Ask to Learn: A Study on Curiosity-driven Question Generation

Thomas Scialom
reciTAL, Paris, France
Sorbonne Université, CNRS, LIP6
thomas@recital.ai

Jacopo Staiano
reciTAL, Paris, France
jacopo@recital.ai

Abstract

We propose a novel text generation task, namely *Curiosity*-driven Question Generation. We start from the observation that the Question Generation task has traditionally been considered as the dual problem of Question Answering, hence tackling the problem of generating a question given the text that contains its answer. Such questions can be used to evaluate machine reading comprehension. However, in real life, and especially in conversational settings, humans tend to ask questions with the goal of *enriching* their knowledge and/or *clarifying* aspects of previously gathered information.

We refer to these inquisitive questions as *Curiosity*-driven: these questions are generated with the goal of obtaining new information (the answer) which is not present in the input text. In this work, we experiment on this new task using a conversational Question Answering (QA) dataset; further, since the majority of QA dataset are not built in a conversational manner, we describe a methodology to derive data for this novel task from non-conversational QA data. We investigate several automated metrics to measure the different properties of *Curious Questions*, and experiment different approaches on the *Curiosity*-driven Question Generation task, including model pre-training and reinforcement learning. Finally, we report a qualitative evaluation of the generated outputs.

1 Introduction

The growing interest in Machine Reading Comprehension (MRC) has sparked significant research efforts on Question Generation (QG), the dual task to Question Answering (QA). In QA, the objective is to produce an adequate response given a query and a text; conversely, for QG, the task is generally defined as generating relevant questions given a source text and, optionally, a specific target answer included therein. To our knowledge, all works tackling QG have thus far exclusively focused on generating relevant questions which can be answered given the source text: for instance, given *The 1st COLING conference took place in 1965* as input, a question likely to be automatically generated would be *When did the 1st COLING conference take place?*, where the answer *1965* is a span of the input. Such questions are useful to evaluate reading comprehension for both machines (Hermann et al., 2015; Eyal et al., 2019) and humans (Mani et al., 1999).

However, the human ability of asking questions goes well beyond evaluation: asking questions is essential in education (Gall, 1970) and has been proven to be fundamental for children cognitive development (Chouinard et al., 2007). Curiosity is baked into the human experience: it allows to extend one's comprehension and knowledge by asking questions that, while being relevant to context, are not directly answerable by it, thus being *inquisitive* and *curious*. The significance of such kind of questions is two-fold: first, they allow for gathering novel relevant information, *e.g.* a student asking for clarification; second, they are tightly linked to one's understanding of the context, *e.g.* a teacher testing a student's

Proceedings of the 28th International Conference on Computational Linguistics, pages 2224–2235
Barcelona, Spain (Online), December 8-13, 2020

knowledge by asking questions whose answers require a deeper understanding of the context and more complex reasoning.

From an applicative point of view, we deem the ability to generate curious, inquisitive, questions as highly beneficial for a broad range of scenarios: *i)* in the context of human-machine interaction (e.g. robots, chat-bots, educational tools), where the communication with the users could be more natural; and *ii)* during the learning process itself, which could be partially driven in a self-supervised manner, reminiscent of how humans learn by exploring and interacting with their environment. To our knowledge, this is the first paper attempting to tackle *Curiosity*-driven neural question generation.

The contributions of this paper can be summarized as follows:

- we propose a new natural language generation task: *curiosity*-driven question generation;
- we propose a method to derive data for the task from popular *non-conversational* QA datasets;
- we experiment using language model pre-training and reinforcement learning, on two different datasets;
- we report a human evaluation analysis to assess both the pertinence of the automatic metrics used and the efficacy of the proposed dataset-creation method above.

2 Related Works

Deep learning models have been widely applied to text generation tasks such as machine translation (Kalchbrenner and Blunsom, 2013), abstractive summarization (Rush et al., 2015) or dialog (Henderson et al., 2013), providing significant gains in performance. The state of the art approaches are based on sequence to sequence models (Cho et al., 2014; Sutskever et al., 2014). In recent years, significant research efforts have been directed to the tasks of Machine Reading Comprehension (MRC) and Question Answering (QA) (Hermann et al., 2015; Rajpurkar et al., 2016). The data used for tackling these tasks are usually composed of $\{context, question, answer\}$ triplets: given a context and the question, a model is trained to predict the answer.

Following QA, research on Question Generation (QG) (Amidei et al., 2018) has also seen increasing interest from the community: the QG task (Du et al., 2017; Zhou et al., 2017) can be considered as the dual task for QA (Duan et al., 2017): given a context and an answer span, the model is trained to generate the corresponding question. One of the main motivations is that an effective QG model can be used to generate synthetic data in order to augment existing QA datasets (Yuan et al., 2017; Alberti et al., 2019). For instance, Yuan et al. (2017) proposed a reinforcement learning setup trained using a QA-based metric reward: given a paragraph and an answer, the model first generates questions; then, the paragraph and the corresponding generated questions are given to a pre-trained QA model which predicts an answer; finally, the reward is computed as the number of overlapping words between the ground truth answer and the predicted answer – in other words, the reward to maximize, for the QG model, corresponds to the QA score. For an extensive evaluation of models trained with different rewards we refer the reader to the work of Hosking and Riedel (2019).

Most of these works follow the approach by Ranzato et al. (2015), who applied reinforcement to neural machine translation: first, a sequence to sequence model is trained under teacher forcing (Williams and Zipser, 1989) to optimize cross-entropy, hence helping to reduce the action space (i.e. the vocabulary size); then, the model is finetuned with a mix of teacher forcing and REINFORCE (Williams, 1992). For automatic evaluation, all previous works on QG resort to BLEU metrics (Papineni et al., 2002), originally developed and widely used in Machine Translation. However, how to evaluate text generation models remains an open research question: Nema and Khapra (2018) pointed out that, on QG tasks, the correlation between BLEU and human evaluation was poor.

A thorough investigation of the behavior of open-domain conversational agents has been recently presented by See et al. (2019). Using controllable neural text generation methods, the authors control important attributes for chit-chat dialogues, including question-asking behavior. Among the take-away messages of this work, is that question-asking represents an essential component in an engaging chit-chat pipeline: the authors find, via a large-scale human validation study, that agents with higher rates of question-asking obtain qualitative improvements in terms of inquisitiveness, interestingness and engagingness.

Indeed, in a conversational setting, it can be expected that the nature of follow-up questions significantly differs from those used as target in a traditional QG training setup: as mentioned earlier, QG has so far been framed as the dual task to QA, hence training models to generate questions whose answer is present in the input context. In contrast, we argue that in natural conversations the questions *follow* the input context but are rather a means to augment one's knowledge (as their answer is *not* explicit in the input context). In this work, we thus define the task as *Curiosity*-driven Question Generation.

3 Dataset

Question Answering datasets are usually composed of a set of questions associated with reading passages (the *context*) and the corresponding answers contained therein. The QA task is defined as finding the answer to a question given the context. As opposed, the Question Generation (QG) task is to generate the question given the input and (optionally) the answer. Most previous efforts on the QG task have resorted to the widely used Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). It contains roughly 100,000 questions posed by crowd-workers on a selected sample of Wikipedia articles. Several other QA datasets have also been recently published accounting for characteristic such as requiring multi-passage or discrete reasoning (Yang et al., 2018; Dua et al., 2019); further, *conversational* QA datasets have been made available: CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) have the desirable property to be in a dialogue-like setting.

In our scenario, *Curiosity*-driven QG, the reading passage associated with a question should *not* contain the answer, but rather pave the way for asking a new, *curious*, question – whose answer would eventually enrich the knowledge on the matter at hand. Therefore, a natural choice to build QG data would be to rely on existing datasets for *conversational* QA. A detailed comparison of the above-mentioned CoQA and QuAC datasets is provided by Yatskar (2019), who reports the proportion of *Topic Error* (*i.e.* questions unlikely to be asked in the context) and *Entity Salad* (*i.e.* questions unanswerable for any context):[1] compared to QuAC, CoQA is found to include significantly more *Topic Error* and *Entity Salad* questions. For this reason, we resort to QuAC in order to derive data *Curiosity*-driven QG.

Furthermore, recognizing the fact that the great majority of QA datasets available does not account for conversational characteristics, we propose a methodology to derive data for *Curiosity*-driven Question Generation from standard QA datasets, and apply it to the popular SQuAD (Rajpurkar et al., 2016).

For both our data sources, and consistently with standard QA and QG tasks, we encode each sample as a triplet $\{P, q, a\}$ where the paragraph P comprises n sentences $[s_0, ..., s_n]$, and a represents the answer to the question q. A canonical QG approach would thus use s_a, *i.e.* the sentence of P that contains the answer, as source, and q as generation target. On the contrary, for *Curiosity*-driven QG, any sentence s_x from P can potentially be used as the source sequence, as long as it does not contain the answer – *i.e.* under the necessary constraint of $x \neq a$. In the following subsections, we elaborate on additional constraints depending on the nature of the source data.

In general, we define samples as triplets

$$t = \{s_x, P', y\} \tag{1}$$

where s_x and P' are, respectively, the input sentence and the paragraph P modified according to the appropriate dataset-depending constraint, as detailed in the following, and y is the reference (target) question.

3.1 Conversational QA Data

As mentioned above, we first derive our data from the QuAC dataset, which is built from Wikipedia articles by iterating over the following procedure: given a sentence, a student annotator asks a relevant

[1] see section 2.1 in Yatskar (2019)

	Train	Dev	Test
Learning to ask	86,635	8,965	8,964
Unconstrained	342,768	27,624	27,807
Constrained	25,356	2,076	2,087

Table 1: Data distributions over the train-validation-test splits. *Learning to ask* refers to the original split by Du et al. (2017), from which our data is derived. The bottom rows refer to the data we obtain using our methodology, with and without NER constraining.

question for which he does not have the answer; then, the teacher (annotator) retrieves a sentence that contains the answer. Thus, the logical conversational ordering in QuAC makes each question *curious* by design, given the text that precedes it. More formally, for a question q (our target), we consider the source s_x as the text P' preceding the sentence s_a that contains the answer. In other words, our QuAC-derived dataset is built by applying the stricter constraint $x < a$. Numerically, QuAC compounds to 83,568 questions (on 11,567 articles) for the train set, 7,354 for the validation set and 7,353 for the test set (each covering 1,000 articles). Since the test set is not public, we use the original QuAC validation set for testing. From the training set, we randomly drop 1,000 articles (hence, 7,224 samples) which we use to derive our validation set, thus resulting in 76,345 questions for training.

3.2 Standard QA Data

As discussed in section 2, most of the available QA datasets are not conversational. Thus, we propose a simple method to obtain data for *Curiosity*-driven QG from standard QA datasets. For this, we use the widely popular SQuAD (Rajpurkar et al., 2016), and specifically the original splits released by Du et al. (2017), which are commonly used for Question Generation. As opposed to QuAC, the questions in SQuAD do not follow a logical ordering. Therefore, any sentence s_x from P can potentially be used as the source sequence, as long as it does not contain the answer a (constraint: $x \neq a$). Nonetheless, as is reasonable for factoid QA datasets, several questions are so specific to their associated sentence s_a that they would be extremely unlikely to be asked without knowing the contents of s_a itself. To exemplify this issue, take the following paragraph from SQuAD:

> *Nikola Tesla was the fourth of five children. Nikola had an older brother named Dane [..]*

Given *"Nikola had an older brother named Dane."* as s_a, and operating under the sole constraint of $x \neq a$, the sentence *"Nikola Tesla was the fourth of five children"* would be eligible as a source s_x for the target question *"Who was Dane?"*. This question can only be asked if either contextual information or background knowledge is available, since it requires to know that *Dane* was among Tesla's four siblings. To overcome this problem, we added an additional constraint based on Named Entity Recognition (NER): s_x is an acceptable input only if all the entities present in the question q are also present in the input sentence s_x. In the previous example, this would thus filter out the target *"Who was Dane?"* while allowing for *"How much brothers and sisters Nikola have?"*. For our experiments we used `spaCy`.[2]

In Table 1 we report the number of samples we obtained from SQuAD before and after applying NER filtering. After applying the above methodology to construct a dataset for *Curiosity*-driven QG, our training dataset contains 25,356 samples for training, 2,076 for development, and 2,087 for testing.

4 Metrics

Automatic evaluation of Natural Language Generation (NLG) systems is a challenging task (Nema and Khapra, 2018). For QG, n-gram based similarity metrics are commonly used. These measures evaluate how similar the generated text is to the corresponding reference(s). While they are known to suffer from several shortcomings (Liu et al., 2016; Paulus et al., 2017; Scialom et al., 2019a), they allow to evaluate specific properties of the developed models. In this work, we use various automatic metrics detailed below, and we assess their quality for our task through a human evaluation – see Section 6.

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

BLEU One of the most popular metrics for QG, BLEU (Papineni et al., 2002) provides a set of measures to compare automatically generated texts against one or more references. In particular, BLEU-N is based on the count of overlapping n-grams between the candidate and its corresponding reference(s).

Self-BLEU Within the field of Computational Creativity, *Diversity* is considered a desirable property (Karampiperis et al., 2014). Indeed, generating always the same question such as *"What is the meaning of the universe?"* would be an undesirable behavior, reminiscent of the "collapse mode" observed in Generative Adversarial Networks (GAN) (Goodfellow et al., 2014). Therefore, we adopt Self-BLEU, originally proposed by Zhu et al. (2018), as a measure of diversity for the generated text sequences. Self-BLEU is computed as follows: for each generated sentence s_i, a BLEU score is computed using s_i as hypothesis while the other generated sentences are used as reference. When averaged over all the references, it thus provides a measure of how diverse the questions are: low Self-BLEU scores indicate high diversity.

QA-based metrics Given a text, a question can be considered *curious* if the answer is not contained in the input text. In our framing, this implies that a question q should not be answerable given its corresponding input sentence s_x. Thanks to the recent improvements obtained on Question Answering tasks – for instance, human-level performance has been achieved on SQuAD-v1[3] – the *answerability* of a question can be automatically measured. Therefore, given a question-context pair as input to a QA model, two type of metrics can be computed as: *n-gram-based*, measuring the average n-gram overlap between the retrieved answer and the ground truth; and, *probability-based*: the confidence of the QA model for its retrieved answer; this corresponds to the probability of being the correct answer assigned by the QA model to the retrieved answer. This latter metric is more abstractive, allowing more flexibility beyond n-grams.

Since several diverse questions can be generated for a given input, we consider the latter metric (*probability-based*) to better fit the *Curiosity*-driven QG task. Hence, given the evaluated question q and the input text s_x, we define a metric *QA_prob* as the confidence of the QA model that its predicted answer is correct. This metric measures *answerability* of q given s_x: therefore, the lower this score, the less likely the answer is contained in the input text.

While being non-answerable represents a necessary condition for q being a *curious* question with respect to its context s_x, we also want q to be as relevant and useful as possible. To this end, we compute the above *QA_prob* for question q on P', which represents the source paragraph stripped from the sentence containing the answer (see Eq. 1). The higher this score, the more likely the question is relevant and useful to augment the knowledge provided by s_x. Thus, the two proposed metrics are defined as $QA_{source} = QA_{prob}(q, s_x)$ and $QA_{context} = QA_{prob}(q, P')$. Hence, under our definition, *Curiosity*-driven questions are those that minimize QA_{source} while maximizing $QA_{context}$. In other words, we want a curious question to not be answerable given its input, while being answerable given the context.

To compute these QA-based metrics, we use the HuggingFace implementation[4] of BERT (Devlin et al., 2018).

5 Experiments

Baseline model As baseline architecture we adopt the popular Transformer (Vaswani et al., 2017), which proved to perform well on a wide range of text generation tasks. Among these, neural machine translation (Ott et al., 2018b), automatic summarization (Gehrmann et al., 2018), and question generation (Dong et al., 2019; Scialom et al., 2019b). It can be briefly described as a sequence-to-sequence model with symmetric encoder and decoder based on a self-attention mechanism, which allows to overcome the inherent obstacles to parallelism present in recurrent models such as Long Short Time Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997).

[3] https://rajpurkar.github.io/SQuAD-explorer/

[4] https://github.com/huggingface/pytorch-transformers

The copy mechanism (Gulcehre et al., 2016) proved beneficial for QG (Zhao et al., 2018; Scialom et al., 2019b); indeed, the QG task is very sensitive to rare and out of vocabulary words such as named entities and this mechanism helps deal with it efficiently: more than 50% of the answers in SQuAD, for instance, correspond to named entities – see Table 2 in Rajpurkar et al. (2016). Hence, following (Gehrmann et al., 2018; Scialom et al., 2019b), we include a copy mechanism in our Transformer architecture.

For our experiments, we used the following hyper-parameters for the transformer: `N=2` (number of blocks); `d_model=256` (hidden state dimension); `d_ff=512` (position-wise feed-forward networks dimension); and, `h=2` (number of attention heads). Experiments run with the original hyper-parameters[5] as proposed by Vaswani et al. (2017) obtained consistent and numerically similar results. During training, we used mini-batches of size 64 and the Adam optimizer (Kingma and Ba, 2014). At generation time, the decoding steps are computed via beam search with $k = 5$.

5.1 Reinforcement

Reinforcement Learning (RL) is an efficient technique to maximize discrete metrics for text generation. Previously, Ranzato et al. (2015) used the REINFORCE algorithm (Williams, 1992) to train RNNs for several generation tasks, showing improvements over previous supervised approaches. Moreover, Paulus et al. (2017) combined supervised and reinforcement learning, demonstrating improvements over competing approaches both in terms of ROUGE and on human evaluation. However, the metrics used as reward are often found to overfit, leading to numerical improvements which do not correspond to increased output quality – and rather contribute to degrading, leading to reduced effectiveness of the trained models for practical applications. On this matter, and with a particular focus on QG, Hosking and Riedel (2019) performed a human evaluation of RL models trained with several metrics as reward, finding them to be indeed poorly aligned with human judgments: the models appear to learn to exploit the weaknesses of the reward source. In particular, the model learns to generate questions which are adversarial to a QA model: while meaningless, the QA would systematically be duped into assigning a high probability for their answerability. For more details on adversarial probing of QA systems we refer to Jia and Liang (2017). To overcome this issue, we propose to use a balanced reward:

$$r(q, P, P') = QA_{context} - QA_{source} \tag{2}$$

thus maximizing the probability of finding an answer to the generated question within the input paragraph but not in the source sentence. We hypothesize that such a metric might lead the model to avoid generating adversarial questions, having to find a balance between $QA_{context}$ or $-QA_{source}$.

In our experiments, we follow the approach proposed by (Ranzato et al., 2015; Paulus et al., 2017), considering a mixed loss L_{ml+rl} which combines supervised and reinforcement learning schemes:

$$L_{ml+rl} = \gamma L_{rl} + (1 - \gamma) L_{ml} \tag{3}$$

where the maximum likelihood L_{ml} is defined as $L_{ml} = -\sum_{t=0}^{m} log(p(y_t|y_0, ..., y_{t-1}, X))$, with $X = [x_1, ..., x_n]$ representing the source text of length n and $Y = [y_1, ..., y_m]$ the corresponding reference question of length m. Conversely, we define the reinforcement loss L_{rl} to be *minimized* according to the standard RL actor-critic scheme, where $r(q, P, P')$ is the reward function defined in Section 2:

$$L_{rl} = (r(\widehat{Y}) - r(Y^s)) \sum_{t=0}^{m} log(p(y_t^s|y_0^s, ..., y_{t-1}^s, X)) \tag{4}$$

Greedy decoding according to the conditional distribution $p(y|X)$ is used to obtain a sequence $\widehat{Y}$. The model is sampled using its Markov property, *i.e.* one token at a time, producing the output sequence Y^s.

[5] `N=6`, `d_model=512`, `d_ff=2048`, `h=8`.

	human	base_b1	base_b3	base_b5	RL_b1	RL_b3	RL_b5
BLEU1	-	31.94	26.92	22.26	30.19	32.15	26.06
BLEU2	-	14.45	14.76	13.55	13.19	16.01	15.28
BLEU3	-	7.49	10.59	10.84	6.81	9.04	11.52
BLEU4	-	4.31	8.79	9.59	3.72	6.1	9.85
Self-BLEU1	96.09	99.84	99.88	99.95	99.96	99.94	99.96
Self-BLEU2	84.55	99.64	99.75	99.91	99.91	99.89	99.93
Self-BLEU3	70.55	99.39	99.63	99.87	99.86	99.84	99.9
Self-BLEU4	57.57	99.09	99.5	99.83	99.79	99.79	99.87
QA$_{\text{source}}$	44.5	48.86	35.8	29.88	57.54	41.36	35.03
QA$_{\text{context}}$	48.94	48.32	40.96	38.48	55.38	42.95	41.63

Table 2: Results obtained on QuAC-derived data. *b1*, *b3*, *b5* suffixes indicate the beam size used.

Pretraining (PT) As shown in Table 1, the constrained dataset amounts to roughly three times less samples than both QuAC and the original SQuAD dataset it derives from. We thus investigate, for this dataset, the impact of pretraining the model under the traditional (i.e. not *Curiosity*-driven) QG training setup, using the training set provided by Du et al. (2017)). Then, we resume training using the data obtained after applying the NER-based constraints for *Curiosity*-driven QG to the same training samples. For the QuAC *Curiosity*-driven dataset, the amount of data is comparable to the original dataset, given the *conversational* nature of QuAC. Therefore, we do not use pretraining for the experiments on QuAC.

6 Results

Automatic metrics In Table 2 we report the results of our experiments on QuAC for the baseline model (*base*) and the RL model. We use a beam k, and compute the results for $k = [1, 3, 5]$. In addition the generated questions with a beam $k = 5$, we also computed the results for $k = 1$ and $k = 3$.

While one would expect to see for all the metrics a slight improvement, with increasing beam size, we observe a strong divergence among the results: increasing values for k correspond to a significant improvements in terms of *BLEU-4* and notable drops for *BLEU-1*. A similar phenomena was observed by Ott et al. (2018a) in the context of machine translation: they found that the presence of 1 or 2% of noisy data is enough to significantly degrade the beam search results. In our case, one of most frequent generated question is *Are there any other interesting aspects about this article ?*. Indeed, the frequency of this question in our training set amounts to 4.18% of the questions. On the test set we see that roughly 80% of the generated questions start with the token *"are"* . This sequence is not very likely to be generated with a greedy search ($k = 1$): at any time step during the generation, if any other token has a higher probability, this question will be dismissed. On the other hand, with a higher beam, it is likely to be kept and eventually result as the most probable sequence, among the different remaining beams at the end of the inference, consistently with what observed by Ott et al. (2018a).

Moving to our SQuAD-based experiments, we observe that the models trained on SQuAD do not seem to suffer from this issue since all the metrics improved when increasing the beam size from $k = 1$ to $k = 5$. This is consistent with the results reported by (Zhao et al., 2018) where improving the beam improve slightly all the metrics. Thus, we only report the results with $k = 5$ in Table 3. A possible explanation is that SQuAD only contains factoid questions, as opposed to QuAC wherein, for instance, the open-ended question "Are there any other interesting aspects about this article" covers 4.18% of the samples.

We observe that the models trained with RL obtain, as could be expected, higher scores for *QA$_{context}$* with respect to those trained without RL. A higher *QA$_{context}$* implies that the QA model is more likely to find an answer in the near context of the source. *QA$_{source}$* is lower, as expected, for SQuAD based models, though comparatively higher than the models trained with RL on QuAC. We identify two possible reasons for this: first, the QA model is trained on answerable questions; second, the nature of the QUaC questions is less factoid than the SQuAD ones, and non-factoid questions can arguably be harder for the QA model

	human	**base**	**RL**	**PT**	**PT+RL**
BLEU1	-	32.81	31.71	**33.02**	32.13
BLEU2	-	14.31	13.67	**14.9**	14.58
BLEU3	-	7.57	7.21	**8.1**	7.81
BLEU4	-	4.12	3.88	**4.61**	4.53
Self-BLEU1	95.85	**93.80**	94.37	95.80	95.42
Self-BLEU2	87.96	**87.00**	88.80	91.29	90.71
Self-BLEU3	81.75	**79.59**	82.64	86.47	85.66
Self-BLEU4	77.60	**72.60**	76.48	81.63	80.52
$\mathbf{QA_{source}}$	54.12	57.85	**55.87**	63.13	58.46
$\mathbf{QA_{context}}$	74.93	52.11	**55.98**	50.81	**56.36**

Table 3: Results obtained on SQuAD-derived data.

	Answerability	**Correctness**	**External Knowledge**	**Relevance**	**Soundness**
base	1.23	4.07	2.41	2.54	3.21
RL	**1.14**	4.07	**2.66**	**2.65**	3.09
PT	**1.16**	**4.22**	2.30	2.43	3.13
PT+RL	1.35	**4.23**	2.21	2.53	3.06
human	*1.42*	*4.61*	*2.90*	*3.91*	*4.49*

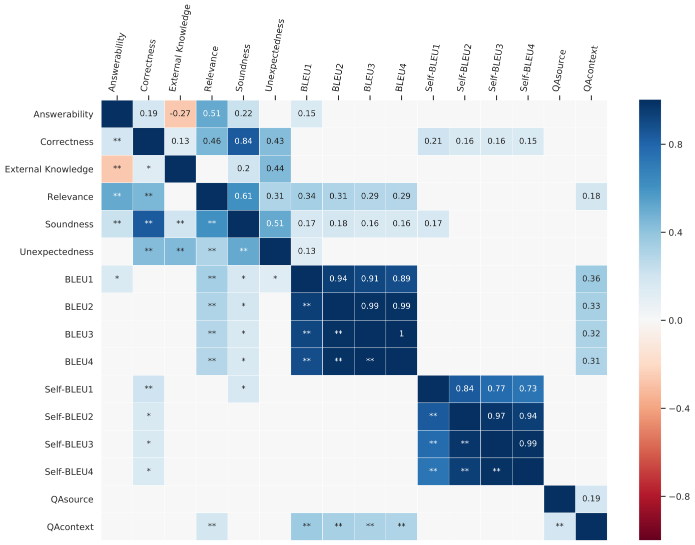
Table 4: Qualitative results obtained via human evaluation.

to evaluate. This could explain why, in the RL setting, $QA_{context}$ (the evaluation on answerable questions) is higher for both SQuAD and QUaC models, but only SQuAD models achieve a lower *QA_source* (the evaluation on non-answerable questions). Further, we observe that pretraining allows to achieve higher BLEU scores at the cost of lower Self-BLEU, thus showing an increased accuracy but less diversity in the generated questions. Indeed, we find that pretrained models tend to generate a higher number of questions starting with "*What*" compared to both other models and the references; the distribution for the first words of the human questions appears closer to that of non-pretrained models.

Human Evaluation In addition to the automatic metrics, we proceeded to a human evaluation. We chose to use the data from our SQuAD-based experiments in order to also to measure the effectiveness of the proposed approach to derive *Curiosity*-driven QG data from a standard, non-conversational, QA dataset. We randomly sampled 50 samples from the test set. Three professional English speakers were asked to evaluate the questions generated by: humans (*i.e.* the reference questions), and models trained using pre-training (PT) or (RL), and all combinations of those methods. Before submitting the samples for human evaluation, the questions were shuffled. Ratings were collected on a 1-to-5 likert scale, to measure to what extent the generated questions were: *answerable* by looking at their context; grammatically *correct*; requiring *external knowledge* to be answered; *relevant* to their context; and, semantically *sound*. The results of this human evaluation are reported in Table 4.

7 Discussion

What is the impact of the pretraining? We observe that for pretrained models (*i.e. PT* and *PT+RL*) the *Correctness* is significantly higher than the models without pretraining (*i.e. base* and *RL*). This is consistent with the higher BLEU observed for these models in Table 3. Additionally, we observe that for pretrained models the *External Knowledge* required to answer the generated questions is lower, while the *Relevance* is slightly higher. This might be due to the nature of the pretraining, during which the models learn to generate non-curious questions that focus on their inputs. Again, this is consistent with the significantly higher QA_source scores obtained by these models (see Table 3).



	Answerability	Correctness	External Knowledge	Relevance	Soundness	Unexpectedness	BLEU1	BLEU2	BLEU3	BLEU4	Self-BLEU1	Self-BLEU2	Self-BLEU3	Self-BLEU4	QAsource	QAcontext
Answerability		0.19	-0.27	0.51	0.22		0.15									
Correctness	**		0.13	0.46	0.84	0.43					0.21	0.16	0.16	0.15		
External Knowledge	**	*			0.2	0.44										
Relevance	**	**			0.61	0.31	0.34	0.31	0.29	0.29						0.18
Soundness	**	**	**	**		0.51	0.17	0.18	0.16	0.16	0.17					
Unexpectedness		**	**	**	**		0.13									
BLEU1	*			**	*	*		0.94	0.91	0.89						0.36
BLEU2				**	*		**		0.99	0.99						0.33
BLEU3				**	*		**	**		1						0.32
BLEU4				**	*		**	**	**							0.31
Self-BLEU1		**			*							0.84	0.77	0.73		
Self-BLEU2		*									**		0.97	0.94		
Self-BLEU3		*									**	**		0.99		
Self-BLEU4		*									**	**	**			
QAsource																0.19
QAcontext				**			**	**	**	**					**	

Figure 1: Correlation matrix obtained from the human assessment data ($*: p < .05$, $**: p < .005$).

Does Reinforcement help? From the human assessment we conducted – see Table 4, we observe that the models trained via RL obtain higher scores for *Relevance* and lower ones for *Soundness*, as compared to their non-reinforced counterparts. Further, the results reported in Table 3 show the reinforced models obtaining lower BLEU and QA_{source} source; conversely, they score higher when it comes to $QA_{context}$. We thus conclude that reinforcement brings improvements in terms of diversity of the generated questions, at the price of slightly degraded formulations in the outputs.

How effective is our dataset creation methodology? Looking at the bottom row of Table 4, which shows the scores obtained by the reference (human) questions, we observe the highest relative values for all dimensions, with the exception of *Answerability*. This indicates that the data we derived from a non-conversational QA dataset (SQuAD) fits well the task of *Curiosity*-driven question generation. As a sidenote, we remark that the models we built obtain lower scores than humans in terms of *Answerability*, a fact we hypothesize due to the lower quality of the generated questions: the less *sound* and *correct*, the less *answerable* a question would be, regardless of its context.

How well do the metrics fit human judgement? We report the pairwise Spearman correlation and p-value among all the different metrics and human measures in Figure 1. Our analysis shows that BLEU metrics correlate positively with *Relevance* (B4: .29, $p < .005$) and *Soundness* (B4: .19, $p < .005$), and to a weaker extent with *Answerability* (B1: .15, $p < .05$) and *Unexpectedness* (B1: .13, $p < .05$).[6] Self-BLEU metrics correlate significantly with *Soundness* (Self-B1: .17, $p < .05$) and *Correctness* (Self-B4: .15, $p < .05$), while $QA_{context}$ is associated with *Relevance* (.18, $p < .005$). The only human measure that does not correlate with any automatic metric is *External knowledge*. It is indeed one of the most challenging aspect to evaluate, even for humans. However, as expected, it correlates negatively with *Answerability*.

8 Conclusions

Asking inquisitive questions allows humans to learn from each other and increase their knowledge. We thus proposed a new task: *Curiosity*-driven Question Generation, which attempts to address such a key component for several human-machine interaction scenarios. In absence of data directly usable for this task, we proposed an automatic method to derive it from conversational QA datasets. Further, recognizing that the great majority of QA datasets are not conversational, we also extended the method to standard QA data. Our experiments, which include learning strategies such as pretraining and reinforcement, show promising results under both automatic and human evaluation. In future works, we plan to extend the approach to conditional generation of *Curiosity*-driven questions.

[6] For a standard QG task, Nema and Khapra (2018) report a Pearson correlation of 0.258 for BLEU-1 and 0.233 for BLEU-4.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA Corpora Generation with Roundtrip Consistency. *arXiv:1906.05416 [cs]*, June. arXiv: 1906.05416.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in Automatic Question Generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands, November. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs, stat]*, June. arXiv: 1406.1078.

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, October-November. Association for Computational Linguistics.

Michelle M Chouinard, Paul L Harris, and Michael P Maratsos. 2007. Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, pages i–129.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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. *arXiv:1905.03197 [cs]*, May. arXiv: 1905.03197.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. *arXiv:1705.00106 [cs]*, April. arXiv: 1705.00106.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*.

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.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question Answering as an Automatic Evaluation Metric for News Article Summarization. *arXiv:1906.00318 [cs]*, June. arXiv: 1906.00318.

Meredith D Gall. 1970. The use of questions in teaching. *Review of educational research*, 40(5):707–721.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the Unknown Words. *arXiv:1603.08148 [cs]*, March. arXiv: 1603.08148.

Matthew Henderson, Blaise Thomson, and Steve Young. 2013. Deep Neural Network Approach for the Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 467–471, Metz, France, August. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tom Hosking and Sebastian Riedel. 2019. Evaluating Rewards for Question Generation Models. *arXiv:1902.11049 [cs]*, February. arXiv: 1902.11049.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October. Association for Computational Linguistics.

Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 508–512. IEEE.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December. arXiv: 1412.6980.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. *arXiv:1603.08023 [cs]*, March. arXiv: 1603.08023.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a Better Metric for Evaluating Question Generation Systems. *arXiv:1808.10192 [cs]*, August. arXiv: 1808.10192.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018a. Analyzing Uncertainty in Neural Machine Translation. *arXiv:1803.00047 [cs]*, February. arXiv: 1803.00047.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018b. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Belgium, Brussels, October. 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:1606.05250 [cs]*, June. arXiv: 1606.05250.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence Level Training with Recurrent Neural Networks. *arXiv:1511.06732 [cs]*, November. arXiv: 1511.06732.

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

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv:1509.00685 [cs]*, September. arXiv: 1509.00685.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019a. Answers unite! unsupervised metrics for reinforced summarization models. 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 3237–3247.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019b. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6027–6032, Florence, Italy, July. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. *arXiv:1902.08654 [cs]*, February. arXiv: 1902.08654.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *arXiv:1409.3215 [cs]*, September. arXiv: 1409.3215.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*, June. arXiv: 1706.03762.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mark Yatskar. 2019. A Qualitative Comparison of. In *Proceedings of the 2019 Conference of the North*, pages 2318–2323, Minneapolis, Minnesota. Association for Computational Linguistics.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Sandeep Subramanian, Saizheng Zhang, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. *arXiv preprint arXiv:1705.02012*.

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, October. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural Question Generation from Text: A Preliminary Study. *arXiv:1704.01792 [cs]*, April. arXiv: 1704.01792.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. *arXiv preprint arXiv:1802.01886*.