Ask what's *missing* and what's *useful*: Improving Clarification Question Generation using Global Knowledge

Bodhisattwa Prasad Majumder[†♣] **Sudha Rao**[♢]
Michel Galley[♢] **Julian McAuley**[♣]
[♣]Department of Computer Science and Engineering, UC San Diego
{bmajumde, jmcauley}@eng.ucsd.edu
[♢]Microsoft Research, Redmond
{sudha.rao, mgalley}@microsoft.com

Abstract

The ability to generate clarification questions i.e., questions that identify useful missing information in a given context, is important in reducing ambiguity. Humans use previous experience with similar contexts to form a global view and compare it to the given context to ascertain what is missing and what is useful in the context. Inspired by this, we propose a model for clarification question generation where we first identify what is missing by taking a difference between the global and the local view and then train a model to identify what is useful and generate a question about it. Our model outperforms several baselines as judged by both automatic metrics and humans.

1 Introduction

An important but under-explored aspect of text understanding is the identification of *missing information in a given context* i.e., information that is essential to accomplish an underlying goal but is currently missing from the text. Identifying such missing information can help to reduce ambiguity in a given context which can aid machine learning models in prediction and generation (De Boni and Manandhar, 2003; Stoyanchev et al., 2014). Rao and Daumé III (2018, 2019) recently proposed the task of clarification question generation as a way to identify such missing information in context. They propose a model for this task which while successful at generating fluent and relevant questions, still falls short in terms of usefulness and identifying missing information. With the advent of large-scale pretrained generative models (Radford et al., 2019; Lewis et al., 2019; Raffel et al., 2019), generating fluent and coherent text is within reach. However, generating clarification questions requires going beyond fluency and relevance. Doing so requires understanding what is missing, which if included could be useful to the consumer of the information.

TITLE:	Sony 18x Optical Zoom 330x Digital Zoom Hi8 Camcorder
DESC:	Sony Hi-8mm Handycam Vision camcorder 330X digital zoom, Nightshot(TM) Infrared 0 lux system, Special Effects, 2.5" SwivelScreen color LCD and 16:9 recording mode, Laserlink connection. Image Stabilization, remote, built in video light.
QUESTION:	Can I manually control the video quality?

Table 1: Product description from amazon.com paired with a clarification question generated by our model.

Humans are naturally good at identifying missing information in a given context. They possibly make use of *global knowledge* i.e., recollecting previous similar contexts and comparing them to the current one to ascertain what information is *missing* and if added would be the most *useful*. Inspired by this, we propose a two-stage framework for the task of clarification question generation. Our model hinges on the concept of a "schema" which we define as the key pieces of information in a text. In the first stage, we find *what's missing* by taking a difference between the global knowledge's schema and schema of the local context (§3.1). In the second stage we feed this missing schema to a fine-tuned BART (Lewis et al., 2019) model to generate a question which is further made more *useful* using PPLM (Dathathri et al., 2019) (§3.2).[1]

We test our proposed model on two scenarios (§2): *community-QA*, where the context is a product description from amazon.com (McAuley and Yang, 2016) (see e.g. Table 1); and *dialog* where the context is a dialog history from the Ubuntu Chat forum (Lowe et al., 2015). We compare our model to several baselines (§4.2) and evaluate outputs using both automatic metrics and human evaluation to show that our model significantly outperforms baselines in generating useful questions that identify missing information in a given context (§4.4).

[†] Work done during an internship at Microsoft Research

[1] The code is available at https://github.com/microsoft/clarification-qgen-globalinfo

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4300–4312
June 6–11, 2021.

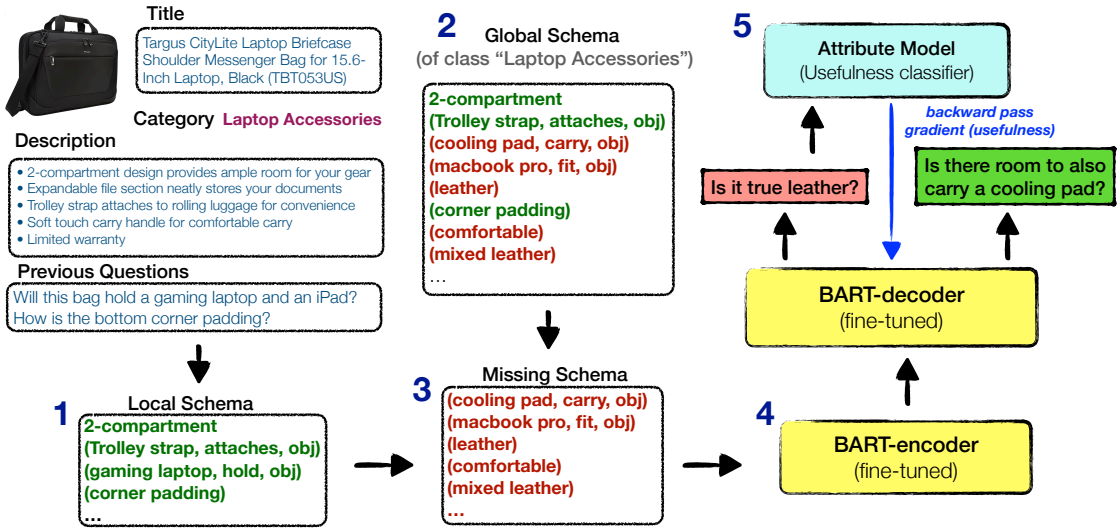


Figure 1: Test-time behaviour of our proposed model for *useful* clarification question generation based on *missing* information in a Community-QA (amazon.com) setup. 1. We obtain a local schema from the available context for a product: description and previously asked questions. 2. We obtain the global schema of the category of the product. 3. We estimate the *missing* schema that is likely to guide clarification question generation. 4. A BART model fine-tuned on (missing schema, question) pairs to generate a question ("*Is it true leather?*"). 5. A PPLM model with usefulness classifier as its attribute model further tunes the generated question to make it more *useful* ("*Is there room to also carry a cooling pad?*").

Furthermore, our analysis reveals reasoning behind generated questions as well as robustness of our model to available contextual information. (§5).

2 Problem Setup and Scenarios

Rao and Daumé III (2018) define the task of clarification question generation as: given a context, generate a question that identifies missing information in the context. We consider two scenarios:

Community-QA Community-driven question-answering has become a common venue for crowdsourcing answers. These forums often have some initial context on which people ask clarification questions. We consider the Amazon question-answer dataset (McAuley and Yang, 2016) where context is a product description and the task is to generate a clarification question that helps a potential buyer better understand the product.

Goal Oriented Dialog With the advent of high quality speech recognition and text generation systems, we are increasingly using dialog as a mode to interact with devices (Clark et al., 2019). However, these dialog systems still struggle when faced with ambiguity and could greatly benefit from having the ability to ask clarification questions. We explore such a goal-oriented dialog scenario using the Ubuntu Dialog Corpus (Lowe et al., 2015) consisting of dialogs between a person facing a technical issue and another person helping them resolve the issue. Given a context i.e a dialog history, the task is to generate a clarification question that would aid the resolution of the technical issue.

3 Approach

Figure 1 depicts our approach at a high level. We propose a two-stage approach for the task of clarification question generation. In the first stage, we identify the missing information in a given context. For this, we first group together all similar contexts in our data[2] to form the *global schema* for each high-level class. Next, we extract the schema of the given context to form the *local schema*. Finally, we take a difference between the local schema and the global schema (of the class to which the context belongs) to identify the missing schema for the given context. In the second stage, we train a model to generate a question about the most useful information in the missing schema. For this, we fine-tune a BART model (Lewis et al., 2019) on (missing schema, question) pairs and at test time, we use PPLM (Dathathri et al., 2019) with a usefulness classifier as the attribute model to generate a useful question about missing information.

3.1 Identifying Missing Information

Schema Definition Motivated by (Khashabi et al., 2017) who use essential terms from a

[2]See §4.1 for details to combine data splits

question to improve performance of a Question-Answering system, we see the need of identifying important elements in a context to ask a better question. We define schema of sentence s as set consisting of one or more triples of the form (key-phrase, verb, relation) and/or one or more key-phrases.

$$schema_s = \{\ element\ \};\ \text{where}$$
$$element \in \{(key\text{-}phrase, verb, relation),\ key\text{-}phrase\} \quad (1)$$

Schema Extraction Our goal is to extract a schema from a given context. We consider (key-phrase, action verb, relation) as the basic element of our schema. Such triples have been found to be representative of key information in previous work (Vedula et al., 2019). Given a sentence from the context, we first extract bigram and unigram key-phrases using YAKE (Yet-Another-Keyword-Extractor) (Campos et al., 2020) and retain only those that contain at least a noun. We then obtain the dependency parse tree (Qi et al., 2020b) of the sentence and map the key-phrases to tree nodes.[3] Now, to obtain the required triple, we need to associate a verb and a relation to each key-phrase. This procedure is described in Alg 1. At a high-level, we use the path between the key-phrase and the closest verb in the dependency tree to establish a relation between the key-phrase and the verb. In cases where there is no path, we use only the key-phrase as our schema element. Figure 2 shows an example dependency tree for a sentence.

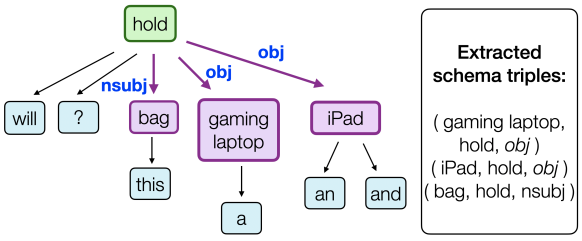


Figure 2: Dependency tree and paths showing how we obtain schema triples for a sentence: *"Will this bag hold a gaming laptop and an iPad?"* (from Figure 1).

Creating local schema Given a context, we extract a schema for each sentence in the context. The local schema of a context c is a union of schemata of each sentence s in the context.

$$local\_schema_c = \cup_{s \in c}\ schema_s \quad (2)$$

[3]In the case of bigram phrases, we merge the tree nodes.

Algorithm 1 Pseudocode for extracting (key-phrase, verb, relation) triple.

```
Initialize with empty path (path length ∞) for all possible
pairs of verbs (∈ {VB, VBG, VBZ}) and key-phrases in the
sentence
for Each verb and key-phrase pair do
    Search for the key-phrase among the children of the verb
    in the dependency tree
    if A key-phrase is found and path is shorter than the
    stored path then
        Update the path between the key-phrase and the verb
        pair
    end if
end for
for Each verb and key-phrase pair do
    if The key-phrase is the immediate child of the verb
    then
        Create the triple (key-phrase, verb, relation) using the
        relation in the path
    else
        Traverse backward from the key-phrase, stop at the
        first verb, use the relation with its immediate child in
        the path to create (key-phrase, verb, relation)
    end if
end for
```

Creating global schema We define global schema at the class level where a 'class' is a group of similar contexts. For *Amazon*, classes consist of groups of similar products and for *Ubuntu*, classes consist of groups of similar dialogs (see §4.1 for details). The global schema of a class K is a union of local schemata of all contexts c belonging to K.

$$global\_schema_K = \cup_{c \in K}\ local\_schema_c \quad (3)$$

A naive union of all local schemata can result in a global schema that has a long tail of low-frequency schema elements. Moreover, it may have redundancy where schema elements with similar meaning are expressed differently (e.g. *OS* and *operating system*). We therefore use word embedding based similarity to group together similar key-phrases and retain only the most frequent elements (see appendix).

Creating a missing schema Given a context c, we first determine the class K to which the context belongs. We then compute its missing schema by taking the set difference between the global schema of class K and the local schema of the context c:

$$missing\_schema_c = global_K \setminus local_c \quad (4)$$

More specifically, we start with the elements in the global schema and remove elements that have a semantic match (see appendix) with any element in the local schema to obtain the missing schema.

3.2 Generating Useful Questions

Our goal is to generate a useful question about missing information. In §3.1, we explained how we compute the missing schema for a given context; here we describe how we train a model to generate a useful question given the missing schema.

BART-based generation model Our generation model is based on the BART (Lewis et al., 2019) encoder-decoder model, which is also a state-of-the-art model in various generation tasks including dialog generation and summarization. We start with the pretrained base BART model consisting of a six layer encoder and six layer decoder. We fine-tune this model on our data where the inputs are the missing schema and the output is the question. The elements of the missing schema in the input are separated by a special `[SEP]` token. Since the elements in our input do not have any order, we use the same positional encoding for all input positions. We use a token type embedding layer with three types of tokens: key-phrases, verbs, and relations.

PPLM-based decoder We observed during our human evaluation[4] that a BART model fine-tuned in this manner, in spite of generating questions that ask about missing information, does not always generate *useful* questions. We therefore propose to integrate the usefulness criteria into our generation model. We use the Plug-and-Play-Language-Model (PPLM) (Dathathri et al., 2019) during decoding (at test time). The attribute model of the PPLM in our case is a usefulness classifier trained on bags-of-words of questions. In order to train such a classifier, we need usefulness annotations on a set of questions. For the Amazon dataset, we collect usefulness scores (0 or 1) on 5000 questions using human annotation whereas for the Ubuntu dataset we assume positive labels for (true context, question) pairs and negative labels for (random context, question) pairs and use 5000 such pairs to train the usefulness classifier. Details of negative sampling for Ubuntu dataset is in Appendix.

4 Experiments

We aim to answer the following research questions (RQ):

1. Is the model that uses missing schema better at identifying missing information compared to models that use the context directly to generate questions?
2. Do large-scale pretrained models help generate better questions?
3. Does the PPLM-based decoder help increase the usefulness of the generated questions?

[4]See results of BART+missinfo in Table 5

	Train	Validation	Test
Amazon	123,567	4,525	2,361
Ubuntu	102,678	7,864	200

Table 2: Number of data instances in the train, validation and test splits of Amazon and Ubuntu datasets (Both datasets are in English. Links are in appendix)

4.1 Datasets

Amazon The Amazon review dataset (McAuley et al., 2015) consists of descriptions of products on amazon.com and the Amazon question-answering dataset (McAuley and Yang, 2016) consists of questions (and answers) asked about products. Given a product description and N questions asked about the product, we create N instances of (*context*, *question*) pairs where *context* consists of the description and previously asked questions (if any). We use the "Electronics" category consisting of 23,686 products. We split this into train, validation and test sets (Table 2). The references for each context are all the questions (average=6) asked about the product. A class is defined as a group of products within a subcategory (e.g. DSLR Camera) as defined in the dataset. We restrict a class to have at most 400 products, and a bigger subcategory is broken into lower-level subcategories (based on the product hierarchy) resulting in 203 classes. While creating global schema, we exclude target questions from validation and test examples. The product descriptions and associated metadata come as inputs during test time. Hence, including them from all splits while creating the global schema does not expose the test and validation targets to the model during training.

Ubuntu The Ubuntu dialog corpus (Lowe et al., 2015) consists of utterances of dialog between two users on the Ubuntu chat forum. Given a dialog, we identify utterances that end with a question mark. We then create data instances of (context, question) where the question is the utterance ending with a question mark and the context consists of all utterances before the question. We consider only those contexts that have at least five utterances and at most ten utterances. Table 2 shows the number of data instances in the train, validation and test splits. Unlike the Amazon dataset, each context has only one reference question. A class is defined as a

group of dialogs that address similar topics. Since such class information is not present in the dataset, we use k-means to cluster dialogs into subsequent classes. Each dialog was represented using a TF-IDF vector. After tuning the number of clusters based on sum of squared distances of dialogs to their closest cluster center, we obtain 26 classes. We follow a similar scheme as with Amazon for not including target questions from validation and test sets while building the global schema.

4.2 Baselines and Ablations

Retrieval We retrieve the question from the train set whose schema overlaps most with the missing schema of the given context.

GAN-Utility The state-of-the-art model for the task of clarification question generation (Rao and Daumé III, 2019) trained on (context, question, answer) triples.

Transformer A transformer (Vaswani et al., 2017)[5] model trained on (context, question) pairs.

BART We finetune a BART model (Lewis et al., 2019) on (context, question) pairs.

BART + missinfo We compare to a BART model fine-tuned on (missing schema, question) pairs.

BART + missinfo + WD This is similar to the "BART + missinfo" baseline with the modification that, at test time only, we use a weighted-decoding (WD) strategy (Ghazvininejad et al., 2017) by re-defining the probability of words in the vocabulary using usefulness criteria (more in appendix).

BART + missinfo + PPLM This is our proposed model as described in § 3 where we fine-tune the BART model on (missing schema, question) pairs and use a usefulness classifier based PPLM model for decoding at test time.

4.3 Evaluation Metrics

4.3.1 Automatic Metrics

BLEU-4 (Papineni et al., 2002) evaluates 4-gram precision between model generation and references. at the corpus level; **METEOR** (Banerjee and Lavie, 2005) additionally uses stem and synonym matches for similarity; and **Distinct-2** (Li et al., 2016) measures diversity by calculating the number of distinct bigrams in model generations scaled by the total number of generated tokens.

[5]We use original hyperparameters & tokenization scheme.

4.3.2 Human Judgment

Similar to Rao and Daumé III (2019), we conduct a human evaluation on Amazon Mechanical Turk to evaluate model generation on the four criteria below. Each generated output is shown with the context and is evaluated by three annotators.

Relevance We ask "*Is the question relevant to the context?*" and let annotators choose between Yes (1) and No (0).

Fluency We ask "*Is the question grammatically well-formed i.e. a fluent English sentence?*" and let annotators choose between Yes (1) and No (0).

Missing Information We ask "*Does the question ask for new information currently not included in the context?*" and let annotators choose between Yes (1) and No (0).

Usefulness We perform a comparative study where we show annotators two model-generated questions (in a random order) along with the context. For Amazon, we ask "*Choose which of the two questions is more useful to a potential buyer of the product*". For Ubuntu, we ask "*Choose which of the two questions is more useful to the other person in the dialog*".

4.4 Experimental Results

4.4.1 Automatic Metric Results

Amazon Table 3 shows automatic metric results on Amazon. Under BLEU-4 and METEOR, the retrieval model performs the worst suggesting that picking a random question that matches the most with the missing schema does not always yield a good question. This strengthens the need of the second stage of our proposed model i.e. BART + PPLM based learning. GAN-Utility, which is state-of-the-art on Amazon, outperforms the Transformer baseline suggesting that training a larger model (in terms of the number of parameters) does not always yield better questions. BART, on the other hand, outperforms GAN-Utility suggesting the benefit of large-scale pretraining (RQ2). BART+missinfo further outperforms BART showing the value in training on missing schemata instead of training directly on the context (RQ1). A variation of this model that uses weighted decoding performs marginally better on METEOR but slightly worse of BLEU-4. Our final proposed model i.e., BART+missinfo+PPLM performs the best among all baselines across both BLEU-4 and METEOR.

Under diversity (Distinct-2), the retrieval model

Model	BLEU-4	METEOR	Distinct-2
Retrieval	8.76	9.23	**0.92**
GAN-Utility	14.23	16.82	0.79
Transformer	12.89	14.56	0.60
BART	15.98	16.78	0.78
+ missinfo	16.87	17.11	0.82
+ missinfo + WD	16.23	**17.98**	**0.84**
+ missinfo + PPLM	**18.55**	**18.01**	**0.86**
Reference	–	–	0.95

Table 3: Automatic metric results on the full test set of Amazon. The difference between bold and non-bold numbers is statistically significant with $p < 0.001$.

Model	BLEU-4	METEOR	Distinct-2
Retrieval	4.89	5.12	**0.82**
Transformer	6.89	7.45	0.67
BART	8.23	9.67	0.72
+ missinfo	**9.54**	10.78	**0.75**
+ missinfo + PPLM	**10.02**	**11.65**	**0.79**
Reference	–	–	0.87

Table 4: Automatic metric results the full test set of Ubuntu. The difference between bold and non-bold numbers is statistically significant with $p < 0.001$.

produces the most diverse questions (as also observed by Rao and Daumé III (2019)) since it selects among human written questions which tend to be more diverse compared to model generated ones. Among other baselines, transformer interestingly has the lowest diversity whereas GAN-Utility and BART come very close to each other. Model ablations that use missing schema produce more diverse questions further strengthening the importance of training on missing schema. Our model i.e., BART+missinfo+PPLM, in spite of outperforming all baselines (except retrieval), is still far from reference questions in terms of diversity, suggesting room for improvement.

Ubuntu Table 4 shows the results of automatic metrics on Ubuntu.[6] The overall BLEU-4 and METEOR scores are much lower compared to Amazon since Ubuntu has only one reference per context. Under BLEU-4 and METEOR scores, similar to Amazon, we find that the retrieval baseline has the lowest scores. Transformer baseline outperforms the retrieval baseline but lags behind BART, again showing the importance of large-scale pretraining. The difference between the BLEU-4 scores of BART+missinfo and our final proposed model is not significant but their METEOR score difference is significant suggesting that our model produces questions that may be lexically different from references but have more semantic overlap with the reference set. Under Distinct-2 scores, we find the same trend as in Amazon, with the retrieval model being the most diverse and our final model outperforming all other baselines.

4.4.2 Human Judgement Results

Amazon Table 5 shows the human judgment results on model generations for 300 randomly sampled product descriptions from the Amazon test set. Under relevancy and fluency, all models score reasonably with our proposed model producing the most relevant and fluent questions. Under missing information, the BART model, fine-tuned on context instead of missing schema, has the lowest score. GAN-Utility outperforms BART but significantly lags behind BART+missinfo and BART+missinfo+PPLM reaffirming our finding from the automatic metric results that our idea of feeding missing schema to a learning model helps.

We additionally observe that the human-written questions score lower than model-generated questions under 'fluency' and 'missing information' criteria, mirroring similar observations from (Rao and Daumé III, 2018, 2019). We believe the reason for this is that human-written questions often have typos or are written by non-native speakers (leading to lower fluency). Moreover, humans may miss out on reading full product descriptions causing them to ask about details that are already included in the description (leading to lower missing information scores).

Figure 3a shows the results of pairwise comparison on the usefulness criteria. We find that our model wins over GAN-Utility by a significant margin with humans preferring our model-generated questions 77% of the time. Our model also beats BART-baseline 66% of the time further affirming the importance of using missing schema. Finally, our model beats BART+missinfo model 61% of the time suggesting that the PPLM-based decoder that uses usefulness classifier is able to produce much more *useful* questions (RQ3). The annotator agreement statistics are provided in appendix.

Ubuntu Table 6 shows the results of human judgments on the model generations of 150 randomly sampled dialog contexts from the Ubuntu test set. In terms of relevance, we find that the transformer and BART baselines produce less relevant

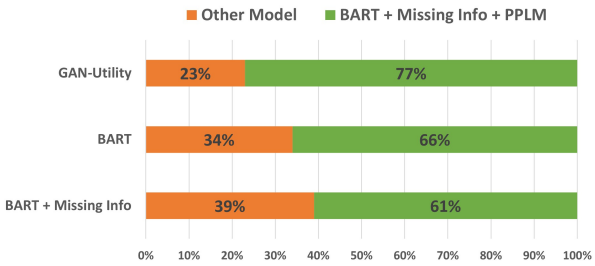
[6] We do not experiment with the GAN-Utility model (since it requires "answers") and the BART+missinfo+WD model (since usefulness labels are not obtained from humans).

Model	Relevancy	Fluency	MissInfo
GAN-Utility	0.9	0.86	0.81
BART	0.94	0.92	0.77
+ missinfo	0.97	0.92	0.87
+ missinfo + PPLM	**0.99**	**0.93**	**0.89**
Reference	0.96	0.83	0.89

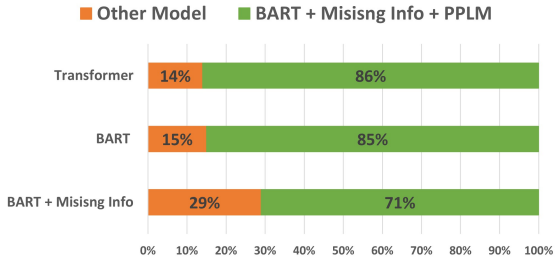
Table 5: Human judgment results (0-1) on 300 randomly sampled descriptions from the Amazon test set

Model	Relevancy	Fluency	MissInfo
Transformer	0.74	0.99	0.99
BART	0.69	0.99	0.96
+ missinfo	0.81	0.95	0.98
+ missinfo + PPLM	0.91	0.83	0.99
Reference	0.85	0.83	0.96

Table 6: Human judgment results (0-1) on 150 randomly sampled dialog contexts from Ubuntu test set



(a)



(b)

Figure 3: Results of a pairwise comparison (on usefulness criteria) between our model and baseline generated question on (a) 300 randomly sampled product descriptions from the Amazon test set, (b) 150 randomly sampled dialogs from the Ubuntu test set as judged by humans.

questions. With the addition of missing schema (i.e., BART+missinfo), the questions become more relevant and our proposed model obtains the highest relevance score. The reference obtains slightly a lower relevance score which can possibly be explained by the fact that humans sometimes digress from the topic. Under fluency, interestingly, the transformer and BART baselines obtain high scores. With the addition of missing schema, fluency decreases and the score reduce further with the PPLM model. We suspect that the usefulness classifier trained with a negative sampling strategy (as opposed to human labelled data, as in Amazon) contributes to fluency issues. Under missing information, all models perform well which can be explained by the fact that in Ubuntu, the scope of missing information is much larger (since dialog is much more open-ended) than in Amazon.

Figure 3b shows the results of pairwise comparison on usefulness criteria. We find that humans choose our model-generated questions 85% of time when compared to either transformer or BART generated questions. When compared to BART+missinfo, our model is selected 71% of the time, further affirming the importance of using the PPLM-based decoder.

5 Analysis

Robustness to input information We analyze how a model is robust toward the amount of information present. To measure the amount of information, we look for context length (description length for Amazon, dialog context length for Ubuntu) and the size of global schema since these two directly control how much knowledge regarding potential missing information is available to the model. We measure the difference in BLEU score between two groups of data samples where context length/size of global schema is either high or low. Figure 5 shows that our model is the least variant toward the information available hence more robust for the Amazon dataset.[7]

Owing to our modular approach for estimating missing information, we seek to analyze whether a question is really asking about missing information in an automatic fashion. This also allows us to explain the reasoning behind a particular generation as we are able to trace back to the particular missing information that is used to generate the question. We run a YAKE extractor on the generated questions to obtain key-phrases. We calculate the ratio between the number of key-phrases in the output that belong to the original missing schema and the total number of key-phrases present in the output. Table 8 shows that when we use our framework of estimating missing information coupled with BART, both models achieve very high missing information overlap, thus suggesting that we can obtain the reasoning behind a generated question reliably by tracing the missing information overlap, as shown in Table 9.

[7]Ubuntu follows similar trends; figure in appendix.

Amazon	
Category	Binoculars & Scopes
Title	Nikon 7239 Action 7x50 EX Extreme All-Terain Binocular
Description	The Monarch ATB 42mm with dielectric high-reflective Multilayer Prism coating binocular features brighter, sharper colors, crisp and drastically improved low-light performance. A new body style provides unparalleled strength and ruggedness in a package ...
Missing Schema	{mounting, **center focused**, (Nikon, works, obj), (Canon, works, obj), digital camera, . . . }
GAN-Utility	price?
BART	How is the focus quality?
BART+missinfo	Is it **center focused**?
BART+missinfo+PPLM	Is it **center focused**, or do you have to focus each eye individually?
Ubuntu	
Dialog history	User A: I'm having trouble installing nvidia drivers for my geforce 6200, could anyone perhaps assist? User B: i use the drivers from the website, much better User A: which drivers? from the website? User B: I used add/remove software from the menu to install nvidia proprietary drivers
Missing schema	{(driver, update, nsubj), (new version, install, nsubj), (machine, **reboot**, nsubj), ...}
Transformer	Did you try booting your machine?
BART	where did you download them from?
BART+missinfo	Can you tell the output after you install them?
BART+missinfo+PPLM	Can you try **rebooting** from the start and removing the software after installation?

Table 7: Model generations for an example product from Amazon and an example dialog context from Ubuntu.

Model	Amazon	Ubuntu
Retrieval	10.5	6.78
GAN-Utility	73.4	–
Transformer	57.2	45.7
BART	60.3	56.9
+ missinfo	97.3	89.2
+ missinfo + PPLM	**98.3**	**90.1**
Reference	99.7	93.7

Table 8: Missing information overlap (in %) between missing schema and output generations

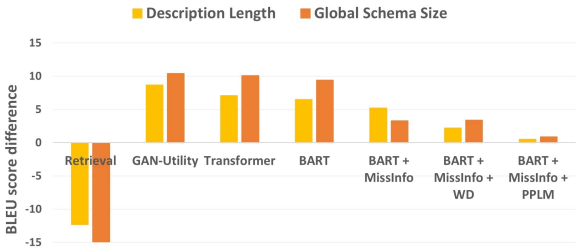


Figure 4: Average BLEU score difference between classes having longer (> 200 (median) words) and shorter descriptions; larger (> 200 (median) key-phrases) and shorter global schema for the Amazon dataset. Lower differences indicate more invariance toward the available information.

Question length We also observe in Table 9 that baseline models tend to generate short and generic questions as compared to our model that often chooses longer schema key-phrases (e.g. bigrams) to generate a more specific question. We further looked into annotated (for usefulness) questions from the Amazon dataset and we observed that 70% of questions that were annotated as useful are longer than not-useful questions. The average length of gold useful questions is 10.76 words and 8.21 for not-useful questions. The average length of generated questions for BART, BART+MissInfo and BART+MissInfo+PPLM (ours) are 5.6, 6.2, 12.3 respectively. We also find a similar trend in the Ubuntu dataset as well.

Dynamic expansion of global schema We anticipate that even if we build the global schema from the available offline dataset, it is possible that new entries may appear in a real application. We investigate how our framework responds to the dynamic expansion of global schema. We simulate a scenario where we extend the "Laptop Accessories" category in the Amazon dataset, with 100 new products (those that appeared on Amazon.com after the latest entry in the dataset). We obtain key-phrases from their product descriptions and include them in the global schema for the category which amounts to a 21% change in the existing global schema. For 50 random products in the test set from the same category, we found that in 28 out of 50 cases (56%), the model picked a new schema element that is added later. This indicates that our framework is capable of supporting dynamic changes in the global schema and reflecting them in subsequent generations without retraining from scratch.

6 Related Work

Most previous work on question generation focused on generating reading comprehension style

questions i.e., questions that ask about information present in a given text (Duan et al., 2017; Zhang and Bansal, 2019). Later, Rao and Daumé III (2018, 2019) introduced the task of clarification question generation in order to ask questions about missing information in a given context. ClarQ (Kumar and Black, 2020) entails clarification questions in a question answering setup. However, unlike our work, these works still suffer from estimating the most useful missing information.

Recent works on conversational question answering also focused on the aspect of question generation or retrieval (Choi et al., 2018; Aliannejadi et al., 2019). Qi et al. (2020a) especially focused on generating information-seeking questions while Majumder et al. (2020) proposed a question generation task in free-form interview-style conversations. In this work, in addition to improving clarification question generation in a community-QA dataset, we are the first to explore a goal-oriented dialog scenario as well.

Representing context and associated global information in a structure format has been shown to improve performance in generation task (Das et al., 2019; Subramanian et al., 2018; Khashabi et al., 2017) in general and summarization (Fan et al., 2019) and story-generation (Yao et al., 2019) in particular. We also derive inspiration from recent works on information extraction from free-form text (Vedula et al., 2019; Stanovsky et al., 2016) and develop a novel framework to estimate missing information from available natural text contexts.

Finally, for question generation, we use BART (Lewis et al., 2019), that is state-of-the-art for many generation tasks such as summarization, dialog generation etc. Furthermore, inspired from recent works that use controlled language generation during decoding (Ghazvininejad et al., 2017; Holtzman et al., 2018), we use Plug-and-Play-Language-Model (Dathathri et al., 2019) to tune generations during decoding. While similar approaches for controllable generation (Keskar et al., 2019; See et al., 2019) have been proposed, we extend such efforts to enhance the usefulness of the generated clarification questions.

7 Conclusion

We propose a model for generating useful clarification questions based on the idea that missing information in a context can be identified by taking a difference between the global and the local view. We show how we can fine-tune a large-scale pre-trained model such as BART on such differences to generate questions about missing information. Further, we show how we can tune these generations to make them more useful using PPLM with a usefulness classifier as its attribute model. Thorough analyses reveal that our framework works across domains, shows robustness towards information availability, and responds to the dynamic change in global knowledge. Although we experiment only with Amazon and Ubuntu datasets, our idea is generalizable to scenarios where it is valuable to identify missing information such as conversational recommendation, or eliciting user preferences in a chit-chat, among others.

Acknowledgements We thank everyone in the Natural Language Processing Group at Microsoft Research, Redmond, with special mention to Yizhe Zhang, Bill Dolan, Chris Brockett, and Matthew Richardson for their critical review of this work. We also thank anonymous reviewers for providing valuable feedback. In addition to this, we want to acknowledge human annotators from Amazon Mechanical Turk for data annotation and human evaluation of our systems. BPM is partly supported by a Qualcomm Innovation Fellowship and NSF Award #1750063. Findings and observations are of the authors only and do not necessarily reflect the views of the funding agencies.

8 Broader Impact

We do not foresee any immediate ethical concerns since we assume that our work will be restricted in domain as compared to free-form language generation. We still cautiously advise any developer who wishes to extend our system for their own use-case (beyond e-commerce, goal-oriented conversations) to be careful about curating a global pool of knowledge for data involving sensitive user information. Finally, since we are finetuning a pre-trained generative model, we inherit the general risk of generating biased or toxic language, which should be carefully filtered. In general, we expect users to benefit from our system by reducing ambiguity (when information is presented in a terse fashion, e.g. in a conversation) and improving contextual understanding to enable them to take more informed actions (e.g. making a purchase).

References

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *EMNLP*.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip R. Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What makes a good conversation?: Challenges in designing truly conversational agents. In *CHI*.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. 2019. Building dynamic knowledge graphs from text using machine reading comprehension. In *ICLR*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. In *ICLR*.

Marco De Boni and Suresh Manandhar. 2003. An analysis of clarification dialogue for question answering. In *NAACL-HLT*.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *EMNLP*.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In *EMNLP-IJCNLP*.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *ACL, System Demonstrations*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *ACL*.

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

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2017. Learning what is essential in questions. In *CoNLL*.

Vaibhav Kumar and Alan W. Black. 2020. Clarq: A large-scale and diverse dataset for clarification question generation. In *ACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pages 110–119.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDial*.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian J. McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *EMNLP*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *WWW*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

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

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020b. Stanza: A python natural language processing toolkit for many human languages. In *ACL Demo*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *ACL*.

Sudha Rao and Hal Daumé III. 2019. Answer-based Adversarial Training for Generating Clarification Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155, Minneapolis, Minnesota. 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. In *NAACL-HLT*.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props. *CoRR*, abs/1603.01648.

Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2014. Towards natural clarification questions in dialogue systems. In *AISB symposium on questions, discourse and dialogue*, volume 20.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *ACL MRQA*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeuRIPS*.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2019. Towards open intent discovery for conversational text. *CoRR*, abs/1904.08524.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *AAAI*.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. *CoRR*, abs/1909.06356.

A Setup and Data

Schema While creating the global schema, we use word embedding[8] based similarity to perform hierarchical clustering of key-phrases[9] and group together key-phrases that have cosine similarity greater than a threshold, a hyperparameter set to 0.6. Finally, we order all key-phrase clusters by their frequencies and retain only the top 60% thus removing low-frequency schema elements.

While creating the missing schema, we do the match based on semantic similarity of key-phrases (even for a tuple we only look at key-phrase similarity) and we consider two key-phrases to be matched if the cosine similarity is above a threshold, that we set as 0.8 since we want to match only highly similar key-phrases.

GloVe embeddings on Amazon and Ubuntu datasets We train 200 dimensional GLoVE embeddings on the vocabulary of both Amazon and Ubuntu dataset separately. We set a vocabulary frequency threshold at 50, i.e. we only obtain embeddings for words that appears at least 50 times in the whole corpus.

Datasets Downloadable links to each datasets are provided here: Amazon[10], Ubuntu[11].

Collecting human annotations for usefulness scores For the Amazon dataset, Rao and Daumé III (2019) define the *usefulness* of a question as the degree to which the answer provided by the question would be useful to potential buyers or current users of the product. We use the annotation scheme defined in Rao and Daumé III (2019) to annotate a set of 5000 questions from the amazon dataset.[12] We show annotators product details (title, category, and description) and a question asked about that product and ask them to give it a usefulness score between 0 to 5.[13] Each question was annotated by three annotators. We average the three scores to get a single usefulness score per question. We use the YAKE extractor to extract the schema elements for each question and assign the usefulness score of the question to each of its schema elements.

Since our aim is to assign a usefulness score to each missing element of each product in our dataset, we train a usefulness classifier on the manually annotated schema elements. Although our usefulness score is a real value between 0 and 5, we find that training a regression model gives us poor performance. Hence we convert the real value

[8] We train GLoVE embeddings separately on Amazon and Ubuntu

[9] For triples, we use only their key-phrase to define similarity.

[10] https://nijianmo.github.io/amazon/index.html

[11] https://github.com/rkadlec/ubuntu-ranking-dataset-creator

[12] We use the Amazon Mechanical Turk platform.

[13] Refer Rao and Daumé III (2019) for an exact description of each score.

into a binary value by threshold at 3 (i.e. values below 3 are assigned label 0 and values above 3 are assigned label 1).

Usefulness classification with negative sampling Collecting usefulness annotation on questions, as we do for the Amazon dataset, can be expensive and may not always be possible in different scenarios. Therefore, for the Ubuntu dataset, we experiment with a classifier where instead of using human annotations are true labels, we use a negative sampling strategy. Specifically, we assume that all (context, question) pairs in the Ubuntu dataset can be labelled 1 and any (context, random question) can be labelled 0. We sample a set of 2500 questions from the Ubuntu dataset and them label 1 and sample an equivalent number of negative samples and assign them label 0.

B Training

BART and PPLM For question generation model, we use BART-base (6 encoder layers, 6 decoder layers, 117M parameters) [14]). For PPLM usefulness classifier, we use a bag-of-word model, that uses the pretrained subword embedding layers from BART-base model. We average the subword embeddings to obtain a sentence representation and a usefulness score is predicted via a linear layer projection with softmax. We use the the PPLM code from official repository[15].

Each BART variant converged in 3 epochs on an average with batch size 4 in a TITAN X (Pascal) GPU that took 12 hours in total. While training, we only observe perplexity on the validation set to employ an early-stopping criteria.

Usefulness Classifier for BART+MissInfo+WD We train an SVM (support vector machines) classifier on this data. We use word emebddings as our features by training a 200 dimensional GLoVE model trained on individual dataset. We average the word embeddings of all words in a schema element and use it as a feature. We obtain an F1-score of 80.6% on a held out test set.[16] We use this classifier to predict a usefulness score for each missing schema element of each instances from a class for each dataset, which was required for the BART+MissInfo+WD model.

[14] https://huggingface.co/transformers/model_doc/bart.html

[15] https://github.com/uber-research/PPLM

[16] In comparison, humans get an F1-score of 82.7% in Amazon dataset

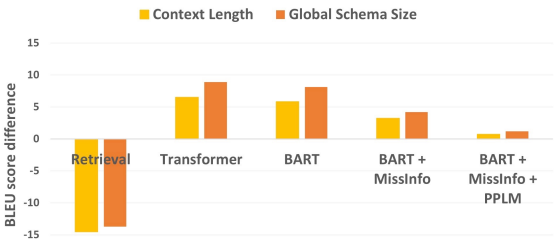


Figure 5: Average BLEU score difference between classes having longer (more than 200 (median) words) and shorter descriptions larger (more than 200 (median) key-phrases) and shorter global schema for Ubuntu dataset. Lower difference indicates more invariance towards information available.

C More Experimental Analysis

We additionally report Krippendorff's alpha, a measure of annotator agreement for our human evaluation, on Amazon dataset.They are : for fluency 0.408, for relevancy 0.177, for missinginfo 0.226, and for usefulness 0.0948. For usefulness, we observe, if the systems are more distinct (GAN-Utility vs BART+missinfo+PPLM), then the agreement is higher i.e. 0.163. For missinginfo, again, 3-way gives higher agreement (0.434), and a probable cause would be that more annotations are going into the undecided category.

Additionally, Figure 5 shows the BLEU difference across different data samples (based on context length and global schema size) that follow a similar trend to Amazon. Table 9 shows generations from all the models, with a case the our best model trades off with missing information to improve the usefulness.

Amazon	
Category	Bookshelf Speakers
Title	Yamaha NS-6490 3-Way Bookshelf Speakers Finish (Pair) Black
Description	Upgrade your current 5.1 home theater to a 7.1-Channel surround sound system by adding a pair of Yamaha NS-6490 bookshelf speakers. This speaker was designed for both professional & home entertainment enthusiasts with the capability to deliver a full, clear,...
Missing Schema	{**speaker wire**, **mounting**, (amplifier, tune, nsubj), wireless, bass, (iPhone, connect, obj), ...}
GAN-Utility	are these speakers compatible with a yamaha satellite speakers?
BART	What are the dimensions?
BART+missinfo	Do the speakers come with **speaker wire**?
BART+missinfo+PPLM	What kind **mounting** does this speaker use?
Category	Camera & Photo
Title	Porta Trace 10 x 12-inches Stainless Steel Frame Lightbox with Two 5000K Lamps
Description	Gagne Porta-Trace light boxes virtually eliminate the hot spots found in competitive lightbox units. Redesigned frame and reflector combine with the thick Plexiglas top to provide uniform and even lighting over the entire durable, stable viewing surface. Durable Stainless Steel frame will maintain its attractive appearance for years...
Missing Schema	{camera, **battery powered**, flash, wireless, canon, nikon, ...}
GAN-Utility	will this work with a canon rebel?
BART	Does it come with the bulbs?
BART+missinfo	Is it **battery powered**?
BART+missinfo+PPLM	Can I replace the bulbs?

Table 9: Model generations for two examples product from Amazon. In the second example, our best model trades off with missing information to make the question more useful.