Abstract

It is of great value to answer product questions based on heterogeneous information sources available on web product pages, e.g., semi-structured attributes, text descriptions, user-provided contents, etc. However, these sources have different structures and writing styles, which poses challenges for (1) evidence ranking, (2) source selection, and (3) answer generation. In this paper, we build a benchmark with annotations for both evidence selection and answer generation covering 6 information sources. Based on this benchmark, we conduct a comprehensive study and present a set of best practices. We show that all sources are important and contribute to answering questions. Handling all sources within one single model can produce comparable confidence scores across sources and combining multiple sources for training always helps, even for sources with totally different structures. We further propose a novel data augmentation method to iteratively create training samples for answer generation, which achieves close-to-human performance with only a few thousand annotations. Finally, we perform an in-depth error analysis of model predictions and highlight the challenges for future research.

1 Introduction

Automatic answer generation for product-related questions is a hot topic in e-commerce applications. Previous approaches have leveraged information from sources like product specifications (Lai et al., 2018a, 2020), descriptions (Cui et al., 2017; Gao et al., 2019) or user reviews (McAuley and Yang, 2016; Yu et al., 2018; Zhang et al., 2019) to answer product questions. However, these works produce answers from only a single source. While a few works have utilized information from multiple sources (Cui et al., 2017; Gao et al., 2019; Feng et al., 2021), they lack a reliable benchmark and have to resort to noisy labels or small-scaled human evaluation (Zhang et al., 2020; Gao et al., 2021). Furthermore, almost none of them make use of pretrained Transformer-based models, which are the current state-of-the-art (SOTA) across NLP tasks (Devlin et al., 2019; Clark et al., 2020).

In this work, we present a large-scale benchmark dataset for answering product questions from 6 heterogeneous sources and study best practices to overcome three major challenges: (1) evidence ranking, which finds most relevant information from each of the heterogeneous sources; (2) source selection, which chooses the most appropriate data source to answer each question; and (3) answer generation, which produces a fluent, natural-sounding answer based on the relevant information. It is necessary since the selected relevant information may not be written to naturally answer a question, and therefore not suitable for a conversational setting.

Most published research on product question answering is based on the AmazonQA dataset (McAuley and Yang, 2016), which takes the community question-answers (CQAs) as the ground truth. This leads to several problems. (1) CQAs, even the top-voted ones, are quite noisy. Many are generic answers or irrelevant jokes (Gao et al., 2021). (2) CQAs are based more on the opinion of the individual customer who wrote the answer rather than on accompanying sources such as product reviews and descriptions. As such, CQAs are not reliable references for judging the quality of answers generated from these sources (Gupta et al., 2019). (3) There are no annotations for assessing the relevance of the information across multiple data sources. This makes it difficult to evaluate the evidence ranker and generator separately. Some works collect annotations for evidence relevance, but only for a single source and with questions formulated post-hoc rather than naturally posed (Lai et al., 2018a; Xu et al., 2019). To address these shortcomings, we collect a benchmark dataset with the following features: (1) It provides clear annotations for both evidence ranking and answer...
generation, enabling us to perform in-depth evaluation of these two components separately. (2) We consider a mix of 6 heterogeneous sources, ranging from semi-structured specifications (jsons) to free sentences and (3) It represents naturally-occurring questions, unlike previous collections that elicited questions by showing answers explicitly.

As sources differ in their volume and contents, collecting training data covering all sources of natural questions and answers is challenging. To get enough positive training signals for each source, we propose filtering community questions based on the model score of a pretrained QA ranker. Questions are only passed for annotation when the confidence scores of top-1 evidence lie within some certain range. This greatly reduces annotation effort by removing most unanswerable questions.

After collecting the data, we apply SOTA Transformer-based models for evidence ranking and answer generation, and present a set of data augmentation and domain adaptation techniques to improve the performance. We show that pretraining the model on the AmazonQA corpus can provide a better initialization and improve the ranker significantly. For evidence ranking, we apply question generation with consistency filtering (Alberti et al., 2019) to obtain large amounts of synthetic QA pairs from unannotated product sources. For answer generation, we propose a novel data augmentation algorithm that creates training examples iteratively. By first training on this augmented data and then finetuning on the human annotations, the model performance can be further enhanced.

As for the model design, we homogenize all sources by reducing them to the same form of input which is fed into a unified pretrained Transformer model, similarly to many recent works of leveraging a unified system for various input formats (Oguz et al., 2020; Su et al., 2020; Komeili et al., 2021). We show that combining all sources within a single framework outperforms handling individual sources separately and that training signals from different answer sources can benefit each other, even for sources with totally different structures. We also show that the unified approach is able to produce comparable scores across different sources which allows for simply using the model prediction score for data source selection, an approach that outperforms more complex cascade-based selection strategies. The resulting system is able to find the correct evidence for 69% of the questions in our test set. For answer generation, 94.4% of the generated answers are faithful to the extracted evidence and 95.5% of them are naturally-sounding.

In summary, our contributions are four-fold: (1) We create a benchmark collections of natural product questions and answers from 6 heterogeneous sources covering 309,347 question-evidence pairs, annotated for both evidence ranking and answer generation. This collection will be released as open source. (2) We show that training signals from different sources can complement each other. Our system can handle diverse sources without source-specific design. (3) We propose a novel data augmentation method to iteratively create training samples for answer generation, which achieves close-to-human performance with only a few thousand annotations and (4) We perform an extensive study of design decisions for input representation, data augmentation, model design and source selection. Error analysis and human evaluation are conducted to suggest directions for future work.

2 Benchmark test set collection

We begin by explaining how we collect a benchmark test set for this problem. The benchmark collection is performed in 4 phases: question sourcing, supporting evidence collection, relevance annotation, and answer elicitation. An annotation example is shown in Table 1.

<table>
<thead>
<tr>
<th>Question: how much weight will it safely hold?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1: Annotation example. <strong>Relevance annotation</strong>: Given one question and evidence from heterogeneous sources, judge if each one is relevant to the question. <strong>Answer elicitation</strong>: annotators produce a natural-sounding answer given the question and the evidence that was marked as relevant.</td>
</tr>
</tbody>
</table>

2.1 Question sourcing

To create a question set that is diverse and representative of natural user questions, we consider two methods of question sourcing. The first method collects questions through Amazon Mechanical Turk, whereby annotators are shown a product image and title and instructed to
ask 3 questions about it to help them make hypothetical purchase decisions. This mimics a scenario in which customers see a product for the first time, and questions collected in this way are often general and exploratory in nature. The second method samples questions from the AmazonQA corpus. These are real customer questions posted in the community forum and tend to be more specific and detailed, since they are usually asked after users have browsed, or even purchased, a product. We then filter duplicated and poorly-formed questions. This yields 914 questions from AmazonQA and 1853 questions from Mturk. These are combined to form the final question set.

Collecting Supporting Evidence We gather “supporting evidence” from 6 heterogeneous sources: (1) Attributes: Product attributes in json format extracted from the Amazon product database. (2) Bullet points: Product summaries from the Amazon product page. (3) Descriptions: Product descriptions from the manufacturer and Amazon. (4) Pub (PUB): Online Publications written about corresponding products (for example here). (5) CQA: Top-voted user-provided answers from the Amazon Customer Questions & Answers. Answers directly replying to questions in our question set are discarded and (6) Review: User reviews written for the product.

Relevance Annotation Annotators are presented with a question about a product and are instructed to mark all the items of supporting evidence that are relevant to answering the product question. Such evidence is defined as relevant if it implies an answer, but it does not need to directly address or answer a question. For evidence items from source 1, we directly present the attribute json to annotators. For sources 2~6, we split the evidence into sentences and present each sentence as a separate item to be considered. There can be a very large number of CQA and Reviews for each product. As manual annotation of these would be impractical, we annotate only the top 40 and 20 evidence from each collection, respectively, as determined by a deep passage ranker pretrained on general-domain QA. Each item of evidence is inspected by 3 annotators and is marked as relevant if supported by at least two of them. In this way, items of evidence are paired with questions for review by annotators. Overall, annotators have inspected 309,347 question-evidence pairs, of which 20,233 were marked as relevant.

Answer Elicitation In the answer elicitation stage, annotators are presented with a question and an item of supporting evidence that has been marked as relevant. They are required to produce a fluent, natural-sounding and well-formed sentence (not short span) that directly answers the question. We sample 500 positive question-evidence pairs from each source for answer elicitation (if that many are available). The annotated answers are evaluated by another round of annotation to filter invalid ones. In the end, we obtain 2,319 question-evidence-answer triples for answer generation.

Table 2 shows the collection statistics. Availability differs across sources. Only 19% of questions have available PUB articles, but all products have corresponding Attributes and Bullet Points. 93.72% of questions are answerable from at least 1 out of the 6 sources, indicating these sources are valuable as a whole to address most user questions.

3 Training data collection

For training data collection, a complete annotation of each set of evidence is not necessary; we need only a rich set of contrastive examples. Therefore, we propose to select questions for annotation based on the confidence score of a pretrained ranker (the same ranker we used to select top evidence for CQA and review). We sample 50k community questions about products in the same domain as the testset. We first select questions whose top-1 item of supporting evidence returned by the pretrained ranker has a prediction score of $>0.8$. In this way the selected questions have a good chance of being answerable from the available evidence and the approach should also yield enough positive samples from all sources to train the ranker. This selection step is crucial to ensure coverage of low-

<table>
<thead>
<tr>
<th>Source</th>
<th>#words</th>
<th>available</th>
<th>answerable</th>
<th>N/P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>5.84</td>
<td>100%</td>
<td>36.10%</td>
<td>22.88</td>
</tr>
<tr>
<td>Bullet</td>
<td>12.55</td>
<td>100%</td>
<td>24.95%</td>
<td>5.39</td>
</tr>
<tr>
<td>Desc</td>
<td>12.86</td>
<td>98.37%</td>
<td>38.59%</td>
<td>23.97</td>
</tr>
<tr>
<td>PUB</td>
<td>17.75</td>
<td>18.98%</td>
<td>4.54%</td>
<td>11.16</td>
</tr>
<tr>
<td>CQA</td>
<td>13.32</td>
<td>99.39%</td>
<td>70.61%</td>
<td>13.85</td>
</tr>
<tr>
<td>Review</td>
<td>18.37</td>
<td>95.64%</td>
<td>61.16%</td>
<td>2.28</td>
</tr>
</tbody>
</table>

93.72% questions are answerable from at least 1 source.
resource sources, like PUB, which otherwise might have zero positive samples. To avoid a selection process that is biased towards easy questions we further include questions whose top-1 evidence has a score within the range of 0.4~0.6. Intuitively these questions will pose more of a challenge in ranking the evidence and their annotation should provide an informative signal.

From each out of the 6 sources, we sample 500 questions with prediction score > 0.8 and another 500 questions with scores in the range of 0.4~0.6. For each question, we then annotate the top-5 (if available) evidence items returned by the pretrained ranker. This reduces annotation cost relative to the complete annotation that was done for the test set. The final dataset contains 6000 questions with 27,026 annotated question-evidence pairs being annotated, 6,667 of which were positive. We then submit the positive question-evidence pairs for answer elicitation. After filtering invalid annotations as was done for the benchmark collection, we obtain a set of 4,243 question-evidence-answer triples to train the answer generator. For both evidence ranking and answer generation, we split the collected data by 9:1 for train/validation.

4 Model

4.1 Evidence Ranking

Evidence ranking aims to get the best evidence from each of the sources. We build our evidence ranker with the Electra-base model (Clark et al., 2020). The question and evidence are concatenated together and fed into the model. We flatten the json structured from the attribute source into a string before feeding it to the encoder, whereas we split evidence from other sources into natural sentences, so it can be encoded as plain text (training detail in appendix D). We present comparison studies in Figure 1 with the best model configuration. Due to space constraints we report only p@1 scores in Fig 1, with full results in appendix C.

Pre-tuning on AmazonQA Pre-tuning the evidence ranker on similar domains has shown to be important when limited in-domain training data is available (Hui and Berberich, 2017; Hazen et al., 2019; Garg et al., 2020; Hui et al., 2022). For our product-specific questions, the AmazonQA corpus is a natural option to pre-tune the model (Lai et al., 2018b). The corpus contains 1.4M question-answer pairs crawled from the CQA forum. We remove answers containing “I don’t know” and “I’m not sure”, and filter questions of more than 32 words and answers of more than 64 words. We construct negative evidence with answers to different questions for the same product. The filtered corpus contains 1,065,407 community questions for training. In the training stage, we first finetune the Electra-base model on the filtered AmazonQA corpus and then finetune on our collected training data. As can be seen, pre-tuning on the AmazonQA corpus improves the p@1 on all sources. The conclusion holds for both training on mixed sources and individual sources separately.

Mixed sources vs split sources We investigate whether different sources conflict with each other by (1) training a single model on the mixed data from all sources, and (2) training a separate model for each individual source. For the second case, we obtain 6 different models, one from each source. The resulting models are tested on 6 sources individually. We can observe that mixing all answer sources into a single training set improves the performance on each individual source. The training signals from heterogeneous sources complement each other, even for sources with totally different structures. p@1 on the semi-structured attribute...
improves consistently through adding training data of unstructured text. This holds for models with and without pre-tuning on AmazonQA.

**Linearization methods** In the above experiment, we use a simple linearization method that flattens the json-formatted attributes into a string. We also compare it with 3 other different linearization methods: (1) key-value pairs: Transform the hierarchical json format into a sequence of key-value pairs. For example, the attribute in Table 1 will be transformed into “item_weight unit pounds | item_weight value 2.2”. (2) templates: Transform the json by pre-defined templates, e.g. “The [attribute_name] of it is [value] [unit]” and (3) NLG: Transform the json into a sentence by a neural data-to-text model. The results show that the best performance is achieved by simply linearizing the json into a string. Although applying the template or neural data-to-text model is closer to a natural sentence, this did not lead to an improvement in p@1. Nonetheless, all these methods have rather similar performance, suggesting the model can adapt quickly to different representations by finetuning on limited training data and that more complex linearization methods are unnecessary.

**Question Generation** Question generation has been a popular data augmentation technique in question-answering. We collect ~50k unannotated pieces of evidence from the 6 sources and apply a question generator to generate corresponding questions. The question generator is finetuned first on the AmazonQA corpus and then on our collected training data. We apply nucleus sampling with p = 0.8 to balance the diversity and generation quality (Sultan et al., 2020). We further filter the generated questions with our evidence ranker by only keeping those with model prediction scores of > 0.5, which has been shown crucial to get high-quality augmented data (Alberti et al., 2019). We try different finetuning methods and report the results on the bottom of Fig 1, where the “+” means the finetuning order. As can be observed, **finetuning on the augmented data brings further improvement to the model.** A three-step finetuning to gradually bring the model to our interested domain leads to the best performance over all sources.

### 4.2 Source Selection

Source aims to select the best source to answer after we obtain the top-1 item of evidence from each source. We show results for the following source selectors: (1) **perfect**: oracle selection of the correct item of evidence (if any) in the top-1 pieces of evidence provided from the 6 sources. (2) **best-score**: evidence item with the highest empirical accuracy in its score range which should yield the upper-bound performance for a selector based on model prediction scores. (3) **highest-score**: evidence with the highest model prediction score. (4) **cascade 1**: prioritizes evidence from the attribute/bullet sources since they have the highest p@1 scores. If the top-1 evidence item from those two sources has a score of more than ϵ, it is selected. Otherwise, the evidence item with the highest prediction score is selected from the remaining sources and (5) **cascade 2**: prioritizes evidence from attribute, bullet, and descriptions sources since these have better official provenance than user-generated data sources. The selection logic is the same as **cascade 1. highest-score** is the most straightforward choice but relies on a comparable score across sources. **cascades 1/2** are also commonly used to merge results from sub-systems. For the **best-score** selector, we split the prediction score range into 100 buckets and estimate the empirical accuracy on the test data. For example the prediction score of 0.924 for the top-1 evidence from an attribute source will fall into the bucket 0.92~0.93. In our test set, evidence items from each source will have an empirical accuracy within each score bin.

Table 3: p@1 using different rankers and source selectors.

<table>
<thead>
<tr>
<th>selector</th>
<th>BM25</th>
<th>AmazonQA</th>
<th>our best</th>
</tr>
</thead>
<tbody>
<tr>
<td>perfect</td>
<td>0.4709</td>
<td>0.7546</td>
<td>0.8338</td>
</tr>
<tr>
<td>best-score</td>
<td>0.2880</td>
<td>0.5370</td>
<td>0.6986</td>
</tr>
<tr>
<td>highest-score</td>
<td>0.2696</td>
<td>0.5089</td>
<td>0.6888</td>
</tr>
<tr>
<td>cascade 1</td>
<td>0.2653</td>
<td>0.5298</td>
<td>0.6791</td>
</tr>
<tr>
<td>cascade 2</td>
<td>0.2638</td>
<td>0.5110</td>
<td>0.6715</td>
</tr>
</tbody>
</table>

As our best “fair” ranker, the highest-score selec-

---

2By continuing to split the confidence range into more buckets we can make an arbitrarily exact approximation to the perfect selector for the test set, but with significant over-fitting.
tor performs remarkably well, with p@1 only 1% lower than that of the best-score-based selectors. It also outperforms the two cascade-based selectors which prioritize official and high-precision sources. This implies the prediction scores across different sources are comparable in our model, which might be because our model is trained on a combination of all sources with the same representation. For the model tuned on AmazonQA, where evidence comes solely from the CQA source, the highest-score selector is not as effective as the cascade selectors. For all rankers, even with the best-score-based selector, there is still a large p@1 gap with the perfect selector, suggesting a further improvement must take into account evidence content, in addition to the prediction scores.

In Figure 2, we visualize the distribution of selected sources by varying the threshold of two cascade-based selectors. We also show the distribution by using the highest-score selector (score) on the left. As the threshold grows, model precision first grows and then degrades, suggesting all sources can contribute to answering product questions. There is no single source that dominates. Although the cascade selection strategy underperforms the highest-confidence selector, it provides us with a flexible way to adjust the source distribution by threshold tuning. In practice, one may want to bias the use of information from official providers, even with a slight reduction in precision.

### 4.3 Answer Generation

After selecting an evidential item from one source, the role of answer generation is to generate a natural-sounding answer based on both the question and the evidence. We build our answer generator with the Bart-large model (Lewis et al., 2020). Similar to the evidence ranker, we take a unified approach for all sources by concatenating both the question and the evidence together (split by the token “|”) as the model input. The model is then fine-tuned on the collected question-evidence-answer (q-e-a) triples. As in training the ranker, we flatten the json structures into strings and process them in the same way as the other sources.

**Mixed sources vs split sources** We experimented with training the generative model on each individual source separately as well as mixing the training data from all sources and training a unified model. We measured the BLEU scores of these systems with results shown in Figure 3, where we also include the results of directly copying the evidence. We can see that training a unified model to handle all sources improves the performance on all sources, as is consistent with our findings in evidence ranking. This is not surprising since previous research on data-to-text has also found that text-to-text generative models are quite robust to different variants of input formats (Kale and Rastogi, 2020; Chang et al., 2021). Directly copying the evidence as the answer leads to very low BLEU scores, especially for json-formatted attributes. This indicates we must significantly rewrite the raw evidence to produce a natural answer.

**Conditional Back-translation (CBT)** In our scenario, the AmazonQA contains a large amount of q-a pairs but these do not have corresponding evidence. We can apply a similar idea as back-translation (Sennrich et al., 2016) but further “condition” on the question. Firstly, we train an evidence generator based on our annotated q-e-a triples. The model is trained to generate the evidence by taking the q-a pairs as input. We then apply the model to generate pseudo-evidence $e'$ from the $q-a$ pairs in AmazonQA. The answer generator is then first finetuned on the pseudo $q-e'-a$ triples and then finetuned further on the real $q-e-a$ annotations. It can be considered as a “conditional” version of back-translation where the model is additionally conditioned on the questions. We use nucleus sampling with p=0.8 to generate the evidence $e'$ since the diversity of inputs is important for back-
translation (Edunov et al., 2018; Zhao et al., 2019). The results are displayed in Table 4. We can see that adding the conditional back-translation step improves the BLEU score by nearly 3 points.

**Noisy Self-training (NST)** Self-training is another popular technique in semi-supervised learning (Scudder, 1965). It uses a trained model to generate outputs for unlabeled data, then uses the generated outputs as the training target. In our scenario, however, the unlabeled input data is not readily available since it requires positive question-evidence pairs. We first apply the same question generation model used for evidence ranking to create “noisy” \( q' - e \) pairs. The current model then generates an answer \( a' \) based on the \( q' - e \) pairs. We use beam search with beam size 5 to generate the answers as the generation quality is more important than diversity in self-training (He et al., 2020). A new model is then initialized from Bart-large, first finetuned on the \( q' - e - a' \) triples, then finetuned on the real training data. We also experimented with adding noise to the input side when training on the \( q' - e - a' \) triples, which has shown to be helpful for the model robustness (He et al., 2020) \(^3\). As shown in Table 4, NST improves the model performance by over 1 BLEU point. Adding the noise to the input further brings slight improvement.

**Iterative Training** We further investigated combining the proposed CBT and NST into an iterative training pipeline. The intuition is that CBT can improve the answer generator which then helps NST to generate higher-quality pseudo answers. The higher-quality triples from NST can in turn be used to ‘warm up’ the evidence generator for CBT. Algorithm 1 details the process. It can be considered a variant of iterative back-translation (Hoang et al., 2018; Chang et al., 2021) with an additional condition on the question and the noisy self-training process inserted in between. It essentially follows a generalized EM algorithm (Shen et al., 2017; Cotterell and Kreutzer, 2018; Graça et al., 2019) where the evidence generator and the answer generator are guaranteed to improve iteratively. We show the results after each iteration in Table 4. As can be seen, the iterative training pipeline further improves generation quality. Most gains are found in the first iteration and the model saturates at iteration 3 with a BLEU score of 34.9.

**Human Evaluation** We run a human evaluation to assess generation quality of our best generator (iteration-3 from Table 4), human reference and the copied evidence. We evaluate from two perspectives: (1) Faithfulness: A sentence is unfaithful to the evidence if it contains extra or contradictory information, and (2) Naturalness: A sentence is unnatural if it is not fluent; contains additional information that not relevant as an answer; or does not directly reply to the question. We show the results after each iteration in Table 4. As can be seen, the iterative training pipeline further improves generation quality. Most gains are found in the first iteration and the model saturates at iteration 3 with a BLEU score of 34.9.

### Table 4: BLEU scores on different methods: copying the input evidence as the answer (copy), finetuning Bart-large on training samples (Bart-large), Bart-large + conditional back-translation (CBT) and Bart-large + noisy self-training (NST).

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy</td>
<td>4.0</td>
<td>47.3</td>
<td>22.4</td>
<td>15.9</td>
<td>12.6</td>
</tr>
<tr>
<td>Bart-large</td>
<td>30.9</td>
<td>57.6</td>
<td>36.1</td>
<td>24.9</td>
<td>17.6</td>
</tr>
<tr>
<td>CBT</td>
<td>33.5</td>
<td>60.3</td>
<td>39.0</td>
<td>27.6</td>
<td>20.5</td>
</tr>
<tr>
<td>NST</td>
<td>32.5</td>
<td>59.5</td>
<td>37.3</td>
<td>26.2</td>
<td>19.2</td>
</tr>
<tr>
<td>NST + noise</td>
<td>33.2</td>
<td>59.8</td>
<td>38.0</td>
<td>26.9</td>
<td>19.9</td>
</tr>
<tr>
<td>Iteration-1</td>
<td>34.3</td>
<td>61.1</td>
<td>39.4</td>
<td>28.0</td>
<td>20.8</td>
</tr>
<tr>
<td>Iteration-2</td>
<td>34.9</td>
<td>61.1</td>
<td>39.8</td>
<td>28.3</td>
<td>21.4</td>
</tr>
<tr>
<td>Iteration-3</td>
<td>34.9</td>
<td>61.3</td>
<td>39.7</td>
<td>28.6</td>
<td>21.6</td>
</tr>
<tr>
<td>Iteration-4</td>
<td>34.7</td>
<td>61.3</td>
<td>39.8</td>
<td>28.3</td>
<td>21.3</td>
</tr>
</tbody>
</table>

### Algorithm 1 (Iterative Training Process):

1. \( G_e = G_a = \text{Bart-large} \);
2. for \( i = 1 \) to \( N \) do
   1. Finetune \( G_e \) on \( \{ q - a - e \}_{\text{real}} \);
   2. Generate \( e' \) with \( G_e \) from \( \{ q - a \}_{\text{AmazonQA}} \);
   3. Finetune \( G_a \) on generated \( \{ q - e' - a \}_{\text{AmazonQA}} \);
   4. Noisy Self-training \( (G_a) \);
   5. Generate \( a' \) with \( G_a \) from \( \{ q' - e \}_{QG} \);
   6. Finetune \( G_e \) on generated \( \{ q' - a' - e \}_{QG} \);
3. end

### Table 5: Human evaluation results.

<table>
<thead>
<tr>
<th>Evaluated</th>
<th>Faithfulness (%)</th>
<th>Naturalness (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>copied evidence</td>
<td>-</td>
<td>15.44</td>
</tr>
<tr>
<td>our best</td>
<td>94.39</td>
<td>95.51</td>
</tr>
<tr>
<td>human reference</td>
<td>97.00</td>
<td>95.82</td>
</tr>
</tbody>
</table>

\(^3\)We apply a similar noise function as in Edunov et al. (2018) that randomly deletes replaces a word by a filler token with probability 0.1, then swaps words up to the range of 3.
### Subjectivity of Relevance

**Question:** what is the snow made of?
**Predicted:** kind of cotton like.

<table>
<thead>
<tr>
<th>Noise in Attribute Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Question:</strong> is it dark green or brown?</td>
</tr>
<tr>
<td><strong>Predicted:</strong> color: { value: &quot;as shown&quot; }</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overfit to String Match</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q:</strong> what is the length of the solder?</td>
</tr>
<tr>
<td><strong>Predicted:</strong> coil of solder is about 1-3/4 long and ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Uncertain Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q:</strong> can it be machine washed?</td>
</tr>
<tr>
<td><strong>Predicted:</strong> not sure if it is machine washable.</td>
</tr>
</tbody>
</table>

Table 6: Common errors of evidence relevance ranking.

### Number accuracy

**Q:** what’s the size of the house?
**E:** item_dimensions: {width:..., height: {unit: centimeters, normalized_value: {unit: inches, value: 15.35}, value: 39.}}
**G:** The house is 10.24 \times 30.31 \times 39 inches

<table>
<thead>
<tr>
<th>Hallucination when Inference needed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q:</strong> are the dolls with it?</td>
</tr>
<tr>
<td><strong>E:</strong> included_components: {value: &quot;my dreamy doll-house&quot;}</td>
</tr>
<tr>
<td><strong>G:</strong> Yes, it comes with a doll.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sensitivity to typographical errors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q:</strong> for he many players?</td>
</tr>
<tr>
<td><strong>E:</strong> ages 7 and up, 2-12 players</td>
</tr>
<tr>
<td><strong>G:</strong> this game is for ages 7 and up.</td>
</tr>
</tbody>
</table>

Table 7: Common errors of answer generations.

Error types 1 and 3 could be alleviated through data augmentation. We can create new samples to let the model learn to copy numbers properly and learn to be robust to common typos. Another way to reduce number sensitivity could to delexicalize numbers in the inputs, a common strategy in data to text generation (Wen et al., 2015; Gardent et al., 2017). Error type 2 is a challenging open problem in neural text generation. Many techniques have been proposed such as learning latent alignment (Shen et al., 2020), data refinement with NLU (Nie et al., 2019), etc. These could potentially be applied to our task, which we leave for future work.

### 4.4 Best Practices

To summarize the best practices, the attribute json strings can be directly flattened and all sources are mixed together and trained with a single unified encoder. The ranker is finetuned on AmazonQA, augmented data obtained by question generation and manually annotated training data in order. Source selection can be performed based solely on the model confidence score and the answer generator can be trained as in Algorithm 1.

### 5 Error analysis

Based on the human evaluation, we identified the following key problems that exist in the current system. For evidence ranking, the major problems are:

(1) **Subjectivity of relevance:** It can be subjective to define whether a piece of evidence is enough to answer a given question. The model will sometimes pick a somewhat relevant piece of evidence, even though there could be other, better options that support a more comprehensive answer.

(2) **Noise in attribute value:** When an attribute value contains uninformative data due to the noise of data sources, the model still may choose it based on its attribute name.

(3) **Overfitting to string match:** The model tends to select strings similar to the question while ignoring their fine semantics, a common problem from the bias to ‘shortcut learning’ of neural networks (Geirhos et al., 2020).

(4) **Uncertain evidence:** The model ranks evidence highly, even if this evidence is an uncertain expression. This can be viewed as a special case of over-fitting to string match. We show examples in Table 6. We can attempt to alleviate errors of type 1 by providing finer-grained labels in the training data instead of only binary signals (Gupta et al., 2019). Error types 2 and 4 could be mitigated by data augmentation, constructing negative samples by corrupting the attribute values or making evidence uncertain.

Error type 3 is more challenging. One possible solution is to automatically detect spurious correlations and focus the model on minor examples (Tu et al., 2020). Nevertheless, a fundamental solution to fully avoid Error 3 is still an open question.

For answer generation, we identify the major problems as:

(1) **Number accuracy:** The model cannot fully understand the roles of numbers from the limited training examples.

(2) **Hallucination if inference is needed:** when it is not possible to generate an answer by simple rephrasing, the model can hallucinate false information.

(3) **Sensitivity to typos:** The model is not robust to typos in the question. A tiny typo can easily break the system.

We provide examples of these errors in Table 7. Error types 1 and 3 could be alleviated through data augmentation. We can create new samples to let the model learn to copy numbers properly and learn to be robust to common typos. Another way to reduce number sensitivity could to delexicalize numbers in the inputs, a common strategy in data to text generation (Wen et al., 2015; Gardent et al., 2017). Error type 2 is a challenging open problem in neural text generation. Many techniques have been proposed such as learning latent alignment (Shen et al., 2020), data refinement with NLU (Nie et al., 2019), etc. These could potentially be applied to our task, which we leave for future work.

### 6 Conclusion

To the best of our knowledge, this work is the first comprehensive study of product answer generation from heterogeneous sources including both semi-
structured attributes and unstructured text. We collect a benchmark dataset with annotations for both evidence ranking and answer generation. It will be released to benefit relevant study. We find that the best practice is to leverage a unified approach to handle all sources of evidence together and further experimented with a set of data augmentation techniques to improve the model performance. Error analysis is provided to illustrate common errors, which we hope will lead to inspire future work.

References


Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Peihao Su, David Vandyke, and Steve J Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In EMNLP.


A Collected Data

All our collected data have also been manually verified to remove sample with private or offensive information.

In Figure 4, we show the ngram distribution of question prefixes in our collected data. As can be seen, a large proportion of questions are boolean questions starting with “is”, “does”, “can”, “are”, “do” and “will”. The rest are mostly factual questions like “how many/tall/long ...” and “what ...”. Most of them should be able to answer with a short span since there are not many opinion questions like “how is ...”, “why ...”.

B Instruction for Human Annotation

All annotators are based on the US. We first perform in-house annotation and then estimate the time needed for each annotation. We then set the payment to be roughly 15 USD per hour. The payment is decided based on the average payment level in the US. All annotators are informed that their collection will be made public for scientific research according to the Amazon Mechanical Turk code of rules. The data collection protocol has been approved by an ethics review board.

B.1 Question Collection

Read the given product name and image, imagine you are a customer and are recommended this product. Write one question about it to decide whether or not to purchase this product.

Examples of questions: is it energy efficient? does it require a hub? can I watch sports on this TV? will the plug work with an extension cord?
B.2 Evidence Selection

At the start of each task, the workflow application will present a product, a question about the product and a set of candidates which describe the product. Your annotation task is to mark the proper candidate that contains information to answer the question from the attribute set. If none of the provided candidates contain the information, select "None of the above".

B.3 Answer Generation

Read the raised product question and provided information, write a natural, informative, complete sentence to answer this question. If the provided information cannot address the question, write "none". Make sure the answer is a natural, informative and complete sentence. Do not write short answers like "Yes", "Right", "It is good", etc. Provide enough information to help the asker understand more about the question. If the provided information can only partially answer the question, only reply to the answerable part.

**Good Examples:**

- question: what age range is this product designed for?
  
  Provided information: age_range_description: value:"3 - 8 years
  
  Answer: It is designed for the age range of 3 - 8 years old.

- question: how many people can play at one time?
  
  provided information: number_of_players: value:"8
  
  answer: It is designed for 8 players at one time.

**Bad Examples:**

- question: what age range is this product designed for?
  
  Provided information: age_range_description: value:"3 - 8 years
  
  Answer: 3 - 8 years.

- question: how many people can play at one time?
  
  provided information: number_of_players: value:"8
  
  answer: 8.

C Full Results of Ranker

We show the full results of our best-performed ranker in Table 8. As can be seen, different sources have different accuracy score. The attribute and bullet point source have the highest accuracy score because the former is more structured, and the latter has a consistent writing style with only a few sentences. User reviews also have a high accuracy score. This might be because the candidates of reviews are already the top ones selected by our pretrained ranker. Many of them are already relevant and the negative-positive ratio is low. The model does not have extreme difficulty in handling the user reviews. The model performs worst on the description, PUB and CQA answer source. This might result from the diversity of their writing styles and the high negative-positive ratio, which increase the difficulty. Moreover, these two sources usually depend more on the context to interpret the evidence than other sources. The text description is extracted from the multi-media web page. Simply extracting the text part might lose richer context to interpret the extracted text. Similarly, the CQA usually depends on the community question. If we only extract a sentence from the answer, it might contains references that is not self-contained.

<table>
<thead>
<tr>
<th>Source</th>
<th>MAP</th>
<th>MRR</th>
<th>NDCG</th>
<th>P@1</th>
<th>HIT@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>0.965</td>
<td>0.966</td>
<td>0.974</td>
<td>0.943</td>
<td>0.996</td>
</tr>
<tr>
<td>Bullet</td>
<td>0.935</td>
<td>0.935</td>
<td>0.952</td>
<td>0.890</td>
<td>0.993</td>
</tr>
<tr>
<td>Description</td>
<td>0.648</td>
<td>0.708</td>
<td>0.747</td>
<td>0.611</td>
<td>0.822</td>
</tr>
<tr>
<td>PUB</td>
<td>0.667</td>
<td>0.708</td>
<td>0.763</td>
<td>0.579</td>
<td>0.873</td>
</tr>
<tr>
<td>Review</td>
<td>0.796</td>
<td>0.860</td>
<td>0.875</td>
<td>0.778</td>
<td>0.966</td>
</tr>
<tr>
<td>CQA</td>
<td>0.643</td>
<td>0.750</td>
<td>0.766</td>
<td>0.636</td>
<td>0.897</td>
</tr>
</tbody>
</table>

Table 8: Performance of our best ranker on different sources.

D Training details

For both the generative Bart-large model and the discriminative Electra-base model, we truncate the total input length to 128 subword tokens and select the learning rate from $5e^{-6}, 1e^{-5}, 3e^{-5}, 5e^{-5}, 1e^{-4}$. The warm-up step is selected from $[5\%, 10\%, 20\%, 50\%]$ of the whole training steps. For the discriminative model, we choose the best configuration based on the F1 score on the validation set. For the generative model, we choose the best configuration based on the perplexity on the validation set. In the end, we set the learning rate of Electra-base as $3e^{-5}$ and that of Bart-large as $1e^{-5}$. The warm-up step is set as 20% for Electra-base and 10% for Bart-large. The batch size is set as 64 for Electra-base and 16 for Bart-large. For Electra-base, we measure the validation F1 score after finishing every 1% of the whole training steps and stop the model when the validation F1 score does not increase for 30% of the whole training steps. For Bart-large, we measure the validation
loss every 200 steps and stop the model when the validation loss stops decreasing for 1000 steps. All models are trained once on 8 Nvidia V100 GPUs and the random seed is set as 42.