7-2023

Product question answering in e-commerce: A survey

Yang DENG
Singapore Management University, ydeng@smu.edu.sg

Wenxuan ZHANG

Qian YU

Wai LAM

Citation

DENG, Yang; ZHANG, Wenxuan; YU, Qian; and LAM, Wai. Product question answering in e-commerce: A survey. (2023). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, 2023 July 9-14*. 11951-11964.

Available at: https://ink.library.smu.edu.sg/sis_research/9132

Product Question Answering in E-Commerce: A Survey

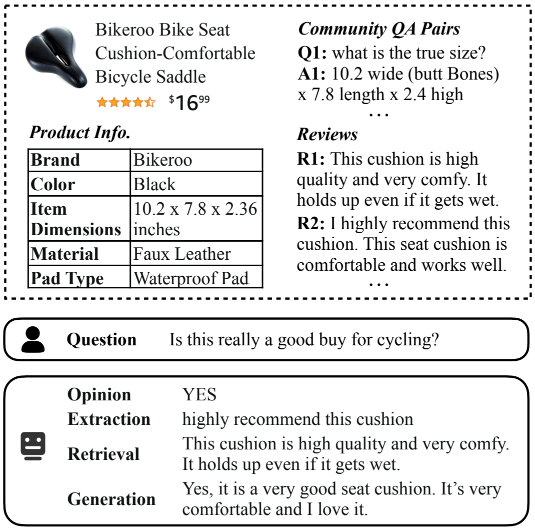
Yang Deng[1], **Wenxuan Zhang**[2,†], **Qian Yu**[3], **Wai Lam**[1]
[1] The Chinese University of Hong Kong, [2] DAMO Academy, Alibaba Group, [3] JD.com
{dengyang17dydy,isakzhang}@gmail.com, yuqian81@jd.com, wlam@se.cuhk.edu.hk

Abstract

Product question answering (PQA), aiming to automatically provide instant responses to customer's questions in E-Commerce platforms, has drawn increasing attention in recent years. Compared with typical QA problems, PQA exhibits unique challenges such as the subjectivity and reliability of user-generated contents in E-commerce platforms. Therefore, various problem settings and novel methods have been proposed to capture these special characteristics. In this paper, we aim to systematically review existing research efforts on PQA. Specifically, we categorize PQA studies into four problem settings in terms of the form of provided answers. We analyze the pros and cons, as well as present existing datasets and evaluation protocols for each setting. We further summarize the most significant challenges that characterize PQA from general QA applications and discuss their corresponding solutions. Finally, we conclude this paper by providing the prospect on several future directions.

1 Introduction

E-Commerce is playing an increasingly important role in our daily life. During the online shopping, potential customers inevitably have some questions about their interested products. To settle down their concerns and improve the shopping experience, many AI conversational assistants have been developed to solve customers' problems, such as Alexa (Carmel et al., 2018) and AliMe (Li et al., 2017a). The core machine learning problem underlying them, namely **Product Question Answering (PQA)**, thus receives extensive attention in both academia and industries recently. Figure 1 depicts an actual PQA example from Amazon. There are a tremendous amount of product-related data available within the product page, which contains natural language user-generated content (UGC) (*e.g.*, product reviews, community QA pairs), structured product-related information (*e.g.*, attribute-value pairs), images, etc. Generally, PQA aims to automatically answer the customer-posted question in the natural language form about a specific product, based on the product-related data.



Bikeroo Bike Seat Cushion-Comfortable Bicycle Saddle

Product Info.

Brand	Bikeroo
Color	Black
Item Dimensions	10.2 x 7.8 x 2.36 inches
Material	Faux Leather
Pad Type	Waterproof Pad

Community QA Pairs

Q1: what is the true size?
A1: 10.2 wide (butt Bones) x 7.8 length x 2.4 high
...

Reviews

R1: This cushion is high quality and very comfy. It holds up even if it gets wet.
R2: I highly recommend this cushion. This seat cushion is comfortable and works well.
...

Question Is this really a good buy for cycling?

Opinion YES
Extraction highly recommend this cushion
Retrieval This cushion is high quality and very comfy. It holds up even if it gets wet.
Generation Yes, it is a very good seat cushion. It's very comfortable and I love it.

Figure 1: An PQA example from Amazon.

Typical QA studies (Rajpurkar et al., 2016) and some other domain-specific QA studies (*e.g.*, biomedical QA (Jin et al., 2023) and legal QA (Gil, 2021)) mainly focus on the questions that ask for a certain factual and objective answer. Differently, product-related questions in PQA typically involve consumers' opinion about the products or aspects of products. Therefore, early studies (Moghaddam and Ester, 2011; Yu et al., 2012) regard PQA as a special opinion mining problem, where the answers are generated by aggregating opinions in the retrieved documents. Most of recent works essentially follow the same intuition, but formulate PQA as different problems in terms of the form of target answers. Accordingly, existing PQA studies

[*] The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14200719).

[†] Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11951–11964
July 9-14, 2023 ©2023 Association for Computational Linguistics

	Method	Document	Extra Data	Backbone	Main Challenge	Dataset	Pros&Cons
Opinion	McAuley and Yang (2016)	PR	-	Feature	Subjectivity	Amazon	**Pro**: tackle a large proportion of questions that ask for certain opinion by using comparatively simple methods. **Con**: only classify the opinion polarity without detailed info.
	Wan and McAuley (2016)	PR	-	Feature	Subjectivity	Amazon	
	Yu and Lam (2018b)	PR	-	Feature	Subjectivity	Amazon	
	Fan et al. (2019)	PR	-	NN	-	Amazon	
	Zhang et al. (2019)	PR	-	PLM	-	Amazon	
	Rozen et al. (2021)	PR	QA	PLM	Low-resource	Amazon+	
Extraction	Gupta et al. (2019)	PR	-	NN	Answerability	AmazonQA	**Pro**: provide pinpointed answers. **Con**: providing an incomplete answer is less user-friendly.
	Xu et al. (2019)	PR	MRC	PLM	Low-resource	ReviewRC	
	Bjerva et al. (2020)	PR	-	NN/PLM	Subjectivity	SubjQA	
Retrieval	Cui et al. (2017)	PR+QA+PI	-	NN	Multi-type Resources	-	**Pro**: select complete and informative sentences as the answer, based on actual customer experience. **Con**: may not answer the given question precisely since the supporting document (*e.g.*, reviews) is not specifically written for answering the given question.
	Yu et al. (2018b)	PR+QA	-	Feature	Low-resource	Amazon+	
	Yu et al. (2018a)	QA	NLI	NN	Low-resource	-	
	Kulkarni et al. (2019)	PR+QA+PI	-	NN	Multi-type Resources	-	
	Chen et al. (2019a)	PR	QA	NN	Low-resource	Amazon+	
	Zhao et al. (2019)	PR	QA	NN	Interpretability	Amazon	
	Zhang et al. (2020c)	QA	PR	NN	Answerability	Amazon	
	Zhang et al. (2020f)	PR+PI	QA	NN	Multi-type Resources	Amazon+	
	Mittal et al. (2021)	QA	CQA	PLM	Low-resource	-	
	Roy et al. (2022b)	PR	QA	PLM	Low-resource	-	
Generation	Chen et al. (2019c)	PR	-	NN	-	Taobao	**Pro**: provide natural forms of answers, which are specific to the given questions and flexible with different information. **Con**: suffer from hallucination and factual-inconsistency issues, and lack of effective automatic evaluation methods.
	Gao et al. (2019)	PR+PI	-	NN	Multi-type Resources	JD	
	Deng et al. (2020)	PR	-	NN	Subjectivity	Amazon	
	Lu et al. (2020)	PR	-	PLM	Subjectivity	AmazonQA	
	Gao et al. (2021)	PR+PI	-	NN	Multi-type Resources	JD	
	Feng et al. (2021)	PR+PI	-	NN	Multi-type Resources	JD	
	Deng et al. (2022)	PR+PI	-	NN	Personalization	Amazon	
	Shen et al. (2022b)	PI	-	PLM	Multi-type Resources	semiPQA	

Table 1: Summary of PQA studies. "Amazon+" denotes that additional annotations are added into the "Amazon" dataset. "PR", "QA", and "PI" denote product reviews, community QA pairs, and product information, respectively.

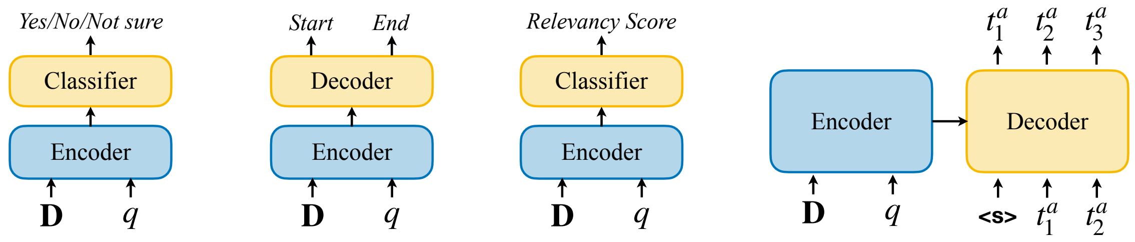
can be categorized into four types: opinion-based, extraction-based, retrieval-based, and generation-based. As shown in Figure 1, opinion-based PQA approaches only provide the common opinion polarity as the answer, while extraction-based PQA approaches extract specific text spans from the supporting documents as the answer. Retrieval-based PQA approaches further re-rank the documents to select the most appropriate one to answer the given question, while generation-based PQA approaches generate natural language sentences based on the available documents as the response. In this paper, we systematically review methods of these four mainstream PQA problem settings, as well as the commonly-used datasets and evaluation protocols.

Besides the task-specific challenges in each type of PQA systems, there are several common challenges across all types of PQA systems, which differentiate PQA from other QA systems. (1) **Subjectivity**. Subjective questions constitute a large proportion of questions in PQA, which requires to aggregate the crowd's opinions about the questions, reflected through related reviews and QAs. (2) **Reliability & Answerability**. Different from those supporting documents constructed by professionals in biomedical or legal QA, product reviews and community QA pairs come directly from non-expert users, which may suffer from some typical flaws as other UGC, such as redundancy, inconsistency, spam, and even malice. (3) **Multi-type resources**. The supporting documents usually consist of heterogeneous information from multi-type data resources, such as text, table, knowledge graph, image, etc. (4) **Low-resource**. PQA systems often encounter the low-resource issue, since different product categories may need different training data, and it is generally time-consuming and costly to manually annotate sufficient labeled data for each domain. Accordingly, we introduce existing solutions to each challenge.

To our knowledge, this survey is the first to focus on Product Question Answering. We first systematically summarize recent studies on PQA into four problem settings as well as introduce the available datasets and corresponding evaluation protocols in Section 2. Then we analyze the most significant challenges that characterize PQA from other QA applications and discuss their corresponding solutions in Section 3. Finally, we discuss several promising research directions for future PQA studies and conclude this paper in Section 4 and 5.

2 Problems and Approaches

Product question answering (PQA) aims to produce an answer a to a given natural language question q based on a set of supporting documents D,



(a) Opinion-based (b) Extraction-based (c) Retrieval-based (d) Generation-based

Figure 2: Four main-stream problem settings in PQA studies.

where the supporting documents can be product reviews, community QA pairs, product information, etc. In terms of the form of provided answers, we systematically categorize the existing PQA studies into four problem settings, including Opinion-based PQA, Extraction-based PQA, Retrieval-based PQA, Generation-based PQA, and introduce corresponding approaches proposed to solve the problem, as summarized in Table 1. We present an overview of the general framework for each problem setting in Figure 2. In addition, the key information of the datasets adopted in existing PQA studies is summarized in Table 2.

2.1 Opinion-based PQA

Opinion-based PQA studies focus on yes-no type questions, *i.e.*, questions that can be answered by “Yes” or “No”, which constitute a large proportion on PQA platforms.

2.1.1 Problem Definition

Given a product-related question q and a set of supporting documents $\mathbf{D}$ (product reviews in most cases), the goal is to predict a binary answer $a \in \{\text{Yes}, \text{No}\}$. Some studies also consider the neutral answer, *e.g.*, “Not Sure".

2.1.2 Datasets & Evaluation Protocols

One of the largest and widely-adopted public PQA datasets is the Amazon Product Dataset (denoted as “Amazon” in Table 1 and hereafter), composed by Amazon Question/Answer Data (McAuley and Yang, 2016; Wan and McAuley, 2016) and Amazon Review Data (He and McAuley, 2016; Ni et al., 2019). It consists of around 1.4 million answered questions and 233.1 million product reviews across over 20 different product categories. The Amazon dataset contains the information of question types (“yes-no" or “open-ended"), answer types (“yes", “no", or “not sure"), helpful votes by customers, and product metadata, which is suitable for opinion-based PQA evaluation.

Due to the existence of a certain proportion of unanswerable questions based on the available reviews, it is difficult to achieve an acceptable performance with the ordinary classification accuracy metric $\text{Acc}(\mathcal{Q})$ for any method. Therefore, McAuley and Yang (2016) propose Acc@k, which has become the de facto metric for evaluating opinion-based PQA methods, which only calculates the classification accuracy of top-k questions ranked by the prediction *confidence*. The *confidence* with each classification is its distance from the decision boundary, *i.e.*, $|\frac{1}{2} - P(a|q, \mathbf{D})|$. A good model is supposed to assign high confidence to those questions that can be correctly addressed.

$$\text{Acc@}k = \text{Acc}(\operatorname*{arg\,max}_{\mathcal{Q}' \in \mathcal{P}_k(\mathcal{Q})} \sum_{q \in \mathcal{Q}'} |\frac{1}{2} - P(a|q, \mathbf{D})|) \quad (1)$$

where $\mathcal{P}_k(\mathcal{Q})$ is the set of k-sized subsets of $\mathcal{Q}$, and k is commonly set to be 50% of the total number of questions.

2.1.3 Methods

McAuley and Yang (2016) propose a Mixtures of experts (MoEs) (Jacobs et al., 1991) based model, namely Mixtures of Opinions for Question Answering (Moqa), to answer yes-no questions in PQA, where each review is regarded as an “expert” to make a binary prediction for voting in favor of a “yes” or “no” answer. The confidence of each review is further weighted by its relevance to the question as follows:

$$P(a|q, \mathbf{D}) = \sum_{d \in \mathbf{D}} \underbrace{P(d|q)}_{\text{how relevant is } d} \cdot \underbrace{P(a|d, q)}_{\text{prediction from } d} \quad (2)$$

Moqa is later enhanced by modeling the ambiguity and subjectivity of answers and reviews (Wan and McAuley, 2016). Yu and Lam (2018b) further improve Moqa by computing the aspect-specific embeddings of reviews and questions via a three-order

Dataset	Language	Answer Form	# Questions	# Categories	Types of Doc.	Additional Info.	Release
Amazon (McAuley and Yang, 2016)	English	Yes-No/Open-ended	~1.4M	21	PR/PI/QA	Timestamps/User/Vote	✓[1]
AmazonQA (Gupta et al., 2019)	English	Yes-No/Open-ended	~923K	17	PR	Answerability	✓[2]
ReviewRC (Xu et al., 2019)	English	Span	2,596	2	PR	Sentiment	✓[3]
SubjQA (Bjerva et al., 2020)	English	Span/Open-ended	10,098	6	PR	Subjectivity	✓[4]
JD (Gao et al., 2019)	Chinese	Open-ended	469,955	38	PR/PI	-	✓[5]
Taobao (Chen et al., 2019c)	Chinese	Open-ended	1,155,530	2	PR	-	×
semiPQA (Shen et al., 2022b)	English	Open-ended	11,243	-	PI	-	×
PAGHS* (Shen et al., 2022a)	English	Open-ended	309,347	-	PR/PI/QA	Relevance of Docs.	×

* PAGHS stands for Product Answer Generation from Heterogeneous Source as there is no specific name for the dataset proposed in Shen et al. (2022a).

Table 2: Summary of existing datasets for product question answering.

auto-encoder network in an unsupervised manner. In these early studies, the features either extracted by heuristic rules or acquired from unsupervised manners may limit the performance and application of opinion-based PQA approaches.

To better model the relation between the question and each review, Fan et al. (2019) and Zhang et al. (2019) explore the utility of neural networks (*e.g.*, BiLSTM (Schuster and Paliwal, 1997)) and pretrained language models (*e.g.*, BERT (Devlin et al., 2019)) to learn the distributed feature representations, which largely outperform previous methods. Recently, Rozen et al. (2021) propose an approach, called SimBA (**Sim**ilarity **B**ased **A**nswer Prediction), which leverages existing answers from similar resolved questions about similar products to predict the answer for the target question.

2.1.4 Pros and Cons

Opinion-based PQA approaches can tackle a large proportion of product-related questions that ask for certain opinion by using comparatively simple and easy-to-deploy methods. However, opinion-based approaches could only provide the classification result of the opinion polarity, based on the common opinion reflected in the supporting documents, without detailed and question-specific information.

2.2 Extraction-based PQA

Similar to typical extraction-based QA (Rajpurkar et al., 2016) (also called Machine Reading Comprehension (MRC)), extraction-based PQA studies aim at extracting a certain span of a document to be the answer for the given product-related questions.

2.2.1 Problem Definition

Given a product-related question q and a supporting document $d = \{t_1, ..., t_n\} \in \mathbf{D}$, which consists of one or more product reviews, the goal is to find a sequence of tokens (a text span) $a = \{t_s, ..., t_e\}$ in d that answers q correctly, where $1 \leq s \leq n$, $1 \leq e \leq n$, and $s \leq e$.

2.2.2 Datasets & Evaluation Protocols

Xu et al. (2019) build the first extraction-based PQA dataset, called ReviewRC, using reviews from SemEval-2016 Task 5 (Pontiki et al., 2016). Similarly, Gupta et al. (2019) conduct extensive pre-processing on the Amazon dataset (McAuley and Yang, 2016; He and McAuley, 2016) to build a dataset for extraction-based PQA, called AmazonQA. It annotates each question as either answerable or unanswerable based on the available reviews, and heuristically creates an answer span from the reviews that best answer the question. Bjerva et al. (2020) propose SubjQA dataset to investigate the relation between subjectivity and PQA in the context of product reviews, which contains 6 different domains that are built upon TripAdvisor (Wang et al., 2010), Yelp[6], and Amazon (McAuley and Yang, 2016) datasets.

Given the same setting as typical MRC, extraction-based PQA adopts the same evaluation metrics, including Exact Match (EM) and F1 scores. EM requires the predicted answer span to exactly match with the human annotated answer, while F1 score is the averaged F1 scores of individual answers in the token-level.

2.2.3 Methods

Due to the limited training data for extraction-based PQA, Xu et al. (2019) employ two popular pretraining objectives, *i.e.*, masked language modeling and next sentence prediction, to post-train the BERT encoder on both the general MRC dataset, SQuAD (Rajpurkar et al., 2016), and E-Commerce review datasets, including Amazon Review (He and McAuley, 2016) and Yelp datasets. In real-world applications, there will be a large number

[3] http://deepx.ucsd.edu/public/jmcauley/qa/

[4] https://github.com/amazonqa/amazonqa

[5] https://howardhsu.github.io/

[6] https://github.com/megagonlabs/SubjQA

[7] https://github.com/gsh199449/productqa

[6] https://www.yelp.com/dataset

of irrelevant reviews and the question might be unanswerable. To this end, Gupta et al. (2019) first extract top review snippets for each question based on IR techniques and build an answerability classifier to identify unanswerable questions based on the available reviews. Then, a span-based QA model, namely R-Net (Wang et al., 2017), is adopted for the extraction-based PQA. Besides, Bjerva et al. (2020) develop a subjectivity-aware QA model, which performs the multi-task learning of the extraction-based PQA and subjectivity classification. Experimental results show that incorporating subjectivity effectively boosts the performance.

2.2.4 Pros and Cons

Extraction-based PQA approaches can provide pinpointed answers to the given questions, but it may be less user-friendly to provide an incomplete sentence to users and may also lose some additional information. Since there are a large proportion of questions that ask for certain user experiences or opinions based on the statistics in (McAuley and Yang, 2016; Deng et al., 2022), extraction-based paradigm is less practical and favorable in real-world PQA applications. Therefore, it can be observed that there are relatively few works in extraction-based PQA studies in recent years.

2.3 Retrieval-based PQA

Retrieval-based PQA studies treat PQA as an answer (sentence) selection task, which retrieves the best answer from a set of candidates to appropriately answer the given question.

2.3.1 Problem Definition

Given a question q and a set of supporting documents $\mathbf{D}$, the goal is to find the best answer a by ranking the list of documents according to the relevancy score between the question q and each document $d \in \mathbf{D}$, *i.e.*, $a = \arg\max_{d \in \mathbf{D}} \mathcal{R}(q, d)$.

2.3.2 Datasets & Evaluation Protocols

Due to the absence of ground-truth question-review (QR) pairs, several efforts (Chen et al., 2019a; Yu et al., 2018b; Zhang et al., 2020f) have been made on annotating additional QR pairs into the Amazon dataset for retrieval-based PQA. Nevertheless, the original Amazon dataset can be directly adopted for retrieval-based PQA studies (Zhang et al., 2020e,c) that aim to select reliable or helpful answers from candidate community answers.

Since the retrieval-based PQA methods are essentially solving a ranking problem, most studies adopt standard ranking metrics for evaluation, including mean average precision (MAP), mean reciprocal rank (MRR), and normalized discounted cumulative gain (NDCG).

2.3.3 Methods

Cui et al. (2017) first demonstrate a retrieval-based PQA chatbot, namely SuperAgent, which contains different ranking modules that select the best answer from different data sources within the product page, including community QA pairs, product reviews, and product information. Kulkarni et al. (2019) further propose a pipeline system that first classifies the question into one of the predefined question categories with a question category classifier, and then uses an ensemble matching model to rank the candidate answers. However, these systems usually contain multiple modules with different purposes, which require a large amount of annotated data from different sources. Therefore, most recent retrieval-based PQA works use one or two sources as the supporting documents and build the model in an end-to-end manner.

When facing a newly posted product-related question, a straight-forward answering strategy is to retrieve a similar resolved question and provide the corresponding answer to the target question. However, such a solution relies heavily on a large amount of domain-specific labeled data, since QA data differs significantly in language characteristics across different product categories. To handle the low-resource issue, Yu et al. (2018a) propose a general transfer learning framework that adapts the shared knowledge learned from large-scale paraphrase identification and natural language inference datasets (*e.g.*, Quora[7] and MultiNLI (Williams et al., 2018)) to enhance the performance of reranking similar questions in retrieval-based PQA systems. Besides, Mittal et al. (2021) propose a distillation-based distantly supervised training algorithm, which uses QA pairs retrieved by a syntactic matching system, to help learn a robust question matching model.

Another approach to obtain answers for new questions is to select sentences from product reviews. The main challenge is that the information distributions of explicit answers and review contents that can address the corresponding questions are quite different and there are no annotated ground-truth question-review (QR) pairs which can

[7] https://www.kaggle.com/c/quora-question-pairs

be used for training. Yu et al. (2018b) develop a distant supervision paradigm for incorporating the knowledge contained in QA collections into question-based response review ranking, where the top ranked reviews are more relevant to the QA pair and are useful for capturing the knowledge of response review ranking. Chen et al. (2019a) propose a multi-task deep learning method, namely QAR-net, which can exploit both user-generated QA data and manually labeled QR pairs to train an end-to-end deep model for answer identification in review data. Zhao et al. (2019) aim at improving the interpretability of retrieval-based PQA by identifying important keywords within the question and associating relevant words from large-scale QA pairs. Zhang et al. (2020f) employ pre-trained language models (*e.g.*, BERT) to obtain weak supervision signals from the community QA pairs for measuring the relevance between the question and heterogeneous information, including natural language reviews and structured attribute-value pairs.

For the situation where multiple user-generated answers have already been posted, Zhang et al. (2020c) propose an answer ranking model, namely MUSE, which models multiple semantic relations among the question, answers, and relevant reviews, to rank the candidate answers in PQA platforms.

2.3.4 Pros and Cons

Retrieval-based approaches select complete and informative sentences as the answer, which may not answer the given question precisely since the supporting document (*e.g.*, reviews) is not specifically written for answering the given question.

2.4 Generation-based PQA

Inspired by successful applications of sequence-to-sequence (Seq2seq) models on other natural language generation tasks, several attempts have been made on leveraging Seq2seq model to automatically generate natural sentences as the answer to the given product-related question.

2.4.1 Problem Definition

Given a product-related question q and a set of supporting documents $\mathbf{D}$ that are relevant to the given question, the goal is to generate a natural language answer $a = \{t_1^a, t_2^a, ...\}$ based on the question q and supporting documents $\mathbf{D}$.

2.4.2 Datasets & Evaluation Protocols

The Amazon dataset can be directly adopted for generation-based PQA. Another popular dataset used for geneartive PQA is from JD (Gao et al., 2019), which is one of the largest e-commerce websites in China. In total, the JD dataset contains 469,953 products and 38 product categories, where each QA pair is associated with the reviews and attributes of the corresponding product.

Evaluating generation-based methods often involves both automatic evaluation and human evaluation. Common automatic evaluation metrics include (i) ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) for evaluating lexical similarity between generated answers and ground-truth answers, (ii) Embedding-based Similarity (Forgues et al., 2014), BertScore (Zhang et al., 2020b), and BleuRT (Sellam et al., 2020) for evaluating semantic relevance, (iii) Distinct scores (Li et al., 2016) for evaluating the diversity of the generated answers. Human evaluation protocols are designed for evaluating different perspectives of the generated answer by human annotations, such as fluency, consistency, informativeness, helpfulness, etc.

2.4.3 Methods

Generation-based PQA studies typically regard the retrieval of relevant documents as a pre-processing step, and build the method upon the retrieved documents. Due to the noisy nature of retrieved documents, Gao et al. (2019) employ a Wasserstein distance based adversarial learning method to denoise the irrelevant information in the supporting reviews, while Chen et al. (2019c) design an attention-based weighting strategy to highlight the relevant words appearing in the retrieved review snippets. Besides identifying relevant information from the retrieved documents, Deng et al. (2020) find that the rich personal opinion information in product reviews also attaches great importance in generation-based methods, as there are a large number of subjective questions in PQA. To this end, a joint learning model of answer generation and opinion mining is proposed to generate opinion-aware answers. Likewise, Lu et al. (2020) propose a cross-passage hierarchical memory network to identify the most prominent opinion across different reviews for answer generation in PQA.

Some recent works focus on leveraging documents from multi-type resources to generate the answer. Feng et al. (2021) model the logical relation between unstructured documents (reviews) and structured documents (product attributes) with a heterogeneous graph neural network. Gao et al. (2021) aim at solving the safe answer problem during the generation (*i.e.*, neural models tend to generate meaningless and general answers), by systematically modeling product reviews, product attributes, and answer prototypes. Shen et al. (2022b) propose present the semiPQA dataset to benchmark PQA over semi-structured data.

2.4.4 Pros and Cons

Generation-based methods can provide natural forms of answers specific to the given questions. However, the hallucination and factual inconsistency issues are prevalent in generation-based methods. In addition, it is still lack of robust automatic evaluation protocols for generation-based methods.

3 Challenges and Solutions

Although the aforementioned PQA methods are developed based on different problem settings, there are some common challenges in PQA, as presented in Table 1. Several main challenges and their corresponding solutions are summarized as follows.

3.1 Subjectivity

Different from typical QA whose answers are usually objective and unique, a large proportion of questions in PQA platforms are asking for subjective information or opinions. Meanwhile, the UGC in E-commerce such as product reviews also provides rich information about other customers' opinion. Therefore, early studies regard PQA as a special opinion mining problem (Moghaddam and Ester, 2011; Yu et al., 2012), which is followed by recent opinion-based PQA studies (McAuley and Yang, 2016; Wan and McAuley, 2016). Ideal answers to this kind of questions require information describing personal opinions and experiences. There are two specific challenges in exploiting such subjective information to facilitate PQA:

- **Detect question-related opinion**. A common solution is to regard the question as the target aspect for aspect-based opinion extraction. For example, Bjerva et al. (2020) use OpineDB (Li et al., 2019c) and some syntactic extraction patterns to extract opinion spans. Deng et al. (2020) employ a dual attention mechanism to highlight the question-related information in reviews for the joint learning with an auxiliary opinion mining task. Zhang et al. (2021) study aspect-based sentiment analysis in PQA, which classifies the sentiment polarity towards certain product aspects in the question from the community answers.
- **Aggregate diverse opinion information**. Since users may differ in opinions towards the same question, a good PQA system should avoid expressing a random opinion, or even being contradictory to the common opinion. To this end, Deng et al. (2020) employ an opinion self-matching layer and design two kinds of opinion fusion strategies to uncover the common opinion among multiple reviews for generation-based PQA. Likewise, Lu et al. (2020) propose a cross-passage hierarchical memory network to identify the most prominent opinion. However, existing studies pay little attention on resolving conflicting user opinions, which is a common issue in opinion summarization of product reviews (Pecar, 2018; Suhara et al., 2020) and worth exploring in the future studies of PQA.

3.2 Answer Reliability & Answerability

Similar to other UGC, product reviews and community answers in E-commerce sites, which are also provided by online users instead of professionals, vary significantly in their qualities and inevitably suffer from some reliability issues such as spam, redundancy, and even malicious content. Therefore, it is of great importance to study the answer reliability and answerability issue when building automatic PQA systems using these UGC. In terms of the availability of candidate answers, existing solutions can be categorized into two groups:

- **Reliability of user-generated answers**. When there are a set of candidate user-generated answers for the concerned question, the reliability measurement of these answers has been investigated from different perspectives. For example, Zhang et al. (2020e) predict the helpfulness of user-generated answers by investigating the opinion coherence between the answer and crowds' opinions reflected in the reviews, while Zhang et al. (2020d) tackle the veracity prediction of the user-generated answers for factual questions as an evidence-based fact checking problem. However, these studies mainly focus on the content reliability while neglecting the reliability degree of the answerer (Li et al., 2017b, 2020).
- **Unanswerable questions based on the available documents**. Question answerability detection has drawn extensive attention in typical QA studies (Rajpurkar et al., 2018). Similarly, Gupta et al. (2019) train an binary classifier to classify the question answerability for PQA. Zhang et al.

(2020a) propose a conformal prediction based framework to reject unreliable answers and return *nil* answers for unanswerable questions. Meanwhile, the answerablity in PQA is also highly related to the reliability of product reviews (Roy et al., 2022a; Shen et al., 2022a).

3.3 Multi-type Resources

Another characteristic of PQA is the necessity of processing heterogeneous information from multi-type resources, including natural language UGC (*e.g.*, reviews, community QA pairs), structured product information (*e.g.*, attribute-value pairs (Lai et al., 2018; Roy et al., 2020), knowledge graph (Li et al., 2019a)), E-manuals (Nandy et al., 2021), images, etc. Early works (Cui et al., 2017; Kulkarni et al., 2019) design separated modules to handle the questions that require different types of data resources. However, these PQA systems rely heavily on annotated data from different types of resources and neglect the relation among heterogeneous data. Therefore, some recent studies focus on manipulating heterogeneous information from multi-type resources in a single model for better answering product-related questions. For instance, Zhang et al. (2020f) design a unified heterogeneous encoding scheme that transforms structured attribute-value pairs into a pesudo-sentence. Gao et al. (2019) employ a key-value memory network to store and encode product attributes for answer decoding with the encoded review representations, which is further combined with answer prototypes (Gao et al., 2021). Feng et al. (2021) propose a heterogeneous graph neural network to track the information propagation among different types of information for modeling the relational and logical information.

3.4 Low-resource

Since there are a large amount of new questions posted in PQA platforms every day and the required information to answer the questions varies significantly across different product categories (even across different single products), traditional supervised learning methods become data hungry in this situation. However, it is time-consuming and labor-intensive to obtain sufficient domain-specific annotations. Existing solutions typically leverage external resources to mitigate the low-resource issue. In terms of the external resources, these solutions can be categorized into two groups:

- **Transfer learning from out-domain data**. This group of solutions typically leverages large-scale open-domain labeled datasets and design appropriate TL strategy for domain adaptation in PQA. For example, Yu et al. (2018a) transfer the knowledge learned from Quora and MultiNLI datasets to retrieval-based PQA models, by imposing a regularization term on the weights of the output layer to capture both the inter-domain and the intra-domain relationships. Xu et al. (2019) perform post-training on the SQuAD dataset to inject task-specific knowledge into BERT for extraction-based PQA.

- **Distant supervision from in-domain data**. Another line of solutions adopt the resolved QA pairs from similar products (Rozen et al., 2021) or products in the same categories (Yu et al., 2018b; Chen et al., 2019a; Zhao et al., 2019; Roy et al., 2022b) as weak supervision signals. For example, Zhang et al. (2020f) and Mittal et al. (2021) employ syntactic matching systems (*e.g.*, BM25) or pre-trained text embeddings (*e.g.*, BERT) to obtain resolved QA pairs for facilitating the distantly supervised training process.

4 Prospects and Future Directions

Considering the challenges summarized in this paper, we point out several promising prospects and future directions for PQA studies:

- **Question Understanding**. Due to the diversity of product-related questions, some attempts have been made on identifying the user's intents (Yu and Lam, 2018a), the question types (Cui et al., 2017), and even the user's purchase-state (Kuchy et al., 2021) from the questions. In addition, some researches investigate the user's uncertainty or the question's ambiguity towards the product by asking clarifying questions (Majumder et al., 2021; Zhang and Zhu, 2021). Despite the extensive studies for QA, question understanding has not been deeply studied in the context of PQA. For example, the system should be capable of identifying the subjectivity from the product-related questions (Bjerva et al., 2020), such as opinionated questions (Deng et al., 2020), comparative questions (Bondarenko et al., 2022), etc.

- **Personalization**. As mentioned before, compared with typical QA studies (Rajpurkar et al., 2016), there is a large proportion of subjective questions (McAuley and Yang, 2016) on PQA platforms, which involve user preference or require personal information to answer, rather than objective or factoid questions that look for a certain answer. Besides, in E-Commerce, different customers often have certain preferences over product aspects or information needs (Chen et al., 2019b; Li et al., 2019b), leading to various expectations for the provided answers. Therefore, Carmel et al. (2018) state that a good PQA system should answer the customer's questions with the context of her/his encounter history, taking into consideration her/his preference and interest. Such personalization can make the answer more helpful for customers and better clarify their concerns about the product (Deng et al., 2022).

- **Multi-modality**. Compared with the widely-studied natural language UGC and structured product knowledge data, image data has received little attention in PQA studies. On E-Commerce sites, there exist not only a great number of official product images, but also increasing user-shared images about their actual experiences, which benefit many other E-Commerce applications (Liu et al., 2021; Zhu et al., 2020). The multimodal data can provide more valuable and comprehensive information for PQA systems.
- **Datasets and Benchmarks**. Despite the increasing attentions on developing PQA systems, the publicly available resources for PQA are still quite limited. Most existing PQA studies are evaluated on the Amazon dataset (McAuley and Yang, 2016), which is directly crawled from the Amazon pages. Some researches (Roy et al., 2022a; Shen et al., 2022a) have discussed several drawbacks of evaluating PQA systems on this dataset: 1) The ground-truth answers are quite noisy, since they are the top-voted community answers posted by non-expert users. 2) There are no annotations for assessing the relevance of the supporting documents, which may cast potential risks on the reliability of the PQA systems. To facilitate better evaluations, many other data resources for PQA studies have been constructed as presented in Table 2. However, due to the privacy or the commercial issues, some of the datasets cannot be publicly released. Therefore, there is still a great demand for a large-scale, high-quality, and publicly available benchmark dataset for the future studies on PQA.
- **Evaluation Protocols**. The types of questions vary in a wide range, from yes-no questions to open-ended questions (McAuley and Yang, 2016), from objective questions to subjective questions (Bjerva et al., 2020), from factual questions to non-factual questions (Zhang et al., 2020d). Different types of questions may involve different specific evaluation protocol. For example, it is necessary to evaluate the precision of opinion in the answers for subjective questions (Deng et al., 2020), while the veracity or factualness is important in factual questions (Zhang et al., 2020d). Especially for generation-based PQA methods, the evaluation is still largely using lexical-based text similarity metrics, which are not correlated well with human judgements.

5 Conclusions

This paper makes the first attempt to overview recent advances on PQA. We systematically categorize recent PQA studies into four problem settings, including Opinion-based, Extraction-based, Retrieval-based, and Generation-based, and summarize the existing methods and evaluation protocols in each category. We also analyze the typical challenges that distinguish PQA from other QA studies. Finally, we highlight several potential directions for facilitating future studies on PQA.

Limitations

Since product question answering (PQA) is actually a domain-specific application in general QA, the scope of the problem may be limited. However, in recent years, PQA has received increasing attention in both academy and industry. (1) From the research perspective, PQA exhibits some unique characteristics and thus brings some interesting research challenges as discussed in Section 3. For example, some studies use PQA as an entrypoint to analyze the subjectivity in QA tasks. (2) From the application perspective, it has great commercial value. Online shopping is playing an increasingly important role in everyone's daily life, so that many high-tech companies develop AI conversational assistants for promptly solving customer's online problems, including but not limited to Amazon, eBay, Alibaba, JD, etc. Regarding the large amount of research efforts that have been made, there is not a systematic and comprehensive review about this research topic. Similar to recent surveys of other domain-specific QA, such as biomedical QA (Jin et al., 2023) and legal QA (Gil, 2021), we hope that this paper can serve as a good reference for people working on PQA or beginning to work on PQA, as well as shed some light on future studies on PQA and raise more interests from the community for this topic.

References

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. Subjqa: A dataset for subjectivity and review comprehension. In *EMNLP 2020*, pages 5480–5494.

Alexander Bondarenko, Yamen Ajjour, Valentin Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. 2022. Towards understanding and answering comparative questions. In *WSDM 2022*, pages 66–74.

David Carmel, Liane Lewin-Eytan, and Yoelle Maarek. 2018. Product question answering using customer generated content - research challenges. In *SIGIR 2018*, pages 1349–1350.

Long Chen, Ziyu Guan, Wei Zhao, Wanqing Zhao, Xiaopeng Wang, Zhou Zhao, and Huan Sun. 2019a. Answer identification from product reviews for user questions by multi-task attentive networks. In *AAAI 2019*, pages 45–52.

Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019b. Towards knowledge-based personalized product description generation in e-commerce. In *KDD 2019*, pages 3040–3050.

Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019c. Review-driven answer generation for product-related questions in e-commerce. In *WSDM 2019*, pages 411–419.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. Superagent: A customer service chatbot for e-commerce websites. In *ACL 2017, System Demonstrations*, pages 97–102.

Yang Deng, Yaliang Li, Wenxuan Zhang, Bolin Ding, and Wai Lam. 2022. Toward personalized answer generation in e-commerce via multi-perspective preference modeling. *ACM Trans. Inf. Syst.*, 40(4):87:1–87:28.

Yang Deng, Wenxuan Zhang, and Wai Lam. 2020. Opinion-aware answer generation for review-driven question answering in e-commerce. In *CIKM 2020*, pages 255–264.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, pages 4171–4186.

Miao Fan, Chao Feng, Mingming Sun, Ping Li, and Haifeng Wang. 2019. Reading customer reviews to answer product-related questions. In *SDM 2019*, pages 567–575.

Yue Feng, Zhaochun Ren, Weijie Zhao, Mingming Sun, and Ping Li. 2021. Multi-type textual reasoning for product-aware answer generation. In *SIGIR 2021*, pages 1135–1145.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *NeurIPS, modern machine learning and natural language processing workshop*, volume 2, page 168.

Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2021. Meaningful answer generation of e-commerce question-answering. *ACM Trans. Inf. Syst.*, 39(2):18:1–18:26.

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In *WSDM 2019*, pages 429–437.

Jorge Martínez Gil. 2021. A survey on legal question answering systems. *CoRR*, abs/2110.07333.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. Amazonqa: A review-based question answering task. In *IJCAI 2019*, pages 4996–5002.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW 2016*, pages 507–517.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2023. Biomedical question answering: A survey of approaches and challenges. *ACM Comput. Surv.*, 55(2):35:1–35:36.

Lital Kuchy, David Carmel, Thomas Huet, and Elad Kravi. 2021. “did you buy it already?", detecting users purchase-state from their product-related questions. In *SIGIR 2021*, pages 1249–1258.

Ashish Kulkarni, Kartik Mehta, Shweta Garg, Vidit Bansal, Nikhil Rasiwasia, and Srinivasan H. Sengamedu. 2019. Productqna: Answering user questions on e-commerce product pages. In *WWW 2019*, pages 354–360.

Tuan Manh Lai, Trung Bui, Sheng Li, and Nedim Lipka. 2018. A simple end-to-end question answering model for product information. In *ECONLP@ACL 2018*, pages 38–43.

Feng-Lin Li, Weijia Chen, Qi Huang, and Yikun Guo. 2019a. Alime KBQA: question answering over structured knowledge for e-commerce customer service. In *CCKS 2019*, pages 136–148.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, Guwei Jin, and Wei Chu. 2017a. *AliMe Assist* : An intelligent assistant for creating an innovative e-commerce experience. In *CIKM 2017*, pages 2495–2498.

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

Piji Li, Zihao Wang, Lidong Bing, and Wai Lam. 2019b. Persona-aware tips generation? In *WWW 2019*, pages 1006–1016.

Yaliang Li, Nan Du, Chaochun Liu, Yusheng Xie, Wei Fan, Qi Li, Jing Gao, and Huan Sun. 2017b. Reliable medical diagnosis from crowdsourcing: Discover trustworthy answers from non-experts. In *WSDM 2017*, pages 253–261.

Yanying Li, Haipei Sun, and Wendy Hui Wang. 2020. Towards fair truth discovery from biased crowdsourced answers. In *KDD 2020*, pages 599–607.

Yuliang Li, Aaron Feng, Jinfeng Li, Saran Mumick, Alon Y. Halevy, Vivian Li, and Wang-Chiew Tan. 2019c. Subjective databases. *Proc. VLDB Endow.*, 12(11):1330–1343.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Junhao Liu, Zhen Hai, Min Yang, and Lidong Bing. 2021. Multi-perspective coherent reasoning for helpfulness prediction of multimodal reviews. In *ACL/IJCNLP 2021*, pages 5927–5936.

Junru Lu, Gabriele Pergola, Lin Gui, Binyang Li, and Yulan He. 2020. CHIME: cross-passage hierarchical memory network for generative review question answering. In *COLING 2020*, pages 2547–2560.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian J. McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *NAACL-HLT 2021*, pages 4300–4312.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *WWW 2016*, pages 625–635.

Happy Mittal, Aniket Chakrabarti, Belhassen Bayar, Animesh Anant Sharma, and Nikhil Rasiwasia. 2021. Distantly supervised transformers for e-commerce product QA. In *NAACL-HLT 2021*, pages 4008–4017.

Samaneh Moghaddam and Martin Ester. 2011. AQA: aspect-based opinion question answering. In *ICDMW 2011*, pages 89–96.

Abhilash Nandy, Soumya Sharma, Shubham Maddhashiya, Kapil Sachdeva, Pawan Goyal, and Niloy Ganguly. 2021. Question answering over electronic devices: A new benchmark dataset and a multi-task learning based QA framework. In *Findings of ACL: EMNLP 2021*, pages 4600–4609.

Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP-IJCNLP 2019*, pages 188–197.

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

Samuel Pecar. 2018. Towards opinion summarization of customer reviews. In *ACL 2018, Student Research Workshop*, pages 1–8.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *SemEval@NAACL-HLT 2016*, pages 19–30.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL 2018*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP 2016*, pages 2383–2392.

Kalyani Roy, Vineeth Balapanuru, Tapas Nayak, and Pawan Goyal. 2022a. Investigating the generative approach for question answering in E-commerce. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 210–216.

Kalyani Roy, Avani Goel, and Pawan Goyal. 2022b. Effectiveness of data augmentation to identify relevant reviews for product question answering. In *Companion Proceedings of the Web Conference 2022*, page 298–301.

Kalyani Roy, Smit Shah, Nithish Pai, Jaidam Ramtej, Prajit Prashant Nadkarn, Jyotirmoy Banerjee, Pawan Goyal, and Surender Kumar. 2020. Using large pretrained language models for answering user queries from product specifications. *CoRR*.

Ohad Rozen, David Carmel, Avihai Mejer, Vitaly Mirkis, and Yftah Ziser. 2021. Answering product-questions by utilizing questions from other contextually similar products. In *NAACL-HLT 2021*, pages 242–253.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *ACL 2020*, pages 7881–7892.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, Bill Byrne, and Adrià de Gispert. 2022a. Product answer generation from heterogeneous sources: A new benchmark and best practices. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 99–110.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, and Adrià Gispert. 2022b. semiPQA: A study on product question answering over semi-structured data. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 111–120.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. Opiniondigest: A simple framework for opinion summarization. In *ACL 2020*, pages 5789–5798.

Mengting Wan and Julian J. McAuley. 2016. Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *ICDM 2016*, pages 489–498.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *SIGKDD 2010*, pages 783–792.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *ACL 2017*, pages 189–198.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT 2018*, pages 1112–1122.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *NAACL-HLT 2019*, pages 2324–2335.

Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018a. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *WSDM 2018*, pages 682–690.

Jianxing Yu, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Answering opinion questions on products by exploiting hierarchical organization of consumer reviews. In *EMNLP-CoNLL 2012*, pages 391–401.

Qian Yu and Wai Lam. 2018a. Product question intent detection using indicative clause attention and adversarial learning. In *ICTIR 2018*, pages 75–82.

Qian Yu and Wai Lam. 2018b. Review-aware answer prediction for product-related questions incorporating aspects. In *WSDM 2018*, pages 691–699.

Qian Yu, Wai Lam, and Zihao Wang. 2018b. Responding e-commerce product questions via exploiting QA collections and reviews. In *COLING 2018*, pages 2192–2203.

Shiwei Zhang, Jey Han Lau, Xiuzhen Zhang, Jeffrey Chan, and Cécile Paris. 2019. Discovering relevant reviews for answering product-related queries. In *ICDM 2019*, pages 1468–1473.

Shiwei Zhang, Xiuzhen Zhang, Jey Han Lau, Jeffrey Chan, and Cécile Paris. 2020a. Less is more: Rejecting unreliable reviews for product question answering. In *ECML-PKDD 2020*, pages 567–583.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *ICLR 2020*.

Wenxuan Zhang, Yang Deng, and Wai Lam. 2020c. Answer ranking for product-related questions via multiple semantic relations modeling. In *SIGIR 2020*, pages 569–578.

Wenxuan Zhang, Yang Deng, Xin Li, Lidong Bing, and Wai Lam. 2021. Aspect-based sentiment analysis in question answering forums. In *Findings of ACL: EMNLP 2021*, pages 4582–4591.

Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam. 2020d. Answerfact: Fact checking in product question answering. In *EMNLP 2020*, pages 2407–2417.

Wenxuan Zhang, Wai Lam, Yang Deng, and Jing Ma. 2020e. Review-guided helpful answer identification in e-commerce. In *WWW 2020*, pages 2620–2626.

Wenxuan Zhang, Qian Yu, and Wai Lam. 2020f. Answering product-related questions with heterogeneous information. In *AACL/IJCNLP 2020*, pages 696–705.

Zhiling Zhang and Kenny Q. Zhu. 2021. Diverse and specific clarification question generation with keywords. In *WWW 2021*, pages 3501–3511.

Jie Zhao, Ziyu Guan, and Huan Sun. 2019. Riker: Mining rich keyword representations for interpretable product question answering. In *KDD 2019*, pages 1389–1398.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for e-commerce product. In *EMNLP 2020*, pages 2129–2139.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section

☒ A2. Did you discuss any potential risks of your work?
It's a survey paper. There is no potential risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.