Cross-Market Product-Related Question Answering

Negin Ghasemi
Radboud University
Nijmegen, The Netherlands
negin.ghasemitaheri@ru.nl

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Hamed Bonab*
Amazon Inc.
Seattle, WA, USA
hamedrab@amazon.com

Evangelos Kanoulas
University of Amsterdam
Amsterdam, The Netherlands
e.kanoulas@uva.nl

Arjen P. de Vries
Radboud University
Nijmegen, The Netherlands
arjen.devries@ru.nl

James Allan
University of Massachusetts Amherst
Amherst, MA, USA
allan@cs.umass.edu

Djoerd Hiemstra
Radboud University
Nijmegen, The Netherlands
djoerd.hiemstra@ru.nl

ABSTRACT

Online shops such as Amazon, eBay, and Etsy continue to expand their presence in multiple countries, creating new resource-scarce marketplaces with thousands of items. We consider a marketplace to be resource-scarce when only limited user-generated data is available about the products (e.g., ratings, reviews, and product-related questions). In such a marketplace, an information retrieval system is less likely to help users find answers to their questions about the products. As a result, questions posted online may go unanswered for extended periods. This study investigates the impact of using available data in a resource-rich marketplace to answer new questions in a resource-scarce marketplace, a new problem we call *cross-market question answering*. To study this problem's potential impact, we collect and annotate a new dataset, XMarket-QA, from Amazon's UK (resource-scarce) and US (resource-rich) local marketplaces. We conduct a data analysis to understand the scope of the cross-market question-answering task. This analysis shows a temporal gap of almost one year between the first question answered in the UK marketplace and the US marketplace. Also, it shows that the first question about a product is posted in the UK marketplace only when 28 questions, on average, have already been answered about the same product in the US marketplace. Human annotations demonstrate that, on average, 65% of the questions in the UK marketplace can be answered within the US marketplace, supporting the concept of cross-market question answering. Inspired by these findings, we develop a new method, CMJim, which utilizes product similarities across marketplaces in the training phase for retrieving answers from the resource-rich marketplace that can be used to answer a question in the resource-scarce marketplace. Our evaluations show CMJim's significant improvement compared to competitive baselines.

CCS CONCEPTS

• **Information systems** → **Question answering**; • **Computing methodologies** → *Transfer learning*.

KEYWORDS

Cross-Market Question Answering, Product-related Question Answering, Similar Question Retrieval

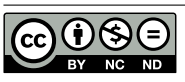
ACM Reference Format:
Negin Ghasemi, Mohammad Aliannejadi, Hamed Bonab, Evangelos Kanoulas, Arjen P. de Vries, James Allan, and Djoerd Hiemstra. 2023. Cross-Market Product-Related Question Answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591658

1 INTRODUCTION

Online shops have become more popular than ever, with an impressive increase in suppliers and customers [16] in many countries. The rapid growth in suppliers can confuse customers, who face an overwhelming number of options when they visit an online shop, leading to many questions about their potential purchases. Unsurprisingly, most e-commerce sites are equipped with a customer questions and answers (Q&A) section for customers to ask questions about the products to help them make a purchase decision. Answering these questions helps customers make informed decisions and benefits suppliers by increasing the number of purchases. Previous work shows that various sources of information on a *resource-rich* marketplace can be leveraged to answer product-related questions on the same marketplace. Examples include product specifications and details [14, 18, 25], reviews [6, 9, 14, 22, 34, 36–38], and the answers given to similar questions for similar products [29].

Online shops like Amazon, eBay, and Etsy operate in many countries (i.e., marketplaces); they regularly expand their operations and sales to new marketplaces in new regions. However, user interaction data is limited in such *resource-scarce* new marketplaces, leading

*Work done prior to joining Amazon.



SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591658

to many products with few reviews and answered questions. Additionally, with relatively few active customers on the e-commerce platform, the number of individuals seeking help outnumbers those answering questions, resulting in long waiting times before questions are answered [1]. This raises the question of whether information from a resource-rich marketplace can complement the information in a resource-scarce marketplace. To exemplify the scenario, imagine a customer who wants to buy an Apple Watch Series 8 from the **uk** marketplace and wants to know if the watch battery has been improved since the last version. They may pose this question on `Amazon.co.uk`'s question-answering section: "How is the battery life compared to Apple Watch 7?" While the question is unanswered on the **uk** marketplace, a similar answered question on the same product exits on `Amazon.com`: "Does the battery life improve compared to the last watch?" with the answer "Yes, the watch battery is much improved compared to the older versions." This question–answer pair would answer the customer's question in the **uk** marketplace.

In this work, we are particularly interested in taking the first step towards modeling the task of *Cross-Market* Question Answering, which is defined as finding relevant answers to a question posted in a resource-scarce main marketplace, utilizing data from a resource-rich auxiliary marketplace, alongside the main marketplace itself. To this end, we first aim to answer the following research question **RQ1**: *Can we utilize a resource-rich marketplace for the task of* cross-market *question answering and to what extent?* To answer this question and facilitate research in this area, we construct a large-scale, real-life Q&A dataset, termed XMarket-QA, based on two different marketplaces: Amazon's **us** marketplace and its **uk** counterpart. XMarket-QA comprises 30,218 unique products in 16 categories, with over 4.8 million question–answer pairs. We further conduct extensive data analysis to answer the research question, **RQ1**, and determine if the data in the **us** marketplace can potentially be leveraged to answer questions in the **uk** marketplace. Our analysis reveals a notable temporal gap (on average 302 days) between the first question that is answered in the **uk** marketplace compared to the **us** marketplace. Also, we see that in our dataset, 70% of questions in **us** are posted before the first question is answered in **uk** (an average of 28 questions per item), indicating the high potential and significance of using the **us** marketplace when not enough data is available in the **uk** marketplace. Moreover, we complement XMarket-QA with human-annotated relevance judgments, assessing the relevance of questions, where we find that, on average, 65% of the relevant answers originate in the auxiliary marketplace (**us**).

Leveraging cross-market data can be challenging as it poses new opportunities as well as challenges. One could argue that augmenting the resource-scarce marketplace data with additional data from the resource-rich marketplace can solve the problem. However, as argued in the cross-market recommendation literature [4, 28], users with different geographical demographics exhibit different behavior that should be considered. We hypothesize that data from the auxiliary (resource-rich) marketplace can improve performance in the main (resource-scarce) marketplace. The goal is to identify the optimal method for extracting relevant answers to new questions. Towards this aim, we examine both single-market and cross-market baselines while considering two main approaches: (i) identifying similar questions for the identical item in the auxiliary marketplace and (ii) identifying similar items and jointly ranking their questions in both the main and auxiliary marketplaces based on the target question. This leads us to our second research question, **RQ2**: *How can we leverage the unique features of cross-market items to enhance product-related question answering?* We focus on the core discriminating feature of the cross-market data: *the exact same item can appear in the other marketplace.* We propose a new model, called CMJim,[1] to jointly rank both items and questions, using a resource-scarce main marketplace and a resource-rich auxiliary marketplace. CMJim consists of a bi-encoder that learns to predict the similarity of two items, not only based on their titles but also on their question–answer pairs. Therefore, it can jointly rank items and questions, learning from their shared knowledge space. We use the *exact-matching items* as positive training samples — data that is unique to the cross-market setting and not available when learning a model for the single-market setting.

Results show that CMJim outperforms competitive baselines over the main marketplace's unanswered questions. In particular, we see that CMJim outperforms the state-of-the-art SimBA [29] product question-answering model by 14%, indicating that leveraging cross-market exact-matching items during training is effective. On the other hand, our results show that using only exact-matching items in inference is not effective, because of the complexity of the problem, such as having market-sensitive questions, which can only be answered using the questions of similar items in the same marketplace.Finally, we are interested in answering research question **RQ3**: *To what extent can question answering in a resource-scarce marketplace benefit from the auxiliary resource-rich marketplace? Moreover, how sensitive is this approach to the amount of data available in the auxiliary marketplace?* Our experimental results show that leveraging the data in a resource-rich marketplace can lead to improved performance; however, the data availability impacts models differently. In particular, CMJim can still outperform the single-market baseline when only 40% of the **us** data is used to train the model, while other models need at least 80% of the **us** data to be saturated under the same setting.

We summarize the contributions of this work as follows:[2]

- We introduce the novel task of cross-market question answering and collect a large-scale real-world dataset of 4.8 million question–answer pairs from Amazon **us** and Amazon **uk**.
- We create a test set consisting of 94 questions with graded relevance judgments for 2430 answers and 2300 pair items with graded relevance judgments about their similarity.
- We conduct an extensive analysis of our data collection and relevance judgments, shedding light on the cross-market question-answering problem and revealing notable data characteristics.
- We propose CMJim, a model that leverages the unique features of the cross-market question-answering task. CMJim learns to rank both items and questions jointly using a resource-scarce marketplace and a resource-rich marketplace.
- Via extensive experiments, we compare the performance of CMJim with six competitive baselines and analyze the results from various angles.

[1] Pronounced as SimJim

[2] Data and code: https://github.com/neginghasemi/XMarket-QA

2 RELATED WORK

Product-related Question Answering (PQA) aims to answer consumers' general questions using various product-related resources, including catalogs, customer reviews, and the existing Q&A sections on a retail platform. The PQA task has introduced novel challenges that have been studied extensively by the research community, thanks to the availability of relevant public datasets [22]. Similar to the Q&A literature [5], possible approaches to PQA can be categorized into abstraction, extraction, and approaches that retrieve answers from existing PA data. In this work, we mainly focus on retrieval-based approaches. To some good degree, both abstraction and extraction-based approaches rely on retrieval-based solutions as the core technology [17].

Among retrieval-based methods, McAuley and Yang [22] presents one of the earliest PQA studies. They introduce a model that uses previously answered questions to automatically detect whether a product review is relevant to a given question. In an extended work, Wan and McAuley [34] combine ambiguity and personalized factors to improve their model's ability. In another study, Yu and Lam [36] considers the latent aspects and aspect-specific embeddings of reviews to improve performance. Following McAuley and Yang [22], more recent work by Zhang et al. [37] uses BERT [10], a pre-trained language model, to address the language mismatch between user queries and reviews. This work has been extended by Zhang et al. [38], where they focused on the issue of answerability in user questions.

There is a different line of previous studies [7, 9, 13, 14] which is different from our task. These studies generate a text as a related answer instead of ranking the existing answers or reviews. Most of the works mentioned rely on a rich set of user reviews; however, this assumption differs for resource-scarce marketplaces or new products. Some existing works for collecting and ranking related reviews using similar product reviews for new products with no reviews, such as work by Pourgholamali [26] and Park et al. [25]; however, none of these address PQA.

Cross-market PQA shares similarities with the Community-based Question Answering (CQA) [8, 11]. CQA aims to facilitate a web service that enables users to post their open-domain questions and obtain answers from other users. Among the sub-problems defined for CQA, answer ranking is an important problem and highly related to our study, which roots back in traditional information retrieval systems [30]. Recent works apply neural answer ranking techniques to address vocabulary mismatch and improve automatic feature extraction, e.g., [27, 31]. Other related work considers using external knowledge, such as knowledge graphs [8] and featuring expert users [19, 23] in combination with retrieval methods.

Studies related to cross-domain and cross-lingual QA are also related to our work, as they aim to utilize data from different domains or languages to compensate for resource scarcity. Yu et al. [35] study the transfer of QA knowledge from a low-resource to a high-resource domain by reformulating the problem into paraphrase identification and natural language inference sub-problems used for finding the most similar question from a QA knowledge-base across different domains. For example, they aimed to utilize Quora's data to improve e-commerce question answering. Their proposed approach simultaneously learns shared representations of questions, reviews, and domain relationships for a hybrid model combining sentence encoding and sentence interaction sub-models. For cross-lingual QA, Asai et al. [3] has introduced Cross-lingual Open-Retrieval Answer generation (CORA) as a unified generative QA system for many languages, targeting open-question answering problems. CORA utilizes a novel dense passage retrieval algorithm for retrieving documents across languages for a question. Most of these cross-lingual studies use recent open-domain Q&A datasets, such as XOR-TYDI QA [2] and MKQA [20], that are collected using Wikipedia in different languages. To our knowledge, the literature does not yet address cross-lingual PQA.

Our work focuses on the answer ranking problem in e-Commerce settings. Our main goal is to rank the existing product-related question-answer pairs across parallel marketplaces in different countries to automatically answer questions in low-resource marketplaces. For this purpose, we utilize similar or exact items in high-resource auxiliary parallel marketplaces. The work introduced by Rozen et al. [29] is the most similar to ours. They focus on new or unpopular products for a given marketplace and propose a model for learning an item representation that can be used to find similar items for answering questions. They utilize answers provided for the same question across similar products for item representation. Their problem setting differs from ours in that they only consider products supplied within a single marketplace. In addition, they restricted their study to answer yes/no questions.

3 PROBLEM DEFINITION

We are given a *main* e-commerce marketplace, $\mathbb{M}_T$, suffering from resource scarcity regarding the number of knowledgeable users answering other users' questions. $\mathbb{M}_T$ comprises a set of items $I = \{I_1, \cdots, I_k, \cdots, I_n\}$. For each product I_k, there is a set of question–answer pairs $QA_k = \{QA_{k1}, \cdots, QA_{km}\}$ exist. For simplicity, we assume a single answer for each question. Beyond the main marketplace, we are provided with at least one parallel high-resource marketplace, that is, a marketplace with a large number of similar if not identical products, named the *auxiliary* marketplace(s) and noted as $\mathbb{M}_S$. The set of question–answer pairs for product $\hat{I}_k$ in this marketplace is noted as $\hat{QA}_k = \{\hat{QA}_{k1}, \cdots, \hat{QA}_{kz}\}$, where $\hat{I}_k$ is the same or similar product to I_k. Given that $\mathbb{M}_S$ is a high-resource marketplace, it can be generally assumed that the number of answered questions (especially answered questions) is much bigger than the main marketplace, i.e., $z \gg m$.

We define the problem of cross-market question answering as automatic answering or suggesting better answers for questions posted in the main marketplace $\mathbb{M}_T$ by finding relevant question–answer pairs (QAs) in either $\mathbb{M}_T$ or $\mathbb{M}_S$. We formulate the problem's input as follows: an unanswered Q, which is related to the product I_u from $\mathbb{M}_T = \{I_1, \cdots, I_i\}$, $I_u = \{QA_{u1}, \cdots, QA_{um}\}$ where $Q \notin I_u$, an auxiliary marketplace $\mathbb{M}_S = \{\hat{I}_1, \cdots, \hat{I}_j\}$. The formulated output would be A, a selected answer for Q, i.e., $A \in \mathbb{M}_T \cup \mathbb{M}_S$.

4 DATA COLLECTION & ANALYSIS

4.1 Data collection and annotation

Data collection. We construct our QA dataset on top of an Amazon product collection called XMarket [4]. XMarket expanded a previous Amazon dataset [22] which only focused on the Amazon

Table 1: General statistics of the XMarket-QA.

	us	uk
# categories	16	16
# unique items	29,976	4,124
# qa pairs	4,491,187	330,145
Median q per item	25	11
Mean q length	15.11 ± 8.84	15.14 ± 8.45
Mean answer length	37.99 ± 41.28	38.92 ± 36.85

U.S. marketplace, collected in 2014 and later updated in 2018. We start with XMarket as an initial seed; our decision is motivated by having a complete dataset consisting of all the features that appear on the Amazon real-world website. We match each item that appeared on the XMarket dataset, using the items' unique identifiers (aka. ASINs), with the items on Amazon's website. For those ASINs, we collect all questions for the existing ASINs on the Amazon QA page, and for each question, the top answer, based on users' votes at the time of the crawling, resulting in a dataset of 4.8 million question–answer pairs on two marketplaces. Based on our initial investigation, we see that if two items have the same ASIN, it often means that they are the same products appearing in two different marketplaces. To the best of our knowledge, this is the first cross-market QA dataset in the community.

Human annotation. Given a question, we ask human annotators to assess the relevance of other question–answer pairs. We follow the typical top-K pooling technique [32], where we pool the top five answers from a variety of question retrieval methods (described in Section 6), aiming to build a reusable test collection. Our test set consists of 94 questions[3] (with an average of 33% relevant documents); for each, we annotate an average of 25 pooled answers, leading to a total of 2430 annotated QA pairs. We create Human Intelligence Tasks (HITs) on Amazon Mechanical Turk and provide the workers with two sets of questions and answers: (i) $Q_t, A_t, Title_I$ in which Q_t shows our target question, A_t and $Title_I$ are the top answer and title of the related item for that question, respectively; and (ii) $\hat{Q}_s, \hat{A}_s, Title_I$ in which $\hat{Q}_s$ shows our predicted question, $\hat{A}_s$ and $Title_I$ are the top answer and title of the related item for that question, respectively.

Crowdsourcing relevance labels.

In this crowdsourcing task, we ask the workers to complete two mini-tasks. Our goal is to cover various aspects and edges of the cross-market Q&A to clarify the task for the workers. We briefly describe the two tasks below:

- In the first step, we instruct the workers to determine the similarity of two given items, judged by their titles. We define five similarity levels to make the item-similarity annotation task as objective as possible; hence we provide measurable quantities to describe the degrees of similarity of two items. The similarity levels consider whether the items are similar, have the same brand, have the same version, or have nothing in common.
- In the second step, we ask the workers to assess two question–answer pairs — one from each item they already assessed in the previous step. We ask them if they could answer the question in the main marketplace (**uk**) if they had access to the question–answer pair in either the main marketplace itself or the auxiliary marketplace (**us**). Note that we also provide the correct answer to the **uk** question to allow the workers to better judge without needing domain knowledge. As shown in Figure 1, we instruct the workers to assess the relevance of the questions at three levels, namely, *Very useful*, *Partially useful*, and *Not useful*.

[3]Each question in the test set corresponds to each topic in typical TREC collections.

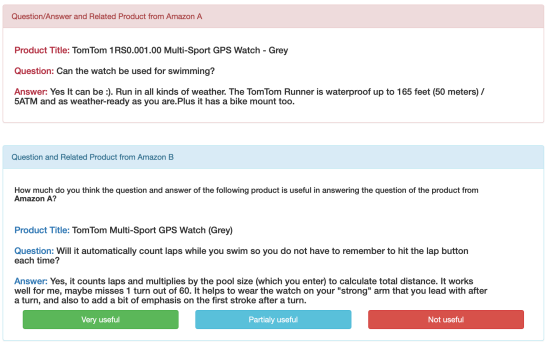
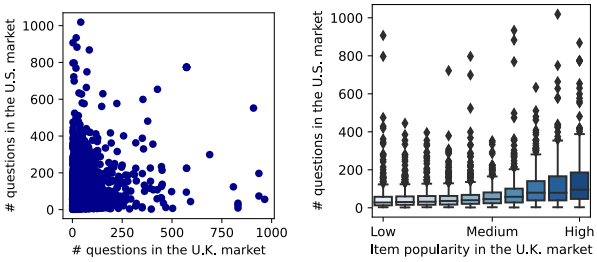


Figure 1: HIT interface for finding the relevant judgment for a question from the main marketplace (the red box) and a question/answer pair from the source marketplace (the blue box). The workers could choose one of the suited options.

Quality assurance. Our annotation task is a complex task with various challenges, where in some cases deep knowledge of the marketplace and the product is required. As an example, an annotator might confuse products with similar titles as being similar. To ensure the high quality of annotations, we instruct the annotators in a step-by-step process and employ various quality assurance techniques. We first launch an onboarding task and only allow the workers who pass this task successfully to take part in the main annotation task. We open the HIT only to workers with at least 10,000 approved HITs and a lifetime approval rate greater than 97%. We limit the workers' geographical location to the U.S. to ensure their English level and familiarity with the Amazon marketplace; however, we instruct them well to distinguish between two marketplaces and avoid any potential biases towards their local Amazon marketplace. Also, we include several test questions with obvious answers in our batches. We observe that qualified workers manage to annotate all the test questions correctly. The Fleiss' κ [12] measure of inter-annotator agreement equals $\kappa = 0.451$, which is considered *moderate agreement*, not bad considering the difficulty of the task. In the first round of annotation, our workers reach an agreed answer on 89% of the annotations, demonstrating the high clarity of the task.

4.2 Data Analysis

Question distribution analysis. Here, we analyze the distribution of questions per item in both marketplaces. We aim to understand the difference in the number of answered questions for the same items in both marketplaces. In Figure 2a, we plot the distribution of question count for the same items in the two marketplaces. On the x-axis, we consider the number of questions related to the items in **uk**. The y-axis represents the number of questions related to the same item in **us**. We observe a low Pearson's correlation ($r = 0.26$, $p \ll 0.001$) regarding the number of questions per item.



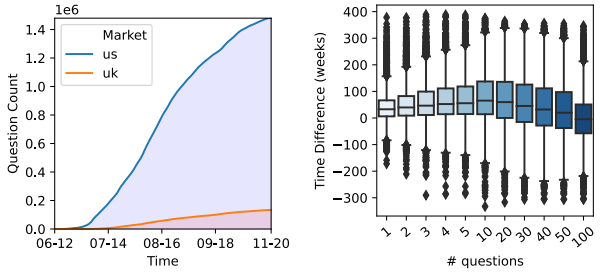
(a) (b)

Figure 2: Comparison of the question distribution per item in uk vs. us. In (a), the x-axis denotes the number of questions per item in uk, and the y-axis, the number of questions of the same items in us. In (b), we group items in uk according to their popularity and plot the distribution of the question counts of the same items in us.

Also, Figure 2a reveals no dependency between the two counts. Therefore, we conclude that popularity in one marketplace does not translate to popularity in another marketplace. Hence, an item with few questions in one marketplace could potentially enjoy numerous answered questions in another marketplace, further motivating the problem of cross-market question answering. To shed more light on this aspect, in Figure 2b, we define ten popularity levels on items belonging to **uk**. To do so, we sort the items based on their question counts, create ten equally sized bins (i.e., Low popularity to High popularity), and plot the question count distribution for the corresponding items that fall into these bins in the **us**. Figure 2b shows that the median **us** question counts increase slightly for the more popular items (hence the low positive correlation). However, if we look closer, we notice many outliers, especially in the items with lower popularity. As we compare the median question count at different bins in the two marketplaces, we can see that the median question count per item in the resource-scarce **uk** items equals 3 (i.e., Low in Figure 2b), whereas the corresponding count in the **us** equals 27. This again supports the potential benefit of transferring QA knowledge from the resource-rich **us** to the **uk**, especially for unpopular/cold items. These items would benefit the most from additional question–answer pairs.

Temporal gap analysis. Next, we consider the temporal aspect of the answered questions to answer the **RQ1**. Our goal is to mimic a realistic scenario where we gain insight into how a young, resource-scarce marketplace could benefit from an older, resource-rich marketplace at each point in time. Figure 3a plots the cumulative sum of the number of questions available on all the items in both marketplaces. Given the nature of the crawled items, we observe that, initially, both marketplaces feature very few questions. However, we see a rapid rise in the number of questions in the **us**, compared to a low pace observed in the **uk**. This reiterates the high potential and higher coverage of resource-rich marketplaces.

Furthermore, we look closer at the item-level temporal distribution to uncover how individual items can benefit from the temporal gap between the two marketplaces. To this aim, we analyze the temporal gap between the same item in the two marketplaces in terms of the question-answering activity of users. Specifically, we compute the time difference between the first questions answered in **us** and **uk** for the same items. We aim to show how long it takes on average



(a) (b)

Figure 3: Temporal analysis of the two datasets in terms of questions and items. (a) the cumulative sum of the number of questions on the Electronics category for us and uk. The plot shows a clear advantage of us in terms of the number of available questions over time. (b) distribution of the time difference between the n-th ($n \in \{1, \ldots, 100\}$) question posted on the same item in us compared with uk. We observe a bigger time difference in the initial posts, suggesting the use of an additional resource-rich marketplace can be beneficial.

for an item to have its first question answered in the **uk**, compared to when its first question appears in the **us**. Figure 3b plots the time difference distribution between the n-th question for the same item in the two marketplaces, with $n \in 1, \ldots, 100$. We see a positive trend in the median temporal difference up to the first ten questions, suggesting a large temporal gap between the two marketplaces. In particular, we observe a temporal gap of, on average, 43.22 weeks (with a median of 33) between the first question being answered in the **uk**, compared to **us**. This huge gap supports the hypothesis that new marketplaces can benefit from the data available in older marketplaces. Note also that the gap increases as the number of questions increases — for the 10th question, mean: 76.30; median: 65 weeks. As the number of questions goes up to 100, we see a smaller time lapse between the two marketplaces, suggesting once more that especially the cold/unpopular items would benefit and that the opportunity to carry over information between marketplaces decreases when items gain enough popularity in both marketplaces.

Finally, we study the number of questions for each item that appears in the **us** before the first question of this same item arrives in the **uk**. For every item, we only keep the questions in the **us** dated before the first question has been posted in the **uk** (for the corresponding item). We find that 70% of the questions in the **us** regarding common items will have been answered there before even the first question receives an answer in the **uk**. This means that on average, each item has 28.11 (median: 11) answered questions in the **us** before the first question is even posted in the **uk**, highlighting the very fact that usually much information is already available about a particular product in other, resource-rich marketplaces.

Cross-market answer distribution. With a final analysis of our data, we aim to test our core hypothesis that an auxiliary resource-rich marketplace (**us**) can help answer questions asked in a resource-scarce marketplace (**uk**). To this end, we analyze our collection of question relevance assessments (see Section 4.1) to see the proportion of relevant question–answer pairs in the **us** for each question in the **uk**. In our judgment pool, 55% of the question–answer pairs originate from the **us**, demonstrating the diversity of the retrieval methods we utilize in the pool. Looking at the question–answer pairs

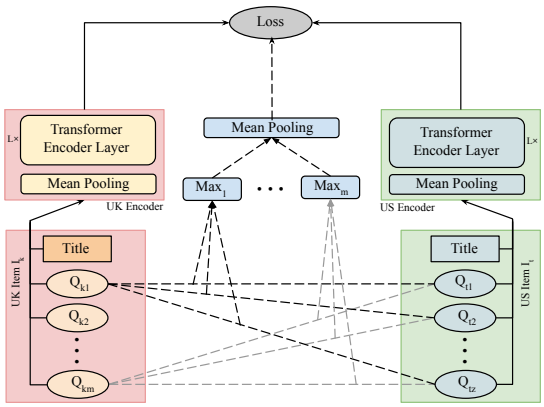


Figure 4: The architecture schema of CMJim. The uk marketplace is the main marketplace (left) and us is the auxiliary marketplace (right).

rated as relevant, we observe that 65% originate from the **us**, demonstrating the high potential of finding a relevant question–answer pair in the auxiliary **us** marketplace. When we count the number of items in the **uk** where additional relevant question–answer pairs can be found in the **us**— more data is available in the auxiliary marketplace, compared to the main marketplace itself — we see that 51% of the **uk** items have more relevant question–answer pairs in the **us**, confirming the significance of cross-market question answering.

5 PROPOSED MODEL

Here we explain our proposed approach for ranking answers to a new product-related question in the main marketplace. We name our method Cross-Market Joint Similarity (CMJim). It ranks products and their corresponding questions across marketplaces. Figure 4 presents the overall schema of our proposed model. In summary, CMJim learns item representations for both main and auxiliary marketplaces based on each item's questions as well as the product title. We first fine-tune a pre-trained BERT [10] model for domain adaptation purposes. Then, we extract dense contextualized embedding vectors for questions and titles as representations for items and related questions. CMJim uses these representations for training and inference. In the following, we provide further details on each step described as well as our training and inference paradigm.

Domain adaptation. Since the existing pre-trained Transformer-based [33] language models are trained using the general text, including Wikipedia, we need to fine-tune the model before using it for the e-commerce domain. For this purpose, we fine-tune a pre-trained BERT model using our data. Since we only use the *[CLS]* output vector in our study, we use a single linear classification layer on top of the *[CLS]* output for fine-tuning. We use the answer relevancy prediction task, so for every question in our train set, we consider the related top answer as positive and two random answers as negative samples. We refer to this model as the DA-BERT model for domain-adapted BERT. We only used **us** data for this domain adaptation due to sufficiently covering the general e-commerce domain. Furthermore, we only used **us** since we assume that the available data size in the source marketplace is much higher than the main marketplace, so in scenarios where the main marketplace is very small, choosing the source marketplace for the domain adaptation task is more realistic.

5.1 Model Architecture and Training

Our model employs a bi-encoder neural model, successfully applied to various product search and recommendation problems [4, 24].

Question Embedding. For a given question, Q, related to an item, we obtain a dense vector embedding using the *[CLS]* output of DA-BERT. For Q, we use EQ as the question's representation vector.

Item Embedding. For a given item, I_k, we obtain the dense vector embedding of the item using its questions, i.e. $\{Q_{k1}, \cdots, Q_{km}\}$ and the product title, i.e. D_k. For I_k, we use EI_k as the item's representation vector and calculate using the mean pooling layer as shown in Eq. 1.

$$EI_k = MeanPooling(EQ_{k1}, \cdots, EQ_{km}, ED_k) \ . \tag{1}$$

CMJim. Our main bi-encoder model, is designed to answer **RQ2** and leverage the core discriminating feature of the cross-market data: *the exact same items*. It takes one item representation from the main marketplace (denoted as EI_t) and the auxiliary marketplace (denoted as $E\hat{I}_k$) and passes through two transformer-based encoders [33]. In our experiments, we use a transformer encoder with six encoder layers (i.e., $L = 6$), six multi-head attention layers, random initial weights, and no parameter sharing. The same products across marketplaces in our data share the same unique item identifier (ASIN). Using this information, every item in the main marketplace which has an identical item in the auxiliary marketplace, is used as a positive sample ($y = 1$). Note that we do not use this information during inference. We also randomly sample two other items as our negative samples ($y = -1$) for training. We use the cosine embedding loss function with the margin of -1 as shown in Eq. 2 with our training for model parameter updates. With this training, we aim to learn similar representations for exact items across marketplaces.

$$Loss = \begin{cases} 1 - \cos(x_1, x_2), & \text{if } y = 1 \\ 1 + \cos(x_1, x_2), & \text{if } y = -1 \end{cases} . \tag{2}$$

5.2 Inference

Finding products with the most similar questions is the key part of our problem. The process of inference is shown in Figure 4 using dashed connecting lines. During inference, for a given question with no answer, we first rank products, and then among top-k products, we rank top-l questions. In order to rank products, we use CMJim's encoders to vectorize all products. Using efficient approximate approaches for vector space ranking (such as Faiss library [15]), we obtain top-k products. By ranking questions associated with these top-k products, we are able to predict the best existing questions that can potentially answer the question in the main marketplace. For ranking questions, we simply use our DA-BERT model to vectorize questions and similarly obtain top-l questions. We use $k = 3$ and $l = 50$, found by experiments, in our setup.

6 EXPERIMENTAL SETUP

Dataset. We use the XMarket-QA dataset for our experiments. We use the **us** marketplace as the auxiliary marketplace and the **uk** marketplace as the main marketplace. We specifically focus on the Electronics category. We prepare each marketplace's data in the same way. For each marketplace, we split the data per item per time,

meaning that for each item, we use the first 70% of the questions as the training set, the next 10% as the validation set, and the final 20% as the test set based on the answer time, from which we randomly sampled 100 questions. In human annotation, we could not collect any relevant question–answer pairs for some of these samples. As a result, the final test set has 94 questions.

We use two versions of the dataset in our experiments: (1) XMarket-QA, in which we use the complete dataset in both training and testing and (2) (XMarket-QA w/o exact items), in which we use the complete training set but we remove all the items that have the matching ASIN with one of our test items during inference.

Baselines. We compare the performance of CMJim to the following models on auxiliary and main marketplaces.

Item-unaware baselines:

- **BM25:** We find the most similar questions to the current question by ranking all the questions available in the training set. We use PyTerrier [21] for the BM25 implementations in this paper.
- **BERT:** We first run their BM25 counterpart and re-rank the top 100 questions using a BERT pre-trained ranker.

Item-aware baselines:

- **Exact-BM25/Exact-BERT:** In these models, we first find the exact same item (matching ASINs) in the auxiliary marketplace and rank only its questions. The question ranking is done using BM25 and BERT, respectively.
- **BM25-BM25:** In this method, we first rank the items based on their similarity to the current item. The similarity is measured based on their title. We rank the items by BM25 score and rank the questions of the top 3 items using BM25.
- **BM25-BERT:** We first run the BM25-BM25 baseline, then take the top 50 questions and re-rank them using BERT.
- **SimBA+APC[29]:** To find similar products, their contextual similarity is calculated based on their answers to the same question. Then, a mixture-of-expert framework predicts the answer by aggregating the answers from contextually similar products. This work was originally designed for yes/no questions and a single marketplace. To make this method similar to our settings, instead of using the yes/no question prediction, we use the list of highly similar questions upon which the answer was predicted. We also mixed both **us** and **uk** data to make the model cross-market.

Hyper parameters. We used learning rate of 2e-5 with no warmup, batch size 16, 1 epoch. Also, the main training parameters used in our experiments to train CMJim is a learning rate of 2e-5 with 3000 warmups, batch size 16, 40 epochs, for training the encoders.

Evaluation metrics. We follow the related work [39] and measure the performance in terms of the following metrics: MRR, P@5. We also use Hit-Rate (HR) for a cut-off of 5, NDCG@3, and NDCG@5.

Significance test. We determined the statistically significant differences using the two-tailed paired t-test with Bonferroni correction at a 95% confidence interval ($p < 0.05$).

7 RESULTS AND DISCUSSION

Performance comparison. In Table 2, we report the performance of CMJim on the XMarket-QA dataset using the annotated test queries compared with other baselines. Table 2 is vertically divided into three parts based on the training data used from each marketplace. The first part of the table (i.e., **uk**) shows the results of the models that only use the main marketplace's data for training. Our results demonstrate the effectiveness of these models when only the data from the main marketplace is available (i.e., no cross-market training is done). Moreover, the table is horizontally divided into two parts based on the test set. The left part of the table reports the results on the full test set, while the right part reports the results on a subset of the test in which all the exact-matching items in the auxiliary **us** marketplace are removed. Our goal is to show how dependent each model is on the existence of the exact same item in the other marketplace — are they generalizable?

Due to space concerns, we do not report the item-unaware baselines in the table. However, it is worth noting that these models perform very poorly as they have no knowledge about either of the products (i.e., neither the current product nor the one with a similar question). Therefore, as long as the questions have a high similarity score, they are considered similar. For example, assume there is a question about an iPhone, "Does it support USB-C?" A similar question can be found on various products such as Android phones, which leads to a wrong answer. Among the item-aware baselines, however, we observe a big improvement in performance in terms of all evaluation metrics compared to the item-unaware baselines. This indicates the necessity of modeling item relevance as part of the ranking process. As another example for comparing the item-aware and item-unaware model, we can mention the following: "This seems to keep my iPad at 1%, shows the charging icon but doesn't charge the iPad? "is a sample question from our ground-truth annotations, related to a charger. The best similar question that an item-unaware BM25 model finds is "Doesn't charge ", which belongs to a charging cable instead of a charger. However, the best similar question that our item-aware BM25-BERT baseline finds is "Can you use the lightning port for audio connections or only to charge? "from the same item in the same marketplace. On the other hand, what CMJim returns is "Does this work with iPad Pro 2020? "from the same item, but in the source marketplace.

Our simplest item-aware baselines are the Exact-BM25 and Exact-BERT. The results show how much improvement we obtain if we refer to the same item in other marketplaces. The matching is done based on the item's ASINs in the dataset. We report the results of the Exact baselines only on the **us** data as we assume that other questions on the same product in the main marketplace do not contain the answer to the user's questions, and hence, finding the exact same ID is only possible in another marketplace. Comparing the results on the full test set (the left part of the table), when we compare the results of exact-matching baselines trained on the **us** data with BM25-BM25 and BM25-BERT baselines trained on the **uk**, we observe that the baselines trained on the **uk** data outperform the exact-matching baselines in a main marketplace. This can be due to multiple reasons. While still stressing the ineffectiveness of the exact-matching solution, it also suggests that some of the questions cannot be answered from other marketplaces. For example, "How fast does it deliver?" is a very market-sensitive question, which can only be answered using the questions of similar items in the same marketplace.

Next, we compare the performance of different models based on the data that we use in training. To do so, we compare the performance of the same models when trained on different marketplaces,

Table 2: Performance comparison of CMJim with baselines on XMarket-QA. Models are separated based on their training data, that is, the uk upper part contains the models that only use the uk data, the us lists the models that only use the us data, and the all part lists the models that use the data from both marketplaces in training. All inferences are done using the uk data. * denotes statistically significant differences compared to BM25-BERT-uk (p-value < 0.05).

market	Method	XMarket-QA					XMarket-QA w/o exact items				
		MRR	HR@5	P@5	NDCG@3	NDCG@5	MRR	HR@5	P@5	NDCG@3	NDCG@5
uk	**BM25-BM25**	0.546*	0.735	0.350	0.352*	0.369*	0.167	0.316	0.095	0.112	0.130
	BM25-BERT	0.605	0.775	0.365	0.393	0.407	0.161	0.337	0.103	0.085	0.131
us	**Exact-BM25**	0.507*	0.632*	0.284*	0.289*	0.305*	–	–	–	–	–
	Exact-BERT	0.458*	0.653*	0.321*	0.271*	0.282*	–	–	–	–	–
	BM25-BM25	0.461*	0.643*	0.262*	0.268*	0.275*	0.167	0.316	0.105	0.111	0.134
	BM25-BERT	0.430*	0.592*	0.239*	0.245*	0.249*	0.203	0.326	0.086	0.143*	0.155
all (us +uk)	**SimBA+APC**	0.519*	0.701	0.342	0.341	0.352	0.178	0.221*	0.113	0.107	0.114
	BM25-BM25	0.568	0.755	0.408	0.393	0.433	0.163	0.275*	0.106	0.106	0.134
	BM25-BERT	0.511*	0.694*	0.355	0.342	0.371	0.184	0.255*	0.094	0.129*	0.140
	CMJim	**0.656***	**0.816***	**0.489***	**0.475***	**0.487***	**0.295***	**0.421***	**0.140**	**0.207***	**0.219***

e.g., BM25-BM25-**uk** vs. BM25-BM25-**us**. We observe that all the models benefit from the data provided by the auxiliary **us** marketplace. This is evident as we compare the results of BM25-BM25 and BM25-BERT trained on **uk** and **us**. We see a 4% and 2% improvement in MRR on BM25-BM25 and BM25-BERT, respectively, suggesting that while similar questions in the main marketplace are useful in answering product-related questions, exploring other resource-rich auxiliary marketplaces can lead to improvements, especially when data of the main marketplace is scarce. When mixing the data from both **us** and **uk**, we see a mixed reaction from the models. Those results can be seen in the **all** (**us** +**uk**) section of Table 2. While we see some slight improvements in some of the evaluation metrics (e.g., Hit@5), overall, we do not observe a big difference in the performance of the item-aware baselines. This indicates that while using both marketplaces is potentially beneficial, the baseline models cannot exploit it.

We see that CMJim outperforms all the baselines significantly as it benefits from joint training on question and product similarity, which enables it to learn the dependency of question similarity on the product. Therefore, it can rank the relevant questions high even if they do not belong to similar products (e.g., asking about the power outlet compatibility), enabling the model to outperform the strong baselines in Table 2. In particular, we observe a 14% improvement compared to SimBA, which is the state-of-the-art model in PQA, indicating the effectiveness of our joint ranking model.

Performance comparison when leaving the exact items out. In this experiment, we aim to examine how dependent each model is on the existence of the exact same items in the auxiliary marketplace. Our goal is to test an extreme case of having no matching item in the auxiliary marketplace and test the ability of CMJim in learning to find relevant questions that belong to other items. We deem these cases to be more challenging. Therefore, in our test set, we remove all the items that have matching ASIN with one of our test items. The left part of Table 2 (XMarket-QA w/o exact items) reports the models' performance under this condition. We see that the performance of all the models drops to a great extent when the exact items are removed because it is very likely to find a relevant question–answer pair in the same item. While using XMarket-QA, results on BM25-BERT-**uk** suggest that this method is a stronger baseline than BM25-BM25-**uk**; however, when we remove the exact items, we can see that the margin in this marketplace lowers. A different thing happens in the **us** marketplace when XMarket-QA is used; we see worse results from BM25-BERT-**us** than BM25-BM25-**uk**. Also, when using XMarket-QA, baselines on **uk** marketplace exhibit stronger performance than their counterparts in the **us** marketplace. However, when we remove the exact items, this pattern changes. Although CMJim's performance drops significantly, just all other models, it still outperforms all the baselines, indicating its ability to the joint question–item ranking leads to a better product representation which enables CMJim to find similar non-exact items more effectively.

Effect of main marketplace (uk) data size. In this experiment, our goal is to understand how dependent a cross-marketplace question–answering system is on the size of the main marketplace. We hypothesize that the more scarce a marketplace is, the more effective using an auxiliary marketplace would be. To perform this experiment, we leave the auxiliary marketplace data (i.e., **us**) unchanged while randomly sampling from the main marketplace (i.e., **uk**). To remove any unwanted model-related effects, we choose the best baseline from Table 2 and compare our model only with BM25-BM25 on three marketplace combinations (i.e., **uk**, **us**, and **all**). Figure 5a shows that BM25-BM25-**all** and CMJim consistently outperform the BM25-BM25-**uk** model. Even as more data in **uk** is available, the models that use the data from both marketplaces effectively outperform the model that only uses the **uk** data. This suggests that the data from a resource-rich marketplace improves the performance even if more data is available in the main marketplace. Note that the relative size of the two marketplaces in XMarket-QA is very different, suggesting that the high volume of **us** data overshadows the fact that more data from **uk** marketplace becomes available to the models. Nevertheless, we observe a constant improvement in the performance as more **uk** data become available, indicating that certain questions are market-dependent (e.g., questions about the delivery time and suppliers). Also, we

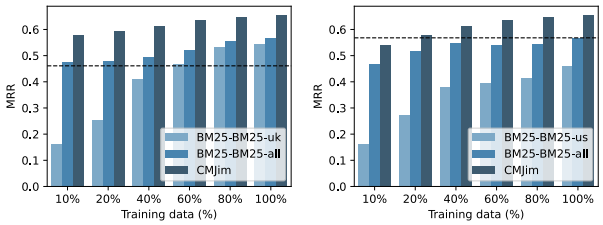
report the performance of the baseline model trained only on the **us** data, BM25-BM25-**us**, using the dashed line.[4] Furthermore, as we increase the size of **uk**, the model trained only on **uk** performs better. After having 80% of the data, it almost reaches the same performance as the model on **all** marketplaces. This indicates that while cross-market training always helps the model to perform better, it exhibits much better performance for cases where the main marketplace data is very scarce (less than 60% in this case).

Effect of training on the incomplete product set in the auxiliary marketplace (us). With this experiment, we want to answer the **RQ3**. We aim to evaluate the performance of the models when trained in a more realistic scenario where the products in the main marketplace (**uk**) do not necessarily have an exact match in the auxiliary marketplace (**us**). To this aim, we randomly sample the **us** products while keeping the main marketplace data (i.e., **uk**) unchanged. Doing so, we simultaneously test the performance of our model when fewer auxiliary data are available and when **uk** products do not have **us** exact matches. Figure 5b compares the performance of BM25-BERT in all three possible marketplaces; we specifically use this method to include all scenarios in comparison. We examine the performance on {20%, 40%, 60%, 80%, 100%} of **us**, which respectively corresponds to {15%, 9%, 6.5%, 6%, 5%} of **uk** items with no **us** exact matches. We see that BM25-BERT-**us** in most cases outperforms the dashed line (i.e., BM25-BERT-**uk**), and the only exception is when only 20% of **us** is fed to the model, leading to 15% of the **uk** products not having an exact match in **us**. We can relate that to the fact that the auxiliary marketplace is resource-rich, where the model can find similar items to cover the exact missing items when enough data is available. Here we aim to examine the cases where an auxiliary marketplace can effectively improve performance. Results show how large an auxiliary marketplace should be to effectively boost the main marketplace's performance. We see in Figure 5b that as the size of the auxiliary marketplace increases, there is a boost in the performance of the cross-market models. Surprisingly, we find consistent improvements for different data sizes of **us**. We attribute this to the fact that the size of the **us** is very large, and hence even 20% of it can improve performance. Furthermore, as more data becomes available (60%), CMJim seems to be saturated, and the performance gain is not as much as before. Also, comparing the performance with the dashed line (i.e., BM25-BERT-**uk**), we see that BM25-BERT-**us** and CMJim outperform BM25-BERT-**uk** when more than 40% of the **us** data is available, indicating the usefulness of the **us** data even when its size is smaller. However, BM25-BERT-**all** performs worse than BM25-BERT-**uk** for smaller **us** data sizes. This can be due to the fact that the model does not effectively separate the two marketplaces and performs worse when the data from both marketplaces are available.

8 CONCLUSIONS & FUTURE WORK

This paper investigates the potential impact of using available data in a resource-rich marketplace to answer questions in a resource-scarce marketplace, a new problem called cross-market product-related question answering. We collect and annotate the first cross-market question-answering dataset called XMarket-QA, providing

[4]Note that, in this experiment, since the data of the **us** is not altered, the performance does not change for different **uk** data sizes. Therefore, we show its performance as a dashed line.



(a) uk **(b) us**

Figure 5: Performance comparison of changing the marketplace size across (a) uk marketplace, (b) us marketplace. The dashed lines denote the performance of the same model using the data from the other marketplace, that is, (a) BM25-BERT-us, (b) BM25-BERT-uk.

4.8 million questions and their top answers across Amazon's **uk** and **us** marketplaces. Our data analysis shows that there is a significant temporal gap between the first question answered in resource-rich and resource-scarce marketplaces and that a significant percentage of questions in resource-scarce marketplaces can be answered using data from resource-rich marketplaces.

We propose a model named CMJim that learns item representations and how items and their questions relate in two marketplaces and jointly ranks items and questions using one resource-scarce marketplace and one resource-rich marketplace. Experiments demonstrate the effectiveness of utilizing item ranking in models by CMJim. We observe up to 5% improvement in terms of MRR compared to the competitive baselines selected for our analysis. We thoroughly analyze the collected data, showing that using the data available in a resource-rich marketplace improves performance in a resource-scarce marketplace. This is even more highlighted when considering the temporal gap between the two marketplaces. This gap may occur due to a higher activity of users in one marketplace or by earlier availability of products (i.e., early adoption).

Numerous potential directions can be explored to exploit further and understand cross-market product-related question answering. One limitation of the introduced task is that marketplaces must be of the same language. The data can be extended for cross-lingual and multi-market problems in the future. Another limitation of the approach is using the exact same products across different marketplaces in training. An interesting next step would be to modify the model to use similar products without considering the ASINs. In another direction, the models can be modified for answer generation instead of answer selection and clarifying question generation. In addition, our designed model can be extended to leverage other auxiliary resources such as users' reviews and product descriptions.

Acknowledgments. This work was supported in parts by Radboud University, the EU project OpenWebSearch.eu under GA 101070014, the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), and the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Muhammad Asaduzzaman, Ahmed Shah, Mashiyat Chanchal, K. Roy, and Kevin A. Schneider. 2013. Answering questions about unanswered questions of stack overflow. In *Proceedings of the 10th International Workshop on Mining Software Repositories*. IEEE Press, 97–100.

[2] Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual Open-Retrieval Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 547–564.

[3] Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems* 34 (2021).

[4] Hamed Bonab, Mohammad Aliannejadi, Ali Vardasbi, Evangelos Kanoulas, and James Allan. 2021. Cross-Market Product Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 110–119.

[5] B Barla Cambazoglu, Mark Sanderson, Falk Scholer, and Bruce Croft. 2021. A review of public datasets in question answering research. In *ACM SIGIR Forum*, Vol. 54. ACM New York, NY, USA, 1–23.

[6] Long Chen, Ziyu Guan, Wei Zhao, Wanqing Zhao, Xiaopeng Wang, Zhou Zhao, and Huan Sun. 2019. Answer identification from product reviews for user questions by multi-task attentive networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 45–52.

[7] Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019. Review-Driven Answer Generation for Product-Related Questions in E-Commerce. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (2019).

[8] Yang Deng, Ying Shen, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge as A Bridge: Improving Cross-domain Answer Selection with External Knowledge. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3295–3305.

[9] Yang Deng, Wenxuan Zhanng, and Wai Lam. 2020. Opinion-aware Answer Generation for Review-driven Question Answering in E-Commerce. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[11] Hanyin Fang, Fei Wu, Zhou Zhao, Xinyu Duan, Yueting Zhuang, and Martin Ester. 2016. Community-based question answering via heterogeneous social network learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[12] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (1971), 378–382.

[13] Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2021. Meaningful Answer Generation of E-Commerce Question-Answering. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–26.

[14] Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 429–437.

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734* (2017).

[16] Rae Yule Kim. 2020. The Impact of COVID-19 on Consumers: Preparing for Digital Sales. *IEEE Engineering Management Review* 48, 3 (2020), 212–218.

[17] Bernhard Kratzwald, Anna Eigenmann, and Stefan Feuerriegel. 2019. RankQA: Neural Question Answering with Answer Re-Ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6076–6085.

[18] T. Lai, Trung Bui, Nedim Lipka, and Sheng Li. 2018. Supervised Transfer Learning for Product Information Question Answering. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2018), 1109–1114.

[19] Xiaoyong Liu, W Bruce Croft, and Matthew Koll. 2005. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM international conference on Information and knowledge management*. 315–316.

[20] Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *Transactions of the Association for Computational Linguistics* 9 (2021), 1389–1406.

[21] Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation in information retrieval using pyterrier. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 161–168.

[22] Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*. 625–635.

[23] Sara Mumtaz, Carlos Rodriguez, and Boualem Benatallah. 2019. Expert2vec: Experts representation in community question answering for question routing. In *International Conference on Advanced Information Systems Engineering*. Springer, 213–229.

[24] Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. 2019. Semantic product search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2876–2885.

[25] Dae Hoon Park, Hyun Duk Kim, ChengXiang Zhai, and Lifan Guo. 2015. Retrieval of relevant opinion sentences for new products. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 393–402.

[26] Fatemeh Pourgholamali. 2016. Mining information for the cold-item problem. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 451–454.

[27] Jinfeng Rao, Hua He, and Jimmy Lin. 2017. Experiments with convolutional neural network models for answer selection. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1217–1220.

[28] Kevin Roitero, Ben Carterette, Rishabh Mehrotra, and Mounia Lalmas. 2020. Leveraging Behavioral Heterogeneity Across Markets for Cross-Market Training of Recommender Systems. In *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*. ACM / IW3C2, 694–702.

[29] Ohad Rozen, David Carmel, Avihai Mejer, Vitaly Mirkis, and Yftah Ziser. 2021. Answering Product-Questions by Utilizing Questions from Other Contextually Similar Products. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 242–253.

[30] Aliaksei Severyn and Alessandro Moschitti. 2013. Automatic Feature Engineering for Answer Selection and Extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA.

[31] Taihua Shao, Fei Cai, Honghui Chen, and Maarten De Rijke. 2019. Length-adaptive neural network for answer selection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 869–872.

[32] Nicola Stokes. 2006. Book Review: TREC: Experiment and Evaluation in Information Retrieval, edited by Ellen M. Voorhees and Donna K. Harman. *Computational Linguistics* 32, 4 (2006). https://aclanthology.org/J06-4008

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[34] Mengting Wan and Julian McAuley. 2016. Modeling Ambiguity, Subjectivity, and Diverging Viewpoints in Opinion Question Answering Systems. *2016 IEEE 16th International Conference on Data Mining (ICDM)* (2016), 489–498.

[35] Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 682–690.

[36] Qian Yu and Wai Lam. 2018. Review-Aware answer prediction for product-related questions incorporating aspects. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 691–699.

[37] Shiwei Zhang, Jey Han Lau, Xiuzhen Zhang, Jeffrey Chan, and Cecile Paris. 2019. Discovering relevant reviews for answering product-related queries. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1468–1473.

[38] Shiwei Zhang, Xiuzhen Zhang, Jey Han Lau, Jeffrey Chan, and Cecile Paris. 2021. Less Is More: Rejecting Unreliable Reviews for Product Question Answering. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*. Springer International Publishing, 567–583.

[39] Wenxuan Zhang, Yang Deng, and Wai Lam. 2020. Answer Ranking for Product-Related Questions via Multiple Semantic Relations Modeling. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).