Addressing Vocabulary Gap in E-commerce Search

Subhadeep Maji[1], Rohan Kumar[1], Manish Bansal[1], Kalyani Roy[2], Mohit Kumar[1], Pawan Goyal[2*]
[1]Flipkart, [2]IIT Kharagpur
{subhadeep.m,rohankumar,manish.bansal}@flipkart.com,kroy@iitkgp.ac.in,mohitkum@gmail.com,pawang@cse.iitkgp.ac.in

ABSTRACT

E-commerce customers express their purchase intents in several ways, some of which may use a different vocabulary than that of the product catalog. For example, the intent for "women maternity gown" is often expressed with the query, "ladies pregnancy dress". Search engines typically suffer from poor performance on such queries because of low overlap between query terms and specifications of the desired products. Past work has referred to these queries as vocabulary gap queries. In our experiments, we show that our technique significantly outperforms strong baselines and also show its real-world effectiveness with an online A/B experiment.

CCS CONCEPTS

• **Information systems** → **Query reformulation**.

KEYWORDS

Query rewriting; product search; siamese networks

ACM Reference Format:
Subhadeep Maji[1], Rohan Kumar[1], Manish Bansal[1], Kalyani Roy[2], Mohit Kumar[1], Pawan Goyal[2]. 2019. Addressing Vocabulary Gap in E-commerce Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331323

1 INTRODUCTION

A significant fraction of queries in E-commerce search suffer from *vocabulary gap* (VG). VG expresses the difficulty faced by users in expressing their need, in a manner which could best match products from product catalog. For example, the query "ladies pregnancy dress" expresses the same need as "women maternity gown". But, the query does not perform well due to vocabulary mismatch between query terms and product catalog definition of relevant products. Query Rewriting has been applied as an effective technique to bridge this gap between user queries and the documents to improve retrieval performance [1, 7, 15, 17]. Most recent techniques either rely on sufficient implicit feedback [6, 7, 14] or restrictive dataset specific assumptions [15, 17] limiting their general applicability.

Null and low recall VG queries are mostly tail queries as also observed in [17]. From an editorial analysis of a random sample of queries exhibiting VG from query logs from Flipkart, it was observed that approximately 97% of these queries convey intent similar to a *well performing* (WP) query. WP queries are defined as head queries with high click-through rate [4]. These results are corroborated with past research, which indicates that a significant fraction of *tail* queries are *head* queries expressed differently [8, 16]. Motivated by this observation, we propose a supervised classification technique by modelling the probability of rewriting VG queries to semantically similar WP queries. A measure of semantic similarity between queries is learned by projecting similar queries to nearby points in space using Manhattan-LSTM (MaLSTM) [12]. We propose an improvement of this base similarity measure by incorporating Co-Attention [11] on MaLSTM states which captures an interdependent notion of similarity between queries. More specifically, our model for semantic similarity learns an attention distribution on terms of one query dependent on the other query as an additional layer in MaLSTM. Our formulation ensures retrieval performance by limiting the rewrite candidates to this set of known WP queries. Our experiments show that the proposed technique is significantly better in addressing vocabulary gap in comparison to strong baselines in multiple experimental settings. We also report a large-scale online A/B experiment run at Flipkart, where we achieved 1.37% improvement in click-through rate, 6.74% improvement in add-to-cart ratio and a reduction of null search ratio by 15.84% over the production system in place.

In summary, we make the following contributions: (i) We propose a novel formulation for addressing VG in e-commerce search by rewriting VG queries to semantically similar WP queries. (ii) We propose a novel Co-Attentive MaLSTM semantic similarity measure, which incorporates an interdependent measure of semantic similarity on query pairs. (iii) We demonstrate effectiveness of our technique in multiple experimental settings including an online A/B experiment at Flipkart.

2 RELATED WORK

Techniques for Query Rewriting (QRW), Reformulation, and Expansion have been shown to improve the retrieval performance of search engines on queries with vocabulary gap. Automatic relevance feedback based techniques [18] require multi-phase retrieval which is prohibitively expensive for commercial search engines. Later techniques incorporate implicit user feedback in the form of click-through rate [2], query co-occurrence [9], and co-clicked query similarity [1, 5] to rewrite queries. More recent techniques pose query rewriting as a machine translation task [6, 14]. However, they do not explicitly optimize for retrieval performance and a two stage framework to handle this is developed in [7]. A fundamental drawback of these techniques is their need for sufficient implicit user feedback which limits their applicability to vocabulary gap queries that are either infrequent, null or have low recall. Recent work on such queries [15] infers taxonomy constraints for relaxed

[*]Rohan Kumar and Manish Bansal contributed equally to this work.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331323

versions of the query to retrieve relevant products. Subsequent work [17] uses domain specific attribute taggings and hand crafted rules. These techniques are limited by their dataset specific assumptions which are not easily extendable in a general e-commerce setting.

3 PRELIMINARIES & DATASET CREATION

Dataset Creation: A team of domain experts at Flipkart periodically analyses a large random sample of queries from query logs to identify and categorize (e.g., spell mistakes, ranking-issues, vocabulary gap) poorly performing queries. From this analysis, we obtained 5k queries identified to be VG, which is a substantial fraction of poorly performing queries. The human experts also provided an alternate query for each VG query which reflects the same user intent expressed in the original query and has better term overlap with product catalog. For example, in the query pair ("scratch jeans, distressed jeans"), "scratch jeans" is a VG query and "distressed jeans" is an alternate query which expresses the same user intent. We observed that approximately 97% of these alternate queries were WP queries. We consider frequent queries (> 70 impressions per week) from past 1 year query logs from Flipkart with high click-through rate (approximately > 30%) as the set of WP queries. In our formulation, we treat the human labeled dataset of VG to WP queries as ground truth. We will refer to this ground truth dataset as $\mathcal{D} = \{(x_1, x_2)\}$, where x_1 is a VG query and x_2 is a WP query, and this pair is called a query rewrite pair.

Past work on web query understanding highlights that a significant fraction of tail queries are head queries expressed differently [8, 16]. Since most VG queries are tail queries [17], our findings corroborate with the past work.

Limitations of product co-clicks and user query reformulation behavior We evaluate the applicability of recent work on QRW [7] by comparing co-clicked products on 5k labelled query-rewrite pairs. A huge fraction (89%) of these pairs have no co-clicked products. Of the remaining pairs, 70% have a Jaccard similarity < 0.2 on the clicked product set. We also evaluate whether query-rewrite pairs from the above 5k set co-occur in query logs from Flipkart as part of the same search session. We observe that only 8.24% of the 5k query pairs exhibit such co-occurrence. This limits the applicability of techniques based on user reformulation behaviour in query logs.

4 REWRITING VOCABULARY GAP QUERIES: PROPOSED FRAMEWORK

We formulate the problem of rewriting VG queries to WP queries as a supervised classification task. The probability of rewriting a VG query x_1 to a WP query x_2 is given by,

$$p(y = 1 \,|\, (x_1, x_2)) = \sigma\left(w f_{\texttt{attn}}(x_1, x_2) + b\right) \tag{1}$$

where w, b are parameters of the model and $f_{\texttt{attn}}(x_1, x_2)$ denotes a similarity measure between queries. Below we describe the proposed Co-Attentive MaLSTM to model this similarity.

4.1 Modelling Semantic Similarity using Co-Attentive MaLSTM

Consider a query pair $x = (x_1, x_2)$, where $x_1 = (x_1^{(1)}, x_1^{(2)}, \ldots, x_1^{(n)})$ and $x_2 = (x_2^{(1)}, x_2^{(2)}, \ldots, x_2^{(m)})$, n and m denoting the number of words in these queries, respectively. We consider two queries to be *similar* if they express the same user intent, e.g., "nike hood tshirts, nike hoodies". More specifically, we are interested in learning a similarity measure between VG and WP queries, which we learn from the ground truth dataset $\mathcal{D}$. Our first attempt at modelling the similarity is based on MaLSTM [12], which also serves as building block for the proposed Co-Attentive MaLSTM .

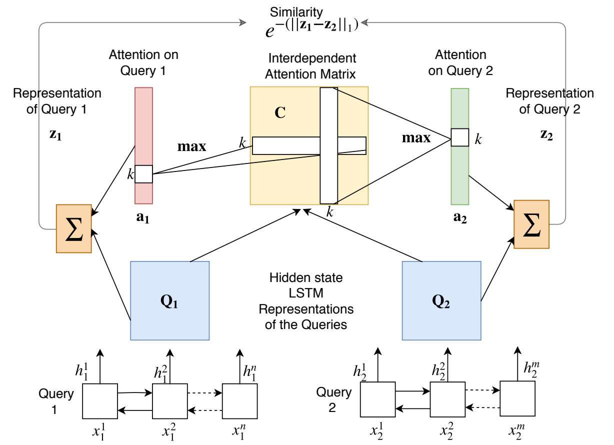

Figure 1: Co-Attentive MaLSTM architecture. The C matrix captures the interdependence between LSTM representations of queries Q_1, Q_2. The max pooling results in attention on a particular term of one query to depend on the most similar terms in the other query.

In MaLSTM, queries x_1 and x_2 are passed through bi-directional Siamese LSTMs (i.e., parameters across LSTMs are shared). The $\ell 1$ norm of the final hidden state representations of the two queries $h_1^{(n)}$ and $h_2^{(m)}$ serves as a measure of similarity in the following function

$$f(x_1, x_2) = \exp\left(-\left\|h_1^{(n)} - h_2^{(m)}\right\|_1\right) \tag{2}$$

where $f \in [0, 1]$ is a function of similarity between the two queries. Equation (2) however falls short in modelling similarity because while the final hidden state alone might not be fully representative of the query, this representation is also independent of the other query. Further, we observe that all terms across the query pairs do not contribute equally towards the similarity measure based on the following characterizations of the similarity :

(i) Consider the *similar* query pair "nike hood tshirts, nike hoodies". The similarity is really expressed between "hood tshirts, hoodies" (i.e., it holds for brands other than "nike"). One way of addressing this redundancy is by focusing less on the term "nike" while modelling the similarity.

(ii) Consider the *similar* query pair "ladies pregnancy dress, women maternity gown". It is apparent that strict subsets of two queries are related to each other, i.e., "pregnancy dress" to "maternity gown" and "ladies" to "women". One way to address this is to learn a similarity measure across subsets of the two queries, such that combined measure is representative of the overall similarity.

Addressing the above characterization, we propose an improved measure of similarity by incorporating an attention distribution over the hidden state vectors of the two queries. Intuitively, across the query pair, the attention on a particular term of a query depends upon how *similar* it is to terms of the other query. Therefore, the

Query pair	women maternity gown	maternity gown	gown
ladies pregnancy dress	0.98 (0.78)	0.56 (0.47)	0.11 (0.15)
pregnancy dress	0.67 (0.44)	0.97 (0.91)	0.31 (0.16)
dress	0.05 (0.05)	0.11 (0.05)	0.27 (0.35)

Query pair	nike hoodies	hoodies	adidas hoodies
nike hood tshirts	0.99 (0.95)	0.28 (0.16)	0.96 (0.35)
hood tshirts	0.44 (0.19)	0.98 (0.90)	0.23 (0.29)

Table 1: The measure of similarity learned by Co-Attentive MaLSTM and MaLSTM (in brackets) across subsets of two query pairs "ladies pregnancy dress, women maternity gown" and "nike hood tshirts, nike hoodies". In both cases, proposed measure of similarity is high for both original as well as all the relevant subset query pairs (highlighted in color).

inter-dependent notion of attention between queries models similarity at a more granular level of terms in context of queries. This interdependent notion of attention was first proposed in computer vision literature on visual question answering [11] and named Co-Attention. Figure 1 provides a block level illustration of our model.

Concretely, we introduce $W_A \in \mathbb{R}^{d \times d}$ as a learnable attention matrix. Let, $Q_1 \in \mathbb{R}^{d \times n}$ be n hidden state vectors of x_1 and $Q_2 \in \mathbb{R}^{d \times m}$ be m hidden state vectors of x_2. We define the following similarity matrix $C \in \mathbb{R}^{n \times m}$

$$C = \tanh\left(Q_1^T W_A Q_2\right)$$

The similarity matrix C is the bilinear form corresponding to the attention matrix W_A, with element-wise $tanh$ introduced to rescale similarities in $[-1, 1]$ range. From the similarity matrix, we compute the attention on the k^{th} term of x_1 as its maximum similarity to terms in x_2 and vice versa.

$$a_1^{(k)} = \max_j \left\{C_{k,j}\right\} \quad a_2^{(k)} = \max_j \left\{C_{j,k}\right\}$$

Intuitively, the max pooling operation aids in modelling the overall query similarity as combination of similarities over term pairs by co-attending on term pairs most similar to each other across the queries. We apply a softmax on attentions to form distributions per query,

$$a_1 = \texttt{softmax}\left(a_1^{(k)}\right) \quad a_2 = \texttt{softmax}\left(a_2^{(k)}\right)$$

Combining the attention distributions with hidden state vectors of x_1 and x_2 we have the following representations,

$$z_1 = \sum_k a_1^{(k)} h_1^{(k)} \quad z_2 = \sum_k a_2^{(k)} h_2^{(k)}$$

Using z_1, z_2 in Equation (2), we define the Co-Attentive MaLSTM similarity function as

$$f_{\texttt{attn}}(x_1, x_2) = \exp\left(-\|z_1 - z_2\|_1\right) \tag{3}$$

We use Equation (3) as a measure of similarity to rewrite VG queries to WP queries. Table 1 illustrates our model's ability to address the aforementioned characterizations (i) and (ii). Specifically, the learned measure of similarity is high between relevant query pair subsets (e.g "pregnancy dress, maternity gown") and by focusing less on brand term "nike", it learns a high similarity between "nike hood tshirts, adidas hoodies" as well, while MaLSTM is not able to generalize to this case.

4.2 Rewriting Vocabulary Gap Queries

The similarity measure in Equation (3) is trained in an end-to-end manner with the task of rewriting VG to WP queries. For each VG query x_1 corresponding to the positive pair $(x_1, x_2) \in \mathcal{D}$, 50 negative examples are randomly selected from WP queries for supervised training. We refer to the dataset thus constructed as $\mathcal{D}^L = \{((x_1, x_2), y)\}$, where $y = 1$ if $(x_1, x_2) \in \mathcal{D}$ or 0 otherwise. The model is trained using binary cross-entropy loss w.r.t. $y \in \mathcal{D}^L$, as modelled using Equation (1).

To improve model's performance, we employ max negative sampling [13]. Specifically, out of 50 negative query pairs for each positive pair, we select 20 pairs having the highest probability of rewrite (hence, incorrectly classified) as defined by Equation (1) as negatives in every training epoch. We will refer to our model as Co-Attentive MaLSTM -QRW.

5 EXPERIMENTS

5.1 Baselines & Methods

We compare our proposed model against four baselines: Word Centroid Distance (WCD), Word Mover's Distance (WMD) [10], BERT Sentence Pair Classification [3] and MaLSTM [12]. WCD measures the cosine similarity between query vector representations calculated by summing/averaging over embedded word vectors. WMD measures the notion of similarity as the minimum amount of distance that the embedded words of first query need to "travel" to reach the embedded words of the other. We compare against BERT fine-tuned on the labelled query-rewrite dataset for the binary Sentence Pair Classification task using aggregate classification embeddings ($[CLS]$). The fine-tuning was done with the same experimental settings as in the original paper [3]. The model MaLSTM-QRW is obtained by replacing f_{attn} in (3) with f from (2) corresponding to MaLSTM. We train a 100 dimensional word embedding while treating queries occurring in a single session (query chain) [9] as a document. We use the same embedding across all the competing models except BERT, which uses BERT pre-trained embeddings. We set an appropriate class weight ratio to account for the class imbalance in our training data $\mathcal{D}^L$ (20 is to 1), while training all the supervised models.

5.2 Experimental Settings

We evaluate our model against baselines in three experimental settings. First, we report recall@k on a random holdout subset of 1000 query pairs from $\mathcal{D}$. Second, we report human labeled product retrieval quality scores for VG queries rewritten to WP queries. Third, we report the results of an online A/B experiment conducted at Flipkart. We observe that VG is most exhibited by queries belonging to Clothing category at Flipkart, thus we conduct product quality evaluation and online A/B in the Clothing category. In all our experiments, for each VG query, the entire WP query set (roughly 80k for Clothing category) is considered as candidate set for rewriting and is ranked by probability of rewrite. For WMD and WCD, ranking is based on the similarity scores.

Model	R@1	R@3	R@5	R@30
WCD	11.60	11.60	21.25	43.80
WMD	14.39	14.39	22.53	40.38
BERT	43.79	51.60	53.21	54.13
MaLSTM-QRW	44.82	59.31	65.48	70.68
Co-Attentive MaLSTM -QRW	**47.12**	**62.06**	**67.93**	**74.48**

Table 2: Recall of baselines and Co-Attentive MaLSTM on random holdout set of 1000 query pairs from $\mathcal{D}$.

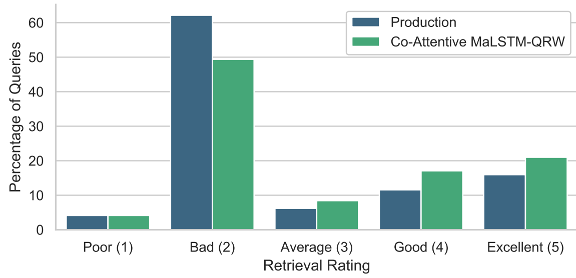

Figure 2: Comparison of product retrieval quality rating on scale from 1 (Poor) to 5 (Excellent) for Production and Co-Attentive MaLSTM -QRW, respectively

Performance on holdout dataset: We perform a 70/10/20 train/ development/ test random split of the ground truth dataset $\mathcal{D}$. For training each baseline and model, we construct $\mathcal{D}^L$ as discussed in Section 4.2. The loss on the development set was used for early stopping of the training procedure, a common practice for regularization of neural networks. Table 2 reports Recall@$\{1, 3, 5, 30\}$ for the competing models. Co-Attentive MaLSTM -QRW outperforms BERT and MaLSTM-QRW consistently at all k. While BERT's R@1 is very competitive, the improvement in recall for large k is very marginal. WMD and WCD fair poorly as baselines demonstrating the efficacy of our formulation of semantic similarity beyond the term level similarity represented by embeddings.

Retrieval Quality Evaluation: To mimic production setting & limitations, we restrict scope of the experiment to rewriting a VG query to a single WP query. This necessitates a high precision operating point. We obtain a large set of queries $\mathcal{D}_{poor}$ from Flipkart search logs with low CTR and low recall, a segment known to exhibit VG [7]. First, we obtain the high precision operating point for classification on $\mathcal{D}_{poor}$ by manual evaluation of rewrite correctness on a random sample of 1000 queries (achieving 87% precision at 37% recall). We use this operating point to obtain 2000 query-rewrite pairs (exclusive from the earlier sample of 1000) from $\mathcal{D}_{poor}$, we call this as $\mathcal{D}_{test}$. We now obtain top 10 retrieved products as per these rewrites (Co-Attentive MaLSTM -QRW) and from a sophisticated production system that handles low recall queries by adding relevant or dropping irrelevant terms to increase recall. The exact details of the production system is out of scope. The retrievals (top 10 products) from both the systems were now rated by domain experts on a 5 point scale from Poor (1) to Excellent (5). The results from this manual evaluation are reported in Figure 2. Clearly, our method results in **12.78%** reduction in Poor and Bad queries (rating $\leq$ 2) and **10.55%** increase in Good and Excellent queries (rating $\geq$ 4) over the production baseline.

Online A/B Experiment: We tested the Co-Attentive MaLSTM -QRW against the production system (discussed above) at Flipkart in a standard A/B testing configuration where production behaves as control and our model as treatment condition. The experiment affected only low recall and low CTR queries. 15% of the user base was randomly assigned to each condition and the experiment ran for 20 days. The query volume affected by the experiment was roughly 75k. We report significant improvements in click-through-rate (CTR) and add-to-cart ratio (fraction of searches leading to products being added to shopping cart, an important metric for e-commerce search relevance) against the production system. In addition, we report a significant reduction in Null search ratio which is defined as fraction of queries resulting in no products retrieved. Table 3 reports the exact improvements in the metrics.

Metric	% Improvement
CTR	1.37%
Add-to-cart ratio	6.74%
Null search ratio	15.84%

Table 3: Results of online A/B experiment comparing production system with Co-Attentive MaLSTM -QRW

6 CONCLUSION & FUTURE WORK

In this paper, we investigated the problem of vocabulary gap in e-commerce queries. Our empirical study suggested that most vocabulary gap queries are well performing queries expressed differently. Using this observation as motivation, we developed a novel inter-dependent measure of semantic similarity between pair of queries to rewrite vocabulary gap queries to well performing queries. Our approach ensures retrieval performance by restricting rewrites to well performing queries. In future, we plan to conduct further experiments, by evaluating our model in more product categories at Flipkart, and explore better ways of modelling similarity. The BERT baseline shows promising results and we would explore ways to incorporate it in our model. To further improve the retrieval performance of VG queries, we would also like to extend our approach to merge the results of multiple WP query rewrites.

REFERENCES

[1] Ioannis Antonellis, Hector Garcia Molina, and Chi Chao Chang. 2008. Simrank++: query rewriting through link analysis of the click graph. *Proceedings of the VLDB Endowment* 1, 1 (2008), 408–421.

[2] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2002. Probabilistic query expansion using query logs. In *WWW (WWW'02)*. ACM, 325–332.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Doug Downey, Susan Dumais, and Eric Horvitz. 2007. Heads and tails: studies of web search with common and rare queries. In *SIGIR*. ACM, 847–848.

[5] Bruno M Fonseca, Paulo Golgher, Bruno Pôssas, Berthier Ribeiro-Neto, and Nivio Ziviani. 2005. Concept-based interactive query expansion. In *CIKM (CIKM'05)*. ACM, 696–703.

[6] Jianfeng Gao, Xiaodong He, Shasha Xie, and Alnur Ali. 2012. Learning lexicon models from search logs for query expansion. In *EMNLP (EMNLP'2012)*. Association for Computational Linguistics, 666–676.

[7] Yunlong He, Jiliang Tang, Hua Ouyang, Changsung Kang, Dawei Yin, and Yi Chang. 2016. Learning to rewrite queries. In *CIKM (CIKM'16)*. ACM, 1443–1452.

[8] Shuai Huo, Min Zhang, Yiqun Liu, and Shaoping Ma. 2014. Improving Tail Query Performance by Fusion Model. In *CIKM (CIKM '14)*. 559–568.

[9] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating query substitutions. In *WWW (WWW'06)*. ACM, 387–396.

[10] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML (ICML'2015)*. 957–966.

[11] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-image Co-attention for Visual Question Answering. In *In Proc. of NIPS (NIPS'16)*. 289–297.

[12] Jonas Mueller and Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. In *AAAI (AAAI'16)*. 2786–2792.

[13] Jinfeng Rao, Hua He, and Jimmy Lin. 2016. Noise-Contrastive Estimation for Answer Selection with Deep Neural Networks. In *CIKM (CIKM '16)*. 1913–1916.

[14] Stefan Riezler and Yi Liu. 2010. Query rewriting using monolingual statistical machine translation. *Computational Linguistics* 36, 3 (2010), 569–582.

[15] Gyanit Singh, Nish Parikh, and Neel Sundaresan. 2012. Rewriting null e-commerce queries to recommend products. In *WWW (WWW'2012)*. ACM, 73–82.

[16] Yangqiu Song, Haixun Wang, Weizhu Chen, and Shusen Wang. 2014. Transfer Understanding from Head Queries to Tail Queries. In *CIKM (CIKM '14)*. 1299–1308.

[17] Zehong Tan, Canran Xu, Mengjie Jiang, Hua Yang, and Xiaoyuan Wu. 2017. Query Rewrite for Null and Low Search Results in eCommerce. In *SIGIR ECOM (ECOM'2017)*.

[18] Jinxi Xu and W Bruce Croft. 1996. Query expansion using local and global document analysis. In *SIGIR (SIGIR'96)*. ACM, 4–11.