This file has been cleaned of potential threats.

If you confirm that the file is coming from a trusted source, you can send the following SHA-256 hash value to your admin for the original file.

cac45d85786a1a839b8d1b562cd0bdf4cd8bf5a8f2874fb9025fbc38148b548

To view the reconstructed contents, please SCROLL DOWN to next page.
Addressing Vocabulary Gap in E-commerce Search

Subhadeep Maji\textsuperscript{1}, Rohan Kumar\textsuperscript{1}, Manish Bansal\textsuperscript{1}, Kalyani Roy\textsuperscript{2}, Mohit Kumar\textsuperscript{1}, Pawan Goyal\textsuperscript{2}\textsuperscript{∗}
\textsuperscript{1}Flipkart, \textsuperscript{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

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
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 \( D = \{(x_1, x_2)\} \), 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,

\[
p(y=1| (x_1, x_2)) = \sigma (w f_{\text{ext}}(x_1, x_2) + b)
\]

where \( w, b \) are parameters of the model and \( f_{\text{attn}}(x_1, x_2) \) 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 \( D \). Our first attempt at modelling the similarity is based on MaLSTM [12], which also serves as building block for the proposed Co-Attentive MaLSTM.

In MaLSTM, queries \( x_1 \) and \( x_2 \) are passed through bi-directional Siamese LSTMs (i.e., parameters across LSTMs are shared). The \( l_1 \) 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

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

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
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-QRST.

### 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-QRST is obtained by replacing \( \text{fatten} \) 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>R@1</th>
<th>R@3</th>
<th>R@5</th>
<th>R@30</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCD</td>
<td>11.00</td>
<td>11.60</td>
<td>21.25</td>
<td>43.80</td>
</tr>
<tr>
<td>WMD</td>
<td>14.39</td>
<td>14.39</td>
<td>22.53</td>
<td>40.38</td>
</tr>
<tr>
<td>BERT</td>
<td>43.79</td>
<td>51.60</td>
<td>54.21</td>
<td>54.15</td>
</tr>
<tr>
<td>MaLSTM-QRST</td>
<td>44.82</td>
<td>59.31</td>
<td>62.48</td>
<td>70.68</td>
</tr>
<tr>
<td>Co-Attentive MaLSTM-QRST</td>
<td>47.12</td>
<td>62.06</td>
<td>67.93</td>
<td>74.48</td>
</tr>
</tbody>
</table>

Table 2: Recall of baselines and Co-Attentive MaLSTM on random holdout set of 1000 query pairs from \( \mathcal{D} \).
which is defined as fraction of queries resulting in no products retrieved. Table 3 reports the exact improvements in the metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>% Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTR</td>
<td>1.37%</td>
</tr>
<tr>
<td>Add-to-cart ratio</td>
<td>6.74%</td>
</tr>
<tr>
<td>Null search ratio</td>
<td>15.84%</td>
</tr>
</tbody>
</table>

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


