Leveraging Customer Reviews for E-commerce Query Generation

Yen-Chieh Lien[1], Rongting Zhang[2], F. Maxwell Harper[2], Vanessa Murdock[2], and Chia-Jung Lee[2]

[1] University of Massachusetts Amherst, USA** ylien@cs.umass.edu
[2] Amazon, Seattle, USA {rongtz,fmh,vmurdock,cjlee}@amazon.com

Abstract. Customer reviews are an effective source of information about what people deem important in products (e.g. "strong zipper" for tents). These crowd-created descriptors not only highlight key product attributes, but can also complement seller-provided product descriptions. Motivated by this, we propose to leverage customer reviews to generate queries pertinent to target products in an e-commerce setting. While there has been work on automatic query generation, it often relied on proprietary user search data to generate query-document training pairs for learning supervised models. We take a different view and focus on leveraging reviews without training on search logs, making reproduction more viable by the public. Our method adopts an ensemble of the statistical properties of review terms and a zero-shot neural model trained on adapted external corpus to synthesize queries. Compared to competitive baselines, we show that the generated queries based on our method both better align with actual customer queries and can benefit retrieval effectiveness.

Keywords: Query generation · Reviews · Weak learning · Zero-shot

1 Introduction

Customer reviews contain diverse descriptions about how people reflect the properties, pros and cons of the products that they have experienced. For example, properties such as "for underwater photos" or "for kayaking recording" were mentioned in reviews for action cameras, as well as "compact" or "strong zipper" for tents. These descriptors not only paint a rich picture of what people deem important, but also can complement and uncover shopping considerations that may be absent in seller-provided product descriptions. Motivated by this, our work investigates ways to generate queries that surface key properties about the target products using reviews.

Previous work on automatic query generation often relied on human labels or logs of queries and engaged documents (or items) [18–22] to form relevance signals for training generative models. Despite the reported effectiveness, the cost of acquiring high quality human labels is high, whereas the access to search logs is often only limited to site owners. As we approach the problem using reviews, it

** Work done while an intern at Amazon.

brings an advantage of not requiring any private, proprietary user data, making reproduction more viable by the public in general. Meanwhile, generation based on reviews is favorable as the outcome may likewise produce human-readable language patterns, potentially facilitating people-facing experiences such as related search recommendation.

We propose a simple yet effective ensemble method for query generation. Our approach starts with building a candidate set of "query-worthy" terms from reviews. To begin, we first leverage syntactic and statistical signals to build up a set of terms from reviews that are most distinguishable for a given product. A second set of candidate terms is obtained through a zero-shot sequence-to-sequence model trained according to adapted external relevance signals. Our ensemble method then devises a statistics-based scoring function to rank the combined set of all candidates, from which a query can be formulated by providing a desired query length.

Our evaluation examines two crucial aspects of query quality. To quantify how readable the queries are, we take the human-submitted queries from logs as ground truth to evaluate how close the generated queries are to them for each product. Moreover, we investigate whether the generated queries can benefit retrieval tasks, similar to prior studies [6,7,17]. We collect pairs of product descriptions and generated queries, both of which can be derived from public sources, to train a deep neural retrieval model. During inference, we take human-submitted queries on the corresponding product to benchmark the retrieval effectiveness. Compared with the competitive alternatives YAKE [1,2] and Doc2Query [6], our approach shows significantly higher similarity with human-submitted queries and benefits retrieval performance across multiple product types.

2 Related Work

Related search recommendation (or query suggestion) helps people automatically discover related queries pertinent to their search journeys. With the advances in deep encoder-decoder models [9,12], query generation [6,18,19,21,22] sits at the core of many recent recommendation algorithms. Sordoni et al. [19] proposed hierarchical RNNs [26] to generate next queries based on observed queries in a session. Doc2Query [6] adapted T5 [12] to generate queries according to input documents. Ahmad et al. [22] jointly optimized two companion ranking tasks, document ranking and query suggestion, by RNNs. Our approach differs in that we do not require in-domain logs of query-document relations for supervision.

Studies also showed that generated queries can be used for enhancing retrieval effectiveness [6, 7, 17]. Doc2Query [6] leveraged the generated queries to enrich and expand document representations. Liang et el [7] proposed to synthesize query-document relations based on MSMARCO [8] and Wikipedia for training large-scale neural retrieval models. Ma et al. [17] explored a similar zero-shot learning method for a different task of synthetic question generation, while Puri et al. [23] improve QA performance by incorporating synthetic questions. Our work resembles the zero-shot setup but differs in how we adapt external corpus particularly for e-commerce query generation.

Customer reviews have been adopted as a useful resource for summarization [24] and product question answering. Approaches to PQA [10, 11, 14, 16] often take in reviews as input, conditioned on which answers are generated for user questions. Deng et al [11] jointly learned answer generation and opinion mining tasks, and required both a reference answer and its opinion type during training phase. While our work also depends on reviews as input, we focus on synthesizing the most relevant queries without requiring ground-truth labels.

3 Method

Our approach involves a candidate generation phrase to identify key terms from reviews, and a selection phrase that employs an unsupervised scoring function to rank and aggregate the term candidates into queries.

3.1 Statistics-based approach

We started with a pilot study to characterize the opportunity of whether and how reviews could be useful for query generation. We found that a subset of terms in reviews resemble that of search queries, which are primarily composed of combinations of nouns, adjectives and participles to reflect critical semantics. For example, given a headphone, the actual queries that had led to purchases may contain nouns such as "earbuds" or "headset" to denote product types, adjectives such as "wireless" or "comfortable" to reflect desired properties, and participles such as "running" or "sleeping" to emphasize use cases.

Inspired by this, we first leverage part-of-speech analysis to scope down reviews to the three types of POS-tags. From this set, we then rely on conventional tf-idf corpus statistics to mine distinguishing terms salient in a product type but not generic across the entire catalog. Specifically, an importance score $I_t^D = \frac{p(t,R_D)}{p(t,R_G)}$ is used to estimate the salience of a term t in a product type D by contrasting its density in review set R_D to generic reviews R_G, where $p(t,R) = \frac{freq(t,R)}{\Sigma_{r \in R}|r|}$. Beyond unigrams, we also consider if the relative frequency of bigram phrases containing the unigrams $\frac{freq([t,t'],R_D)}{freq(t,R_D)}$ is above some threshold; in this case, bigrams will replace unigrams and become the candidates. We apply I_t^D to each review sentence, and collect top scored terms or phrases as candidates.

A straightforward way to form queries is to directly use the candidates as-is. We additionally consider an alternative which trains a seq2seq model using the candidates as weak supervision (i.e. encode review sentences to fit the candidates). By doing so, we anticipate the terms decoded during inference can generalize more broadly compared to a direct application. The two methods are referred to as Stats-base and Stats-s2s respectively.

3.2 Zero-shot generation based on adapted external corpus

Recent findings [7, 17] suggest that zero-shot domain adaptation can deliver high effectiveness given the knowledge embedded in large-scale language models

via pre-training tasks. With this, we propose to rely on fine-tuning T5 [12] on MSMARCO query-passage pairs to capture the notion of generic relevance, and apply the trained model to e-commerce reviews to identify terms that are more probable to be adopted in queries.

This idea has been experimented by Nogueira et al. [6], where their Doc2Query approach focused on generating queries as document expansion for improving retrieval performance. Different from [6], our objective is to generate queries that are not only beneficial to retrieval but also similar to actual queries in terms of syntactic forms. Thus, a direct application of Doc2Query on MSMARCO creates a gap in our case since MSMARCO "queries" predominantly follow a natural-language question style, resulting in generated queries of similar forms[3]. To tighten the loop, we propose to apply POS-tag analysis to MSMARCO queries and retain only terms that satisfy the selected POS-tags (i.e. nouns, adjectives and participles). For example, an original query "what does physical medicine do" is first transformed into "physical medicine" as pre-processing. After the adaptation, we conduct T5 seq2seq model training and apply it in a zero-shot fashion to generate salient terms based on input review sentences.

3.3 Ensemble approach to query generation

For a product p in the product type D, we employ both statistical and zero-shot approaches on its reviews to construct candidates for generating queries, which we denote as C_p. To select representative terms from the set, we devise a scoring function $S_t = freq(t, C_p) \cdot log(\frac{|\{p' \in D\}|}{|\{p'|p' \in D, t \in C_{p'}\}|})$ to rank all candidates, where higher ranked terms are more distinguishable for a specific product based on the tf-idf intuition. Given a desired query length n, we formulate the pseudo queries for a product by selecting all possible $\binom{k}{n}$ combinations from the top-k scored terms in the C_p set[4]. A final post-processing step removes any redundant words after stemming from the queries and adds product types if not already included.

4 Experiments

Our evaluation set is composed of products from three different product types, together with the actual queries[5] that were submitted by people who purchased those products on `Amazon.com`. As shown in Table 1, we consider *headphones*, *tents* and *conditioners* to evaluate our method across diverse product types, for which people tend to behave and shop differently with variances reflected in search queries. The query vocabulary size for conditioners, for instance, is about thrice the size of tents, with headphones sitting in-between the two.

As our approach disregards the actual queries for supervision, we primarily consider competitive baselines that do not involve using query logs. In particular,

[3] Original Doc2Query is unsuitable since question-style queries are rare in e-commerce.
[4] Our experiment sets k=3 and n=1, 2, 3 per its popularity in generic search queries.
[5] Note that we use actual data only for the purpose of evaluation not training.

we compare to the unsupervised approach YAKE [1,2] which reportedly outperforms a variety of seminal key word extraction approaches, including RAKE [4], TextRank [3] and SingleRank [5] methods. In addition, we leverage the zero-shot Doc2Query model on adapted corpus as our baseline to reflect the absence of e-commerce logs. For generation, we initialize separate Huggingface **T5-base** [12] weights with conditional generation head and fine-tune for Stats-s2s and Doc2Query models respectively. Training is conducted on review sentences broken down by NLTK. For retrieval, we fine-tune a Sentence-Transformer [25] **ms-marco-TinyBERT**[6] pre-trained with MSMARCO data, which was shown to be effective for semantics matching. Our experiments use a standard AdamW optimizer with learning rate 0.001 and $\beta_1, \beta_2 = (0.9, 0.999)$, and conduct 2 and 4 epochs training on a batch size of 16 respectively for generation and retrieval.

Table 1. Statistics of the three product types used in the experiments. For each product type, the dev and test split respectively contains 500 disjoint products.

	Headphone		Tent		Conditioner	
	Dev	Test	Dev	Test	Dev	Test
# of reviews	23,165	23,623	19,208	18,734	17,055	17,689
# of sentences	102,281	103,771	97,553	97,320	68,691	70,829

4.1 Intrinsic Similarity Evaluation

Constructing readable and human-like queries is desirable since it is practically useful for applications such as related search recommendation. A natural way to reflect readability is to evaluate the similarity between the generated and customer-submitted queries since the latter is created by human. In practice, we consider customer-submitted queries that had led to at least 5 purchases on the corresponding products as ground-truth queries, to which the generated queries are then compared. We use conventional metrics adopted in generative tasks including corpus BLEU and METEOR for evaluation. The results in Table 2 show that our ensemble approach consistently achieves the highest similarity with human-queries across product types, suggesting that the statistical and zero-shot methods could be mutually beneficial.

4.2 Extrinsic Retrieval Evaluation

We further study how the generated queries can benefit e-commerce retrieval. Our evaluation methodology leverages pairs of generated queries and product descriptions to train a retrieval model and validates its quality based on actual queries. During training, we fine-tune a Sentence-Transformer based on top-3 generated queries of each product. For each query, we prepare its corresponding relevant product description, together with 49 negative product descriptions randomly sampled from the same product type. During inference, instead of

[6] https://www.sbert.net/docs/pretrained-models/ce-msmarco.html

Table 2. The similarity in BLEU and METEOR between generated queries and real queries. $\star$ stands for p-value < 0.05 in T-test compared to the second best performing method in each column. The bottom shows example generated queries by ensemble.

	Headphone		Tent		Conditioner	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
YAKE	0.1014	0.1371	0.2794	0.2002	0.3143	0.1998
Doc2Query	0.1589	0.1667	0.3684	0.2145	0.4404	0.264
Stats-base	0.1743	0.2001	0.3294	0.2201	0.4048	0.2723
Stats-s2s	0.1838	0.2004	0.321	0.2189	0.3931	0.2641
Ensemble	**0.2106***	**0.2024**	**0.394***	**0.2334***	**0.5047***	**0.2956***
Examples	noise cancelling headphone truck driver headphone hearing aids headphone		lightweight tent alps backpacking tent air mattresses queen tent		detangling conditioner shea moisture conditioner dry hair conditioner	

generated queries, we use customer-submitted queries to fetch descriptions from the product corpus, and an ideal retrieval model should rank the corresponding product description at the top. We also include BM25 as a common baseline. Table 3 shows that Doc2Query and the ensemble methods are the most effective and are on par in aggregate, with some variance in different product types. Stats-s2s slightly outperforms Stats-base overall, which may hint a potential for better generalization.

Table 3. The retrieval effectiveness for queries generated by baselines and our method.

	Headphone			Tent			Conditioner		
	MRR	P@1	P@10	MRR	P@1	P@10	MRR	P@1	P@10
BM25	0.28	0.19	0.06	0.43	0.29	0.11	0.56	0.47	0.14
YAKE	0.23	0.11	0.07	0.46	0.34	0.11	0.54	0.43	0.14
Doc2Query	0.28	0.18	**0.08**	**0.49**	**0.40**	0.12	0.58	**0.49**	0.15
Stats-base	0.28	0.16	0.07	0.44	0.29	0.12	0.54	0.42	0.15
Stats-s2s	0.27	0.17	0.07	0.44	0.32	0.12	0.56	0.46	**0.16**
Ensemble	**0.29**	**0.20**	0.07	0.46	0.33	**0.13**	**0.59**	0.48	0.15

5 Conclusion

This paper connected salient review descriptors with zero-shot generative models for e-commerce query generation, without requiring human labels or search logs. The empirical results showed that the ensemble queries both better resemble customer-submitted queries and benefit training effective rankers. Besides MSMARCO, our future plan seeks to incorporate other publicly available resources such as community question-answering threads to generalize the notion of relevance. It is worth to consider ways to combine weak labels with few strong labels and dive deep into the impact of employing different hyper-parameters. A user study that characterizes the extent to which the generated queries can reflect people's purchase intent will further help qualitative understanding.

References

1. Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. A Text Feature Based Automatic Keyword Extraction Method for Single Documents. In: Proceedings of the 40th European Conference on Information Retrieval, pp. 684–691 (2018).
2. Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. YAKE! Keyword extraction from single documents using multiple local features. Information Sciences, vol. 509, pp. 257–289 (2020).
3. Rada Mihalcea and Paul Tarau. TextRank: bringing order into texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004).
4. Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic Keyword Extraction from Individual Documents. In: Text Mining: Theory and Applications. vol. 1, pp. 1–20 (2010).
5. Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence, pp. 855–860 (2008).
6. Rodrigo Nogueira, Wei Yang, Jimmy J. Lin, and Kyunghyun Cho. Document Expansion by Query Prediction. ArXiv (2019).
7. Davis Liang , Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. Embedding-based Zero-shot Retrieval through Query Generation. ArXiv (2020).
8. Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, ArXiv (2016).
9. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880 (2020).
10. Yiren Liu and Kuan-Ying Lee. E-commerce Query-based Generation based on User Review. ArXiv (2020).
11. Yang Deng, Wenxuan Zhang, and Wai Lam. Opinion-aware Answer Generation for Review-driven Question Answering in E-Commerce. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management. pp. 255–264 (2020).
12. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In: Journal of Machine Learning Research (JMLR), vol. 21, pp. 1–67 (2020).
13. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171–4186 (2019).
14. Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. Review-Driven Answer Generation for Product-Related Questions in E-Commerce. In: Proceedings of the 12th ACM International Web Search and Data Mining Conference, pp. 411–419 (2019).

15. Huajie Shao, Jun Wang, Haohong Lin, Xuezhou Zhang, Aston Zhang, Heng Ji, and Tarek Abdelzaher. Controllable and Diverse Text Generation in E-commerce. In: Proceedings of the Web Conference 2021, pp. 2392–2401 (2021).
16. Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. Product-Aware Answer Generation in E-Commerce Question-Answering. In: Proceedings of the 12th ACM International Web Search and Data Mining Conference, pp. 429–437 (2019).
17. Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1075–1088 (2021).
18. Ruey-Cheng Chen and Chia-Jung Lee. Incorporating Behavioral Hypotheses for Query Generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 3105–3110, (2020).
19. Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and JianYun Nie. A hierarchical recurrent encoderdecoder for generative context-aware query suggestion. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, pp. 553–562 (2015).
20. Jyun-Yu Jiang and Wei Wang. RIN: Reformulation Inference Network for Context-Aware Query Suggestion. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 197–206 (2018).
21. Kyungho Kim, Kyungjae Lee, Seung-won Hwang, Young-In Song, and Seungwook Lee. Query Generation for Multimodal Documents. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp. 659–668 (2021).
22. Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. Context Attentive Document Ranking and Query Suggestion. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 385–394 (2019).
23. Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro. Training Question Answering Models From Synthetic Data. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 5811–5826 (2020).
24. Xueying Zhang, Yunjiang Jiang, Yue Shang, Zhaomeng Cheng, Chi Zhang, Xiaochuan Fan, Yun Xiao, and Bo Long. DSGPT: Domain-Specific Generative Pre-Training of Transformers for Text Generation in E-commerce Title and Review Summarization. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2146-2150 (2021).
25. Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982–3992 (2019).
26. David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Tech. rep. ICS 8504. San Diego, California: Institute for Cognitive Science, University of California (Sept. 1985).