A Bespoke Question Intent Taxonomy for E-commerce

Diji Yang[1,*], Omar Alonso[2]

[1]University of California Santa Cruz
[2]Amazon

Abstract

Effective question-intent understanding plays an important role in enhancing the performance of Question-Answering (QA) and Search systems. Previous research in open-domain QA has highlighted the value of intent taxonomies in comprehending data and facilitating answer generation and evaluation. However, existing taxonomies have limitations for specific domains. We're interested in question intent for e-commerce scenarios where questions are specific to shopping activities.

To address such limitations, we propose the adoption of a bespoke strategy for the e-commerce domain. We introduce an E-commerce Question Answering (EQA) taxonomy designed to encapsulate the unique aspects of e-commerce queries. Our empirical analyses validate the EQA taxonomy's ability to more accurately represent users' information needs in shopping scenarios. Further, we employed instruction fine-tuning to develop an intent classifier capable of categorizing questions following EQA taxonomy. Our result shows that EQA can provide clear guidance for intent classification for e-commerce queries. Finally, our approach shows that it is possible to build a domain-specific taxonomy and associated classifiers that can be used in different applications.

Keywords

intent understanding, question taxonomy, question answering

1. Introduction

Question answering (QA) as a longstanding task in NLP has been pushed forward rapidly in recent years with the development of language models. Transformer-based models perform well in most factoid QA datasets, however, they tend to still perform poorly compared to humans in datasets containing more complex problems. In closed-domain QA, such as AmazonQA [1], this problem is more noticeable and requires relevant knowledge. As a result, QA applications are limited in scenarios in which they are deployed for product-level services. At the same time, early research suggests that accurate intent understanding forms the cornerstone for successful information retrieval and contextually relevant answer generation [2, 3]. The goal of question intent understanding is to categorize user queries into distinct intent classes. This categorization aids in facilitating data comprehension, answer generation, and evaluation [4, 2, 5]. It can also be used as a signal for relevance ranking and improving diversity in search results.

In practice, Broder [6] shows the importance of classifying user queries in web search and how it reflects the real word. Intent taxonomies aid in categorizing questions based on their inherent purpose and help in improved answer synthesis and evaluation [7]. However, it has been observed that a single intent taxonomy may not be universally applicable across diverse

eCom'24: ACM SIGIR Workshop on eCommerce, July 18, 2024, Washington, DC, USA
*Work done during internship at Amazon.
dyang39@ucsc.edu (D. Yang); omralon@amazon.com (O. Alonso)

domains due to the specific nuances inherent in different contexts [8, 9, 10]. Bolotova et al. [11] proposed a unified intent taxonomy for non-factoid questions (NFQA). While the NFQA taxonomy is effective in certain contexts, it falls short when applied to the fine-grained features of e-commerce. Human-to-human three-way agreement when using NFQA stands at 49.13%, indicating a lack of consensus in categorizing intent for e-commerce-related queries. Furthermore, a noticeable category imbalance exacerbates the challenges in effectively classifying e-commerce-related questions using the NFQA.

We propose EQA (E-commerce Question Answering) taxonomy, a tailored approach that advocates for the creation and adoption of a bespoke taxonomy dedicated to e-commerce questions. Specifically, recognizing the limitations of the existing NFQA taxonomy in accurately reflecting the intent of e-commerce queries, we eliminate categories that show low inter-rater agreement rates in the e-commerce context and introduce new categories that are more contextually appropriate. EQA is designed to encapsulate the unique characteristics of e-commerce data and user queries within this domain. Our taxonomy demonstrates that it can represent users' real information needs in the context of shopping scenarios. To operationalize the EQA taxonomy, we leverage instruction fine-tuning [12] to train an intent classifier for e-commerce questions. Our experiments demonstrate the effectiveness of this approach in accurately categorizing e-commerce queries.

Our contribution can be summarized as follows:

- We propose a question intent taxonomy for e-commerce questions that can be used in different shopping scenarios. Our quantitative and qualitative analyses confirm the reliability of this taxonomy in e-commerce problems.
- We describe how to build classifiers when introducing a new taxonomy. While EQA is based on e-commerce, we believe this methodology can be generalized to other domains.

2. Question intent

NFQA is a comprehensive question intent taxonomy for open-domain question-answering tasks. To explore the suitability of NFQA trained from open-domain questions for e-commerce questions, two human annotators followed the NFQA taxonomy to label a set of in-domain questions. Meanwhile, using the pre-trained classifiers of NFQA, we obtained NFQA predictions from a deep learning model. The results show that human-to-human agreement stands at a mere 49.13%. These disagreements occur mainly in **experience** and **evidence-based** categories. Moreover, we note that some classes do not faithfully respond to the information needs of the question, e.g., **debate**. More quantitative analyses are covered in Section 4.

3. E-commerce taxonomy

The EQA taxonomy is presented in Table 1. In this section, we describe how EQA fulfills users' information needs in close domains and how to evaluate the taxonomy.

Category	Description	Example
Instruction	The customer wants instructions, guidelines, or procedures to achieve something with respect to a product or service.	`Where is the doorknob? Once the code is entered, how do you open the door?` `How can I tell if this will work on my TV or BluRay player?`
Opinion	The customer wants a subjective piece of information about a product, service, or shopping category.	`Is this product worth buying or will I end up sending it back?` `Any defect complaints?`
Description	The customer wants a definition, description, explanation, or summary of a product, service, or shopping category.	`What are the sizes and types of blades that come with the 5-blade package?` `What are the measurements of this product?`
Comparison	The customer wants a comparison of two or more products or services.	`What's the difference between exclusive Castiel and regular?` `How does this knife compare with a Kershaw?`
Recommendation	The customer wants recommendations for a product or service.	`Looking for a bag for golf for drinks and snacks. Would this be a good choice?` `I need a case for a .357 with a six-inch barrel. Suggestions?`
Factoid	The customer wants an objective piece of information about a product, service, or shopping category.	`Is the price and shipping for one bar or a set of two? Does it fit on Honda CRV 2014?`

Table 1
The proposed EQA taxonomy with examples from the AmazonQA [1] dataset.

3.1. Information Needs and Question Intent

The concept of *information need* refers to the foundational motivation driving users to engage with search systems [3]. Questions are shaped by the askers' specific contexts, akin to how semantics in linguistics rely on contextual understanding [13, 14, 15]. Thus, analyzing questioning requires consideration of the broader context (i.e., question intent). In contrast to open-domain QA or web search queries, users within the e-commerce domain pose questions for more targeted purposes – specifically, to facilitate subsequent purchasing decisions. This inherent focus renders the coarse-grained taxonomy of the general domain insufficient in capturing the nuanced distinctions between e-commerce intents.

Building upon the existing NFQA taxonomy, we first focus on the fact that the **debate** category does not constitute a valid intent in e-commerce. When individuals take part in shopping, they are not likely to anticipate engaging in formal debates with others. Even inquiries that could potentially spark heated discussions on forums or in other contexts, such as "`What's the best graphics card for non-gamers?`", where the user is looking for advice with a purchase. This observation consequently highlights the significance of **recommendation**, which is a prevalent intent within the e-commerce domain. Furthermore, **opinions**, representing subjective insights from other consumers, prove instrumental in guiding purchasing decisions. This intent commonly manifests in the form of queries seeking feedback on product usage or opinions about comparable items. Correspondingly, inquiries for objective information are encompassed by the **description** category. Answers to these queries can often be readily found on the product page, such as technical details provided by the seller. Alongside these,

instruction and **comparison** persist as two enduring question types that are pertinent in e-commerce.

3.2. Evaluation matrix

To quantitatively assess the efficacy of the question intent taxonomy, we adopt two distinct matrices that serve as indicators of its performance across specific datasets.

Distribution of categories We analyze the distribution of each intention as a percentage within the dataset. Although closely tied to dataset characteristics like data source, the distribution offers valuable insights. Extremely unbalanced distributions often imply that current taxonomies struggle to establish effective boundaries for splitting the questions in the given dataset.

Human-to-human agreement Within a given dataset, a well-defined taxonomy should facilitate consistent labeling by diverse human annotators. Agreement implies that the categories in the taxonomy clearly delineate the problem intent. Minimize blurry classification regions as well as controversial labels.

3.3. Intent Classifier Design

3.3.1. Model choice

We chose the encoder-decoder model T5x [16] as the starting point, considering performance and scalability. Benefiting from extensive pre-training data, T5x shows language understanding ability, which is the prerequisite for a language model to predict the intent. Furthermore, encoder-decoder architecture was born with an advantage over encoder-only models in classification tasks [17]. In scalability concerns, decoder-only models that perform well in various NLP tasks tend to be heavy in size and thus difficult to deploy on a lightweight device [18, 19, 20]. For all experiments, we conduct fine-tuning on the Flan-T5-Large model [12].

3.3.2. Data Preparation

Motivated by the success of few-shot learning in NLP [21, 22], we pre-process the training data to better serve the subsequent supervised fine-tuning. Emulating LIMA [23], we prioritize training data quality through stratified sampling for balanced intent representation and manual filtering to eliminate monotonous language patterns. For example, we diversified the **comparison** intent questions to prevent overfitting to repetitive structures like "`What is the difference between A and B?`". Our processed training dataset focuses on both representative across all intent classes and also linguistic diversity, which further enhances the robustness of the fine-tuning process.

EQA		NFQA	
Category	Distribution	Category	Distribution
Factoid	51.98	Debate	55.70
Opinion	16.97	Factoid	16.03
Description	15.07	Instruction	7.61
Instruction	11.50	Not-a-question	7.56
Recommendation	2.23	Experience	5.08
Comparison	2.22	Evidence-based	4.79
Not-a-question	0.04	Comparison	1.61
		Reason	1.63
Variance	275.76	Variance	284.78

Table 2
Intent Distribution on the AmazonQA dataset from EQA and NFQA classifier.

3.3.3. Model Alignment

To better align with the downstream task, i.e., intent classification, we adopt an instruction fine-tuning paradigm [12]. Specifically, we define the task as a 7-classes classification problem and conduct supervised fine-tuning to tailor the model to our specific requirements.

For prediction trustworthiness, in addition to the intent label, we record the model transition probability at the generated token to approximate the confidence score. Particularly, the score x_i, is determined by its conditional probability given all preceding tokens (the given question), $x_{<i}$. The overall score is computed by Equation 1, where s_i is the logit corresponding to x_i, and the denominator is the sum of exponential logits for all tokens in the vocabulary, ensuring normalization.

$$P(x_i|x_{<i}) = \frac{\exp(s_i)}{\sum_j \exp(s_j)} \tag{1}$$

This formulation quantifies the model's certainty or confusion in selecting x_i given the preceding context, with lower scores indicating the model's uncertainty about the prediction. In practical applications, setting a threshold could advise users against placing too much trust in uncertain predictions, thereby enhancing the reliability of classification results. In this work, the threshold is manually set at 0.6. We envision that future research could develop an adaptive, learnable threshold by training a simple neural network, such as a Multilayer Perceptron (MLP), to improve the discernment of prediction reliability.

4. Experiments and Results

4.1. Training details

In line with the prompt design from recent instruction fine-tuning works, our training utilizes an instruction prefix combined with dataset questions as input and human-annotated intent categories as expected labels. The fine-tuning is delivered by the cross-entropy loss and Adam optimizer through 20 training epochs. The entire fine-tuning process was completed in under

EQA		NFQA	
Category	Distribution	Category	Distribution
Factoid	16.29	Factoid	17.43
Opinion	13.71	Debate	0.57
Description	28.86	Evidence-based	40.29
Instruction	5.71	Instruction	4.29
Recommendation	15.14	Experience	12.86
Comparison	20.29	Comparison	18.57
		Reason	6.00
Three-way Agreement	76.88	Three-way Agreement	49.13

Table 3
Intent Distribution on 350 random sampled data from human annotator following EQA and NFQA taxonomy.

two hours on a single NVIDIA A-100 GPU, following the hyperparameter settings recommended by T5x [16].

4.2. Dataset

Built on the top of a product review-based e-commerce dataset, Amazon review data [24], AmazonQA [1] is known as a community QA dataset. All questions, passages, and answers in AmazonQA are extracted from real human interactions, which makes it an ideal dataset for understanding the real information needs of users in the e-commerce domain. The official test split included 92,726 QA pairs. However, due to the fact the human-to-human agreement requires a significant amount of human effort, we use a subset of the original test data, i.e., 350 questions, as our test set.

4.3. Results and Analysis

4.3.1. Intent Classifier

Table 2 presents the intent distribution of questions in the AmazonQA dataset using EQA and NFQA. Our analysis aimed to establish which taxonomy better represents the nature of queries in an e-commerce context. The EQA taxonomy revealed a predominant focus on factoid questions, constituting 51.98% of the dataset. This result aligns well with the nature of e-commerce inquiries, where customers often seek specific, factual information about products. The next significant category in the EQA taxonomy was **opinion** (16.97%) and **description** (15.07%), reflecting the customer's interest in reviews and detailed product descriptions. In contrast, the NFQA taxonomy's most prominent category was debate, accounting for 55.70%. However, the concept of debate is less relevant in an e-commerce setting, as customers typically seek concrete information rather than engage in discussions of a contentious nature. The **factoid** category in NFQA, while still significant, was markedly lower at 16.03%, suggesting a less precise alignment with the nature of e-commerce queries. Of the remaining categories with relatively low occupancy rates, **Not-a-question** accounts for 7.56% of the NFQA and is higher than any other category. This finding points to many cases where the NFQA classifier fails, which did not

Question	EQA	NFQA
I need to replace a defective Julie from brake on a Cannondale Scalpel MTB. Is this a good replacement and will it bolt right on?	Recommendation (99%)	Debate
Compare best 3G/4G internet access plan?	Comparison (99%)	Debate
10′ is L, H or W?	Factoid (99%)	Not-Question
What is the protein in this? Wish they would call that out on the site. Trying to decide between this and Kay's naturals which is protein packed!	Description (99%)	Not-Question
For a more permanent solution, do you super glue the pads along with the sticky adhesive onto the glass or without the sticky adhesive?	Opinion (88%)	Instruction

Table 4
Qualitative analysis of EQA (with confidence score) and NFQA label on the AmazonQA dataset, highlighting three commonly-seen prediction patterns.

occur with EQA. In both taxonomies, **instruction** and **comparison** share similar proportions. NFQA presents experience and evidence-based with high rates. While these categories are relevant in broader information contexts, their specific applicability in e-commerce is less direct compared to the **recommendation** in EQA, which reflects a customer's desire for guidance to make informed purchasing decisions.

Further statistical analysis revealed that the variance for the intent distribution in EQA taxonomy was approximately 275.76, while for NFQA, it was slightly higher at 284.78. This higher variance in the NFQA taxonomy indicates a broader spread in the distribution of question types, which may imply less consistency in categorization relevance for e-commerce data. Our experiment result from AmazonQA emphasizes the adaptability of the EQA taxonomy in capturing the intent of e-commerce customers, providing a more relevant and practical categorization framework for analyzing customer queries on e-commerce scenarios.

4.3.2. Human Evaluation

We performed human annotation of a random sample of 350 data points, where half were from AmazonQA, and another half were from our internal unpublished real e-commerce data. To ensure reliability and reduce the subjectivity inherent in manual labeling, we recruited three independent annotators and adopted a two-stage majority voting process for deriving the final label. In the initial stage, data points where at least two annotators agreed on the label were directly accepted, and these consensus labels were deemed final for those specific data instances. Next, to address the cases with complete annotator disagreement, we calculated the individual accept rate of each annotator, defined as the proportion of their labels being accepted in the first stage. For data points with divergent annotations, the label proposed by the annotator with the highest accept rate was chosen as the final label. We analyzed the distribution of each intent category and calculated the rate of three-way agreement, which is the proportion of three labelers providing the same label.

As reported in Table 3, the human evaluation results reveal a significant difference in the faithfulness of the EQA and NFQA taxonomies in e-commerce contexts. While the EQA taxonomy achieved a substantial three-way agreement rate of 76.88%, NFQA's agreement rate was

notably lower at 49.13%. This difference highlights a key challenge with NFQA in e-commerce: its categories are less tailored to the specific types of queries that arise in this domain. For instance, NFQA's broader categories, like Evidence-based and Reason, may lead to varied interpretations among annotators when applied to the more focused needs of e-commerce customers. The disagreement in NFQA suggests that its categories, possibly well-suited for open-domain questions, are less intuitive and coherent for e-commerce queries, leading to more subjective and inconsistent categorization. In contrast, EQA, with its higher agreement rate, demonstrates a clear alignment with the distinct, often more pragmatic and product-focused nature of e-commerce questions.

4.3.3. Qualitative analysis

This section outlines three patterns in which the NFQA and EQA yield divergent outcomes. As mentioned in Section 4.3.1, one notable issue with NFQA is its tendency to classify a large number of questions as Debate, which is a less reasonable intent in the online shopping context. For example, as illustrated in the first two examples in Table 4, questions that may contain a debating intent in daily discussion (e.g., debates over "the best") [11], but in an e-commerce setting, more accurately interpreted as seeking product recommendations or comparisons. Another commonly seen pattern is the misclassification of questions as **Not-Question** due to the gap between NFQA's pre-training data and the actual shopping queries, particularly failing to recognize questions containing abbreviations. Our analysis of question length shows that data labeled as **Not-Question** by NFQA averaged 27.85 tokens, contrasted with an average of 13.46 tokens for all other intents. This discrepancy further supports the fact that NFQA falls short of processing longer queries in the e-commerce domain. The last example question seeks an opinion, indicating a preference for judgment over direct steps. The distinction between EQA and NFQA highlights their differential capacities to interpret the demands of e-commerce data. Throughout the above-mentioned three patterns, EQA demonstrates alignment with human intuition and consistently delivers high confidence scores.

5. Conclusion and Future Work

We introduce the EQA taxonomy, tailored specifically for e-commerce queries. Our research highlighted the limitations of generic taxonomies like NFQA in the e-commerce context and demonstrated the need for a domain-specific solution. The development and validation of EQA, coupled with an intent classifier trained using instruction fine-tuning, shows a lot of promise for question intent understanding in e-commerce. This approach offers a more accurate framework for question categorization in e-commerce and sets a precedent for developing domain-specific taxonomies in other specialized areas.

EQA has been practiced reliably on e-commerce data; however, the effectiveness of its intent label for downstream tasks is still unproven. Moving forward, we anticipate the integration of EQA classifiers into operational pipelines, enabling systematic evaluation of their efficacy in supporting downstream tasks. Furthermore, the prospect of extending the methodologies employed in this study to other domains, such as healthcare, by adapting domain-specific intent taxonomies for classifier training points in an exciting direction for future research.

References

[1] M. Gupta, N. Kulkarni, R. Chanda, A. Rayasam, Z. C. Lipton, Amazonqa: A review-based question answering task, arXiv preprint arXiv:1908.04364 (2019).

[2] W. G. Lehnert, A conceptual theory of question answering, in: Proc. of IJCAI, 1977, p. 158–164.

[3] B. Shneiderman, D. Byrd, W. B. Croft, Clarifying Search: A User-Interface Framework for Text Searches, Technical Report, 1997.

[4] A. C. Graesser, N. K. Person, Question asking during tutoring, American Educational Research Journal 31 (1994) 104–137.

[5] D. Gupta, R. Pujari, A. Ekbal, P. Bhattacharyya, A. Maitra, T. Jain, S. Sengupta, Can taxonomy help? improving semantic question matching using question taxonomy, in: Proc. of ACL, 2018, pp. 499–513.

[6] A. Broder, A taxonomy of web search, SIGIR Forum 36 (2002).

[7] D. X. Zhou, L. Liu, A. Anubhai, M. Shandilya, S. Sigalas, W. Y. Wang, Z. Huang, Beyond accurate answers: Evaluating open-domain question answering in enterprise search, in: Proc. of CHIIR, 2023, p. 308–312.

[8] J. Suzuki, H. Taira, Y. Sasaki, E. Maeda, Question classification using HDAG kernel, in: Proce. of ACL Workshop on Multilingual Summarization and Question Answering, ACL, 2003, pp. 61–68.

[9] X. Li, D. Roth, Learning question classifiers, in: COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[10] E. Hovy, U. Hermjakob, D. Ravichandran, A question/answer typology with surface text patterns, in: Proc. of HLT, 2002, p. 247–251.

[11] V. Bolotova, V. Blinov, F. Scholer, W. B. Croft, M. Sanderson, A non-factoid question-answering taxonomy, in: Proc. of SIGIR, 2022, p. 1196–1207.

[12] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[13] J. Sarzynska-Wawer, A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, L. Okruszek, Detecting formal thought disorder by deep contextualized word representations, Psychiatry Research 304 (2021) 114135.

[14] Q. Liu, M. J. Kusner, P. Blunsom, A survey on contextual embeddings, arXiv preprint arXiv:2003.07278 (2020).

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[16] A. Roberts, H. W. Chung, G. Mishra, A. Levskaya, J. Bradbury, D. Andor, S. Narang, B. Lester, C. Gaffney, A. Mohiuddin, et al., Scaling up models and data with t5x and seqio, Journal of Machine Learning Research 24 (2023) 1–8.

[17] Y. Kementchedjhieva, I. Chalkidis, An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text, arXiv preprint arXiv:2305.05627 (2023).

[18] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models,

arXiv preprint arXiv:2307.09288 (2023).
[19] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022).
[20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.
[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[22] T. Schick, H. Schütze, Exploiting cloze questions for few shot text classification and natural language inference, arXiv preprint arXiv:2001.07676 (2020).
[23] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al., Lima: Less is more for alignment, arXiv preprint arXiv:2305.11206 (2023).
[24] J. McAuley, A. Yang, Addressing complex and subjective product-related queries with customer reviews, in: Proc. of WWW, 2016, pp. 625–635.