do not use the question portion of this dataset because it consists primarily of straightforward factual questions. A heuristic to determine the proportion of cause-and-effect questions in SQuAD is the number of questions in the dataset which begin with “why.” While this does not capture all possible ways of expressing causality, this can serve as a signal for the prevalence of cause-and-effect questions in each dataset. In SQuAD, only 1.3% of questions begin with “why,” indicating cause-and-effect questions are not a significant component of this dataset. A more extensive analysis which examines hand-labeled question n-grams finds similar results (Appendix A).

TQA consists of 26,260 questions from Life Science, Earth Science and Physical Science textbooks. We choose this dataset because it contains educational textbook text, which often express causal relationships; however, we do not use the TQA questions since many are tied to an entire lesson in the textbook and include visual diagrams.

### 3.2 Causal Extraction

We use and improve a pre-existing causal extractor to identify cause and effect pairs in unstructured input text, see Causal Extractor in Figure 1 (Cao et al., 2016).<sup>2</sup> The system achieves approximately 0.85 recall of hand-labeled cause-and-effect relationships over 3 academic articles, as reported in Cao et al. (2016). The extractor relies on a series of hand-crafted patterns based on syntax cues and

<sup>2</sup>[https://github.com/Angela7126/CE\\_extractor--Patterns-Based](https://github.com/Angela7126/CE_extractor--Patterns-Based)

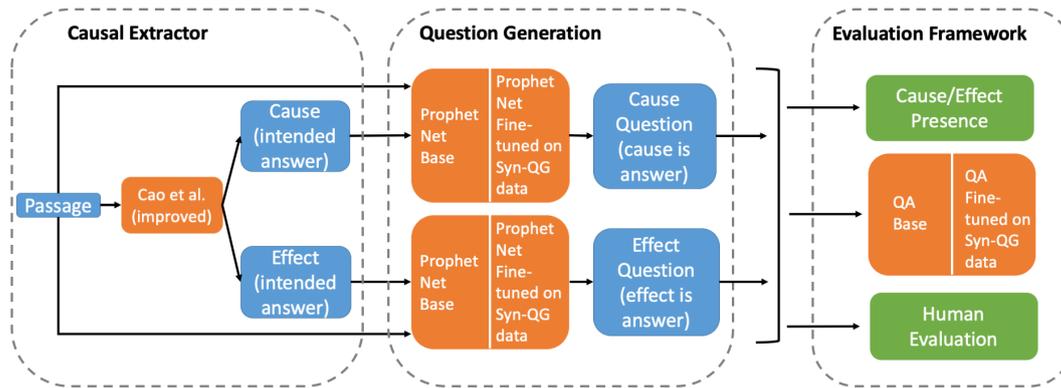


Figure 1: Causal extraction and question generation pipeline. Input passage is passed to the causal extractor, which identifies a cause and effect. The cause and effect are each passed into the Question Generator as intended answers, to generate a resulting question. Questions are evaluated using automatic and human metrics. Example passages and questions can be seen in in Table 1.

part-of-speech tags. An example pattern, where  $\&$  indicates a class of tokens defined by Cao et al. (2016),  $\&C$  is the cause,  $\&R$  is the effect, and optional arguments are shown in parentheses is:

$\&C (,;/./-)$  (&AND) as a (&ADJ) result  
 $(.)$  &R

A match to this pattern from the TQA dataset (**cause bolded**, *effect italicized*, and other matched terms underlined) is:

**Unsaturated fatty acids have at least one double bond between carbon atoms.** *As a result, some carbon atoms are not bonded to as many hydrogen atoms as possible.* They are unsaturated with hydrogens.

We pass 2 or 3 sentence passages into the causal extractor, since a cause and effect may span across adjacent sentences. We always try to select a 3-sentence passage, but when a paragraph boundary would be crossed, we reduce to 2-sentence passage. We use a sliding 3-sentence window to examine all possible sentence combinations. Multiple causal relationships may be found in a single passage; the sliding window captures the multiple relationships across different passages.

### 3.2.1 Adjustments to Cao et al.’s Extractor

After an examination of 160 extracted causal relationships from SQuAD and TQA, we implement targeted modifications to improve the quality of extracted causal relationships. We omit causal relationships which include direct reference to a figure in a textbook or are part of a question. We

additionally find that ambiguous causal linking phrases (“as,” “so,” and “since”) are prevalent but have lower accuracy than more direct phrases (“because”). Thus, we filter out “as,” “so,” and “since” patterns where “as,” “so,” or “since” is not labeled as a conjunction or subordinate conjunction by the part-of-speech tagger. Further details are described in Appendix B. This modification reduces the total number of extracted relations from 3,976 to 3,359 for TQA and from 1,105 to 1,045 for SQuAD. Evaluation of these changes appears in Section 4.

### 3.2.2 Causal Link Typology

One motivation for choosing the Cao et al. system is that each pattern is tied to a typology of causal links (Altenberg, 1984). The typology contains 4 main categories, based on the causal link which is utilized to express the causal relationship: Adverbial (“so”, “hence”, “therefore”), Prepositional (“because of”, “on account of”), Subordination (“because”, “as”, “since”), and Clause-integrated linkage (“that’s why”, “the result was”). Examples of causal relations extracted using each part of the typology, using the TQA dataset, can be seen in Table 2. By incorporating these patterns in our evaluation framework, we can examine potential gaps in model performance that might be tied to syntactic cues, allowing for the design of improved models.

### 3.3 Cause-and-Effect Question Generation

After extracting causes and effects from input text, we use each as an intended answer for a neural question generation system (see Question Generation in Figure 1). This results in two output questions

Type	Example
Adverbial	<b>Different energy levels in the cloud have different numbers of orbitals.</b> <i>Therefore, different energy levels have different maximum numbers of electrons.</i> Table 5.1 lists the number of orbitals and electrons for the first four energy levels.
Prepositional	The function of these scales is for protection against predators. <i>The shape of sharks teeth differ</i> according to <b>their diet</b> . Species that feed on mollusks and crustaceans have dense flattened teeth for crushing, those that feed on fish have needle-like teeth for gripping, and those that feed on larger prey, such as mammals, have pointed lower teeth for gripping and triangular upper teeth with serrated edges for cutting.
Subordination	The spoon particles started moving faster and became warmer, causing the temperature of the spoon to rise. <u>Because</u> <b>the coffee particles lost some of their kinetic energy to the spoon particles,</b> <i>the coffee particles started to move more slowly.</i> This caused the temperature of the coffee to fall.
Clause-Integration	<b>The stars outer layers spread out and cool.</b> <u>The result is</u> <i>a larger star that is cooler on the surface, and red in color.</i> Eventually a red giant burns up all of the helium in its core.

Table 2: Examples of causal relationships from different typology categories (Altenberg, 1984), from the TQA dataset, with causes bolded, effects italicized, and causal link words underlined.

for each causal relationship (one corresponding to cause, and one to effect). We use ProphetNet, a state-of-the-art question generation model, to generate these questions (Qi et al., 2020).<sup>3</sup> The novelty of ProphetNet is its ability to generate text by predicting the next n-gram instead of just the next token, which helps to prevent overfitting on strong local correlations. ProphetNet is fine-tuned for question generation using SQuAD 1.1 (Rajpurkar et al., 2016). Models specifications are described in Section 5, and model evaluation is described in Sections 6 and 7.

#### 4 Evaluating Extracted Causal Relations

In order to ensure the extracted causes and effects are suitable to generate questions from, we evaluated the original Cao et al. (2016) and our improved causal extractor for accuracy via a crowdsourcing task. Workers evaluated if an extracted cause and effect were causal or not, judged from a passage with the extracted cause and effect highlighted (see Appendix C).

We utilized Amazon Mechanical Turk to recruit crowdworkers for data labeling. We required crowdworkers to have a 98% HIT approval rating, have completed at least 5,000 past HITs, and be located in the United States. We estimated the task would require no more than 10 minutes to complete; therefore, we paid \$1.66, equivalent to \$10

<sup>3</sup><https://github.com/microsoft/ProphetNet>

Type	TQA	SQuAD
Adverbial	87.1%	84.6%
Prepositional	92.3%	72.7%
Subordination	77.1%	75.7%
Clause-Int	83.3%	86.7%

Table 3: Percent of extractions labeled as causal by crowdworkers, for samples of 100 each from TQA and SQuAD datasets, by linguistic category, for our improved system.

per hour. We asked crowdworkers to provide a one-sentence explanation for their classification decision for the first and last item in the HIT, which we manually checked to ensure quality. To increase confidence in the labels and to tiebreak disagreements, we acquired 5 labels for every passage.

We sampled 100 passages from TQA and 100 from SQuAD for both the original causal extraction system and our improved system. We used proportionate stratified sampling where the size of sample typology group is proportional to the size of population typology group, while also ensuring that all typology groups are represented in the sample.

After tiebreaking using majority vote, for the original Cao et al. (2016) causal extraction system we find an overall 70% of TQA and 68% of SQuAD are rated as causal. In comparison, for our improved system, we find an overall 83% of TQA and 79% of SQuAD are rated as causal. Results segmented by typology for our improved system can

be seen in Table 3 (results for the original Cao et al. (2016) system can be seen in Appendix C). While the accuracies are higher for TQA than SQuAD, all accuracies range from 72 to 92%. This provides evidence that the extractor is able to reliably identify causes and effects, which we can utilize as intended answers for the downstream question generation system.

Fleiss’s kappa (Fleiss et al., 1971) however, is only slight to fair for this evaluation (0.21 for base extractor and 0.10 for our improved extractor). This could be due to a high prior probability of an extracted cause and effect being causal (62% of questions have 4 or 5 annotators in agreement) (Falotico and Quatto, 2015). However, future work should additionally investigate alternatives to this task to achieve higher agreement. Overall, the high accuracy of our improved extractor provides evidence that we can reliably extract a cause and effect to pass into a question generation system.

## 5 QG and QA Models

This section describes the design of both the question generation (QG) and question answering (QA) models, which occur in the Question Generation and Evaluation Framework components of Figure 1, respectively. We describe them together because they share an auxiliary data source. The QG models are evaluated in Sections 6 (automated) and 7 (manual).

To generate cause-and-effect questions where the intended answers are the causes and effects from our improved extractor, we train and evaluate two question generation models based on ProphetNet. We also make use of question answering models to assess the quality of the generated questions. We choose a transformer-based QA model from the Huggingface Transformers library (Wolf et al., 2020) which is BERT-large fine-tuned on SQuAD 2.0 with whole-word masking.<sup>4</sup> F1 performance for this model on the original SQuAD 2.0 test set is 0.93.

For both QA and QG, we utilize an additional data source for fine-tuning, to examine the effect of augmenting the models with additional cause-and-effect question data. We use a syntactic question generation system combined with backtranslation, Syn-QG<sup>5</sup> (Dhole and Manning, 2020), to generate

<sup>4</sup><https://huggingface.co/deepset/bert-large-uncased-whole-word-masking-squad2>

<sup>5</sup><https://bitbucket.org/kaustubhdhole/syn-qg>

questions and answers for fine-tuning. We limit Syn-QG to two patterns, Purpose (PNC and PRP) and Cause (CAU), which generate cause-and-effect questions. We input three-sentence passages which our system has identified as including a cause and effect to Syn-QG, resulting in 2,082 questions from TQA and 1,753 from SQuAD. While the wording and syntactic structure of Syn-QG questions lack diversity, we hypothesize that ProphetNet will benefit from fine-tuning on a set of cause-and-effect questions as this question type was not well-represented in the SQuAD training set.

Additionally, while Syn-QG uses its own syntactic patterns to generate cause-and-effect questions directly from text, we observe their system is not able to cover all causal relationships included in Cao et al. (2016). From randomly-sampled passages which our improved extractor has identified as including a cause and effect, Syn-QG is able to generate cause-and-effect questions for 309 out of 500 passages for SQuAD and 233 out of 500 passages for TQA.

We compare two models for QG and QA experiments in Sections 6 and 7: Base (with no fine-tuning) and Syn-QG fine-tuned to explore the effect of fine-tuning QG and QA models on a small auxiliary dataset of cause-and-effect questions. For all experiments, we split the Syn-QG auxiliary dataset into 80% training and 20% test data. The QG models’ resulting questions are fed as input into the QA models (see Figure 1). Additional model specifications are in Appendix D. Table 4 contains generated questions for each typology category for both the base and the Syn-QG fine-tuned QG models.

## 6 Automated Evaluation of Generated Questions

As part of our Evaluation Framework, we develop two automated metrics to evaluate the generated cause-and-effect questions: (i) cause/effect presence, which measures whether the cause or effect is present in the question, and (ii) QA system performance, which measures whether a question answering system can answer the generated question.

### 6.1 Cause/Effect Presence

Because causal relationships contain both a cause and an effect, we assume that a question which assesses understanding of this relationship would include a direct mention of one of the two. For instance, a question which has the cause as the

Category	Passage	ProphetNet Base		ProphetNet Syn-QG	
		Cause Question	Effect Question	Cause Question	Effect Question
Adv	Myopia is corrected with a concave lens, which curves inward like the inside of a bowl. <b>The lens changes the focus</b> , so <i>images fall on the retina as they should</i> . Generally, nearsightedness first occurs in school-age children.	what causes images to fall on the retina as they should?	what happens when the concave lens changes the focus?	why do images fall on the retina as they should?	why does the lens change the focus?
Prep	Their activity at night. <i>An elongated, toothed snout used for slashing the fish that they eat</i> , as seen in sawsharks. Teeth used for grasping and crushing shellfish, a characteristic of bullhead sharks.	what is the snout used for?	what is used for slashing the fish that they eat?	for what purpose is an elongated, toothed snout used?	what is used for slashing the fish that they eat?
Sub	The droplets gather in clouds, which are blown about the globe by wind. As <b>the water droplets in the clouds collide and grow</b> , <i>they fall from the sky as precipitation</i> . Precipitation can be rain, sleet, hail, or snow.	how do droplets fall from the sky as precipitation?	what happens when water droplets in the clouds collide and grow?	why do they fall from the sky as precipitation?	what happens as the water droplets in the clouds collide and grow?
C-I	<b>Some mixtures are heterogeneous</b> . This means <i>they vary in their composition</i> . An example is trail mix.	what means some mixtures vary in their composition?	some mixtures are heterogeneous in what way?	why do some mixtures vary in their composition?	some mixtures are heterogeneous in their composition?

Table 4: Randomly sampled generated question from each typology category from TQA, with corresponding base and Syn-QG questions. Passages include causes in **bold** (which are the intended answers for Cause Questions) and effects *italicized* (which are the intended answers for Effect Questions).

intended answer would contain the effect in the question text, and vice versa. Thus, we propose the Cause/Effect Presence metric: for questions where the cause is the intended answer, we measure the recall of the words in the extracted effect present in the question. Likewise, for questions where the effect is the intended answer, we measure the recall of the words in the extracted cause present in the question. Because the question could contain a subset of words expected for the cause/effect, we measure this in terms of recall. Furthermore, a question is necessarily going to have additional words, such as the opening phrase, e.g. “why,” which we do not want to penalize. Formally defined:

$$Recall_{cause} = \frac{|Q_e \cap C|}{|C|}$$

$$Recall_{effect} = \frac{|Q_c \cap E|}{|E|}$$

where  $C$  is a bag of all words in the cause extracted from the passage,  $E$  is a bag of all words in the effect extracted from the passage,  $Q_e$  is a bag of all words in the effect question, and  $Q_c$  is a bag of all words in the cause question. For words that appear  $n$  times in an extracted cause or effect, we

give credit for each appearance in the question, up to  $n$  times.

Results for Cause/Effect Presence on SQuAD and TQA can be seen in Table 5. Results stratified by typology for the best performing model (fine-tuned Syn-QG) can be seen in Appendix E.

We note that after fine-tuning ProphetNet on Syn-QG data, performance increases from 0.55 to 0.72 for TQA and 0.37 to 0.56 for SQuAD. This is somewhat expected, as Syn-QG’s syntax rules result in questions that are similarly-structured and include the cause or effect. We also note slightly higher scores when an effect is the intended answer. Future work should isolate whether this is due to the model’s ability to generate better questions from effects or whether the extractor is better at extracting effects than causes from passages. In Section 7, we explore this further by having humans label the quality of the QG model’s output.

## 6.2 Question Answering System Performance

Following past work on evaluating QG models (Yuan et al., 2017), we measure whether a QA system is able to accurately answer the generated cause-and-effect questions from the passage. If the QG model produces an ill-formed or incorrect ques-

Model	TQA			SQuAD		
	C	E	T	C	E	T
Base	0.52	0.57	0.55	0.32	0.42	0.37
Syn-QG	0.71	0.72	0.72	0.54	0.57	0.56

Table 5: Average cause and effect recall in questions generated by ProphetNet Base and fine-tuned on Syn-QG (out of 1.0, higher is better). C indicates questions where a cause is the intended answer, E indicates questions where the effect is the intended answer, and T indicates the total for both.

QG	QA	Textbook			SQuAD		
		C	E	T	C	E	T
Base	Base	0.24	0.18	0.21	0.22	0.15	0.18
	Syn-QG	0.60	0.49	0.54	0.60	0.46	0.53
Syn-QG	Base	0.24	0.16	0.20	0.19	0.13	0.16
	Syn-QG	0.57	0.48	0.53	0.58	0.43	0.51

Table 6: F1 results for base and fine-tuned QA on questions generated from base and fine-tuned ProphetNet. C indicates questions where a cause is the intended answer, E indicates questions where the effect is the intended answer, and T indicates the total between both.

tion, the QA model will be less likely to produce the correct answer. The QA model must answer a cause question with an effect, and an effect question with a cause. We report the QA model’s F1 performance on the set of cause-and-effect questions in Table 6.

Fine-tuning the QA model with the Syn-QG data provides a large F1 improvement over the base model (0.21 to 0.54 for TQA, 0.18 to 0.53 for SQuAD), showing the benefit of fine-tuning on an auxiliary cause-and-effect dataset. However, the QA models do not perform as well at answering cause-and-effect questions as they do at answering factual ones; we see a 0.4-0.7 drop in F1 values from factual to causal questions. Future work can explore using our dataset to further improve QA performance on cause-and-effect questions.

Results stratified by typology for the best-performing QG and QA pair (both fine-tuned on Syn-QG) can be seen in Table 7. The lowest-performing category for both TQA and SQuAD is Adverbial. However, Adverbial was ranked the second-highest for correct causal links by crowdworkers (in Section 4). It is unexpected that one of the more reliable extraction types resulted in the lowest QA performance. Future work can explore why QA models are not adept at handling this linguistic phrasing and how to improve model performance in this category.

Type	F1	#TQA	F1	#SQuAD
Adv.	0.50	2,868	0.47	510
Prep.	0.57	940	0.56	550
Sub.	0.53	2,546	0.47	870
C-I	0.59	364	0.63	160

Table 7: F1 scores of best QG-QA pair (both fine-tuned on Syn-QG) broken down by typology.

## 7 Human Evaluation of Generated Questions

While automated metrics are scalable, they may not be as reliable as human evaluation. Thus, we conducted a crowdworker task to evaluate generated questions. To ensure we provide the model with high-quality causal relations, we evaluated questions generated from the extracted relations that were labeled as Causal in Section 4 (83 from TQA and 79 for SQuAD). We performed human evaluation for base ProphetNet, to establish a baseline human-rating of generated questions.

Crowdworkers were presented with a passage (with the intended answer highlighted) and a generated question (see Appendix C). Workers were asked to label (1) if the question is a *causal* question and (2) whether the answer highlighted in the passage correctly answers the question. If a question was too ill-formed or ungrammatical to provide these classifications, crowdworkers were asked to select a checkbox and provide a text justification. The conditions for this task were the same as in Section 4.

Results can be seen in Table 8. These high accuracy ratings (80%-90%) indicate ProphetNet is able to reliably generate questions which are correct for the intended answer and are valid cause-and-effect questions. For both tasks, questions generated from the TQA dataset were higher-ranked than those generated from SQuAD. For the task of determining whether a question was causal, there is not a consistent winner between the set of cause and effect questions. This indicates ProphetNet is consistently better at generating questions for both directions of the cause-and-effect relationship.

For the task of determining whether an intended answer is correct, when the effect is the intended answer, the proportion that are rated as correct is higher than for cause. This is consistent with the higher performance for effect questions on the Cause/Effect Presence metric.

Results stratified by typology can be seen in Table 9. For rating answer correctness, the Preposi-

Condition	% Causal	% Answer
TQA Cause	96%	90%
TQA Effect	98%	93%
<b>TQA Total</b>	<b>97%</b>	<b>92%</b>
SQuAD Cause	89%	90%
SQuAD Effect	82%	87%
<b>SQuAD Total</b>	<b>85%</b>	<b>89%</b>
Overall Cause	93%	90%
Overall Effect	90%	90%
<b>Overall Total</b>	<b>91%</b>	<b>90%</b>

Table 8: Percentage of Base ProphetNet generated questions (n=324) classified by crowdworkers as causal (% Causal) and matching the intended answer (% Answer), by dataset and whether the intended answer was cause or effect. “Total” indicates the combination of cause and effect questions and “Overall” indicates the combination of datasets.

Category	TQA			SQuAD		
	%C	%A	Tot	%C	%A	Tot
Adv.	97%	91%	70	84%	86%	44
Prep.	96%	83%	24	91%	84%	32
Sub.	96%	96%	52	82%	95%	56
C-I	100%	100%	20	85%	85%	26

Table 9: Percentage of crowdworking ratings indicating a cause-and-effect question (% C), a correct answer (% A), and the total number of questions evaluated segmented by typology category.

tional category was the lowest for both datasets, indicating a potential model shortcoming. The Subordination questions from SQuAD were the least causal over all categories.

In 22 cases out of 1,620, a crowdworker indicated that a generated question was too ill-formed to provide a rating. Fleiss’s kappa for determining whether a question was causal was 0.07 and for determining whether the answer was correct was 0.14, both slight agreement. However, this is likely due to the high prior probability distribution (Falotico and Quatto, 2015). For 74.1% of questions, 4 or 5 annotators all agreed on the same rating for determining whether a question was causal; likewise, 80.9% of the time 4 or 5 annotators all agreed for determining whether an answer is correct.

Overall, human ratings indicate that 80%-90% of the time, the generated question is both causal and correct for the intended answer.

## 8 Future Work

Our proposed pipeline is generalizable to a wide range of corpora, allowing for the generation of cause-and-effect questions from other educational texts at-scale. Human evaluation ratings indicate our model can reliably generate questions which

are both causal and correct for the intended answer. However, more work needs to be done to generate questions which sound more natural and are less reliant on the wording of the input passage. The types of grammar errors made should also be examined before assessing the educational benefits of the generated questions.

Additionally, our goal in generating cause-and-effect questions was to generate questions which fall higher on Bloom’s taxonomy. While our results indicate cause-and-effect questions are more challenging than straightforward factual questions for a QA model, further work should assess the difficulty of these questions in an educational setting.

The complexity of causal relationships can be further explored. We observe instances of overlapping causal relationships extracted from our system, such as the Subordination example in Table 2. This can be leveraged to generate questions which require knowledge of longer logic chains to answer, similar to past work which accomplished this (Stasaski and Hearst, 2017).

Future work can also examine diversifying question wording. By generating multiple questions testing the same causal relationship, students can have multiple opportunities to solidify their knowledge. Our pipeline allows for straightforward experimentation with the question generation model.

## 9 Conclusion

We propose a new task, cause-and-effect question generation, along with a novel pipeline to utilize extracted causes and effects as intended answers for a question generation system. We provide automatic and manual evaluation metrics and show that fine-tuning QG and QA models on an auxiliary dataset of cause-and-effect questions improves performance. Our publicly-released pipeline can automatically generate cause-and-effect questions for educational resources at scale.

## Acknowledgements

This work was supported by an AWS Machine Learning Research Award, an NVIDIA Corporation GPU grant, an AI2 Key Scientific Challenge Proposal grant, and a National Science Foundation (NSF) Graduate Research Fellowship (DGE 1752814). We thank the three anonymous reviewers as well as Nate Weinman, Philippe Laban, Dongyeop Kang, and the Hearst Lab Research Group for their helpful comments.

## References

- Bengt Altenberg. 1984. [Causal linking in spoken and written english](#). *Studia Linguistica*, 38(1):20–69.
- Lorin W Anderson, Benjamin Samuel Bloom, et al. 2001. [A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives](#). Longman,.
- Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*.
- Ross Beatty Jr. 1975. Reading comprehension skills and bloom’s taxonomy. *Literacy Research and Instruction*, 15(2):101–108.
- Benjamin S. Bloom, David Krathwohl, and Bertram B. Masia. 1956. [Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain](#). Green.
- Manvi Breja and Sanjay Kumar Jain. 2020. Causality for question answering. In *COLINS*, pages 884–893.
- Mengyun Cao, Xiaoping Sun, and Hai Zhuge. 2016. [The role of cause-effect link within scientific paper](#). *12th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 32–39.
- Ying-Hong Chan and Yao-Chung Fan. 2019. [A recurrent BERT-based model for question generation](#). In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Kordula De Kuthy, Madeeswaran Kannan, Haemant Santhi Ponnusamy, and Detmar Meurers. 2020. [Towards automatically generating questions under discussion to link information and discourse structure](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5786–5798, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Kaustubh Dhole and Christopher D. Manning. 2020. [Syn-QG: Syntactic and shallow semantic rules for question generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 752–765, Online. Association for Computational Linguistics.
- Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. [Question generation for question answering](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.
- Rosa Falotico and Piero Quatto. 2015. [Fleiss’ kappa statistic without paradoxes](#). *Quality & Quantity*, 49(2):463–470.
- J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Roxana Girju. 2003. [Automatic detection of causal relations for question answering](#). In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan. Association for Computational Linguistics.
- 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.
- Aniruddha Kembhavi, Minjoon Seo, D. Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. [The NarrativeQA reading comprehension challenge](#). *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. [A Systematic Review of Automatic Question Generation for Educational Purposes](#). *International Journal of Artificial Intelligence in Education*, 30(1):121–204.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [Albert: A lite bert for self-supervised learning of language representations](#). In *International Conference on Learning Representations*.
- Allen Nie, Erin Bennett, and Noah Goodman. 2019. [Learning to explain: Answering why-questions via rephrasing](#). In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 113–120, Florence, Italy. Association for Computational Linguistics.
- 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 NAACL-HLT 2019: Demonstrations*.
- 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.
- Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. [ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410, Online. Association for Computational Linguistics.
- 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.
- Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. [Quizbot: A dialogue-based adaptive learning system for factual knowledge](#). In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.
- Katherine Stasaski and Marti A. Hearst. 2017. [Multiple choice question generation utilizing an ontology](#). In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 303–312, Copenhagen, Denmark. Association for Computational Linguistics.
- Will Thalheimer. 2003. *The learning benefits of questions*. Work Learning Research.
- 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.
- Tong Wang, Xingdi (Eric) Yuan, and Adam Trischler. 2017a. [A joint model for question answering and question generation](#). In *Learning to generate natural language workshop, ICML 2017*.
- Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017b. [Gated self-matching networks for reading comprehension and question answering](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada. Association for Computational Linguistics.
- Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J. Grimaldi, and Richard G. Baraniuk. 2018. [Qg-net: A data-driven question generation model for educational content](#). In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale, L@S '18*, New York, NY, USA. Association for Computing Machinery.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. [Machine comprehension by text-to-text neural question generation](#). In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.
- 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 Manual Analysis of SQuAD Questions

We conduct a manual analysis of the 68 most-frequent n-gram question openers from SQuAD, to investigate the prevalence of causal questions which exist. We restrict the question openers to the most frequent n-grams which cover 80% of the dataset, excluding unigrams other than “why” because they were uninformative. Two researchers hand-labeled each opener as one of: cause-and-effect, not cause-and-effect, and unknown, where unknown question openers could be the start of cause-and-effect questions but would require reading the entire question to determine. A third researcher tie-broke disagreements. The average Cohen’s Kappa (Cohen, 1960) is 0.72 (substantial agreement). Of the 87,599 questions in SQuAD which could be labeled by the 68 most-frequent n-gram question openers, 1,194 (1.4%) were cause-and-effect, 29,540 (33.7%) were not cause-and-effect, and 56,865 (64.9%) were unknown. The full list of labeled question openers is found in our Github repo.

## B Causal Extraction Details

For the causal extraction section of the pipeline, we modified the “as” pattern. The original pattern is formulated as:

```
&R@Complete@ (,) (-such/-same/-seem/-regard/-regards/-regarded/-view/-views/-viewed/-denote/-denoted/-denotes) as (-if/-follow/-follows/-&adv) &C@Complete@
```

Where @Complete@ indicates that the text piece is a clause which must have predicate and subject, “-” indicates tokens followed should not be matched, and “()” indicates tokens that are not required. &R and &C represent the extracted cause and effect. However, the original pattern assumes that the cause is always before “as.” In reality, “as” can be included before both the cause and the effect, such as in the following example:

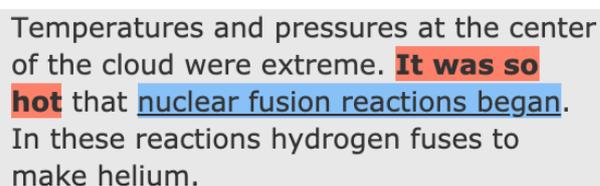
Some renewable resources are too expensive to be widely used. As the technology improves and more people use renewable energy, the prices will come down. The cost of renewable resources will go down relative to fossil fuels as we use fossil fuels up.

For this example, the causal phrase extracted by original pattern is “Some renewable resources are too expensive to be widely used.” The effect phrase extracted by original pattern is “The technology improves and more people use renewable energy, the prices will come down.”

We implement a new pattern (pattern-id = 145): “;/./- As &C , &R”. For each cause-and-effect extracted in the original pattern, if the new pattern is also a match, we replace the cause and effect with the output from the new pattern. For the example sentences above, the causal phrases extracted by our new pattern is “the technology improves and more people use renewable energy.” The corresponding effect phrase extracted by new pattern is “The prices will come down.”

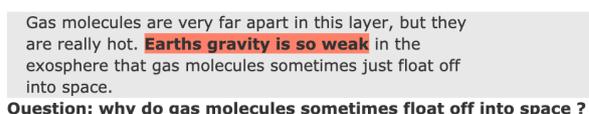
## C Crowdfunding Task Interface

Figure 2 contains the cause (bolded and highlighted in orange) and effect (underlined and highlighted in blue) shown to workers when evaluating the quality of an extracted cause and effect. Figure 3 contains a sample stimulus showing the intended answer and generated question.



Temperatures and pressures at the center of the cloud were extreme. **It was so hot** that nuclear fusion reactions began. In these reactions hydrogen fuses to make helium.

Figure 2: Example crowdfunding presentation of passage, cause, and effect, from TQA dataset.



Gas molecules are very far apart in this layer, but they are really hot. **Earths gravity is so weak** in the exosphere that gas molecules sometimes just float off into space.

Question: why do gas molecules sometimes float off into space ?

Figure 3: Example crowdfunding presentation of passage, intended answer, and generated question, from TQA dataset.

Table 10 contains the crowdworker ratings for the Cao et al. (2016) causal extraction system, stratified by typology. Each main typology category is further stratified by the type and sub-type of link. For example, the Adverbial link category contains two types: (A) Anaphoric and (B) Cataphoric. The Anaphoric category is further segmented into three sub-types: (1) Implicit Cohesion (e.g., “therefore”) (2) Pronominal Cohesion (e.g., “for this reason”),

and (3) Pronominal + Lexical Cohesion (e.g., “because of” *NP*) (Altenberg, 1984). We refer to the main category by name, with the subcategories denoted with codes, e.g., Adv.1.a.

Category	TQA	#T	SQuAD	#S
Adv.1.a	33	34	16	21
Adv.1.b	2	3	2	2
Adv.1.c	0	4	1	3
Adv.2.a	2	2	1	2
Prep.1.a	8	11	16	21
Sub.1.a	12	27	18	29
Sub.2.a	2	3	0	4
Sub.3.a	2	3	3	3
C-I.1.a	3	4	3	5
C-I.1.b	3	3	2	3
C-I.1.c	2	3	2	3
C-I.1.e	*	0	1	1
C-I.1.f	1	1	*	0
C-I.1.h	*	0	1	1
C-I.2.a	0	2	2	2

Table 10: Number of extracted relations that were labeled as causal by crowdworkers for original Cao et al. (2016) system, organized by linguistic category. #T is total number in TQA and #S is total number in SQuAD. ‘\*’ indicates no relation found.

## D Model specifics

**Causal Extraction:** The approximate runtime for this algorithm on an Intel(R) Core(TM) i7-6850K CPU machine is 72 hours for the TQA dataset and 24 hours for SQuAD.

**Question Generation:** ProphetNet has 391,324,672 parameters; our version is unchanged from Qi et al. (2020). We finetune the provided question generation model checkpoint, which is a 16 GB model fine-tuned on SQuAD. The approximate runtime to fine tune this model on an auxiliary dataset on a p3.2xlarge AWS ec2<sup>6</sup> machine is 0.5 hours. For our fine-tuning process, we train for 3 epochs with a learning rate of 1e-6 with a batch size of 1. The rest of parameters are kept the same as what is found in the examples provided by the ProphetNet GitHub repository README. Approximate inference time is 10 minutes for TQA and 5 minutes for SQuAD. We utilize the Fairseq library (Ott et al., 2019) to facilitate the training and inference processes. Comparing the fine-tuned model’s generated

Type	Recall	#T	Recall	#S
Adv.1.a	0.76	1309	0.60	229
Adv.1.b	0.55	28	0.32	8
Adv.1.c	0.37	94	0.39	15
Adv.2.a	0.45	3	0.67	3
<b>Total Adv.</b>	0.73	1434	0.58	255
Prep.1.a	0.72	470	0.56	275
<b>Total Prep.</b>	0.72	470	0.56	275
Sub.1.a	0.69	1146	0.52	385
Sub.2.a	0.73	57	0.54	35
Sub.3.a	0.78	70	0.59	15
<b>Total Sub.</b>	0.70	1273	0.53	435
C-I.1.a	0.71	99	0.57	48
C-I.1.b	0.61	33	0.56	11
C-I.1.c	0.79	29	0.90	13
C-I.1.e	*	0	0.12	1
C-I.1.f	1	1	*	0
C-I.1.h	*	0	0.14	1
C-I.2.a	0.61	20	0.32	6
<b>Total C-I</b>	0.69	182	0.59	80

Table 11: Average cause/effect presence recall in the TQA and SQuAD datasets, categorized by typology. Questions are generated by ProphetNet fine-tuned on Syn-QG. #T refers to number of questions in TQA; #S the same for SQuAD. ‘\*’ indicates no relation found.

questions to the Syn-QG questions, the fine-tuned QG model achieves 57.01 training BLEU and 53.89 test BLEU (Papineni et al., 2002).

**Question Answering:** The QA model we utilize has 334,094,338 parameters. The approximate runtime to fine tune this model on an auxiliary dataset on a p3.2xlarge AWS ec2 machine is 0.5 hours. For our fine-tuning process, we train for 10 epochs with an initial learning rate of 1e-5, a batch size of 4, 500 warm-up steps, and a weight decay of 0.01. The rest of the parameters are the defaults set by the HuggingFace TrainingArguments class. We also truncate each example to a max of 512 tokens. Approximate inference time is 1 minutes for TQA and 1 minute for SQuAD. On the Syn-QG dataset, the fine-tuned QA model achieves 0.97 training F1 and 0.95 test F1.

## E Cause/Effect Present Results

Table 11 shows the results for the automatic cause/effect present metric segmented by typology categories. For SQuAD, the lowest-performing category is Subordination, which corresponds to the category with the lowest proportion of extracted relationships labeled as causal by crowdworkers (Section 4).

<sup>6</sup><https://aws.amazon.com/ec2/>