

Centrality-aware Product Retrieval and Ranking

Hadeel Saadany¹, Swapnil Bhosale², Samarth Agrawal³,
Diptesh Kanojia², Constantin Orăsan³, and Zhe Wu³

¹Institute for People-Centred AI, University of Surrey, United Kingdom

²Centre for Translation Studies, University of Surrey, United Kingdom

³eBay Inc., USA

{hadeel.saadany, s.bhosale, d.kanojia, c.orasan}@surrey.ac.uk,
{samagrwal, zwu1}@ebay.com

Abstract

This paper addresses the challenge of improving user experience on e-commerce platforms by enhancing product ranking relevant to users' search queries. Ambiguity and complexity of user queries often lead to a mismatch between the user's intent and retrieved product titles or documents. Recent approaches have proposed the use of Transformer-based models, which need millions of annotated query-title pairs during the pre-training stage, and this data often does not take user intent into account. To tackle this, we curate samples from existing datasets at eBay, manually annotated with *buyer-centric relevance* scores and *centrality* scores, which reflect how well the product title matches the users intent. We introduce a User-intent Centrality Optimization (UCO) approach for existing models, which optimises for the user intent in semantic product search. To that end, we propose a dual-loss based optimisation to handle hard negatives, *i.e.*, product titles that are semantically relevant but do not reflect the user's intent. Our contributions include curating challenging evaluation sets and implementing UCO, resulting in significant product ranking efficiency improvements observed for different evaluation metrics. Our work aims to ensure that the most buyer-centric titles for a query are ranked higher, thereby, enhancing the user experience on e-commerce platforms.

1 Introduction

Achieving a user-focused experience on e-commerce platforms (eBay, Walmart, Amazon, Etsy, JD) is enabled by ranking products relevant to the user's intent expressed via the search query. However, user queries often do not fully reflect the underlying intent behind the search terms used within the query. For example, ambiguous queries like *'iphone 13'*, or *'i5 pc 1tb 16gb 8gb gpu'* can lead to many variants. To aggravate the challenge

further, user queries can consist of lexical terms with alphanumeric characters, which do not reveal a semantic match within existing product titles. Information Retrieval (IR) systems depend upon semantic similarity/distance between words or phrases used in the search query and the product title. Therefore, ranking the product titles based on only lexical or only semantic query-title match can be a particularly challenging problem, as detailed in the examples below:

Ambiguous Queries Some queries can be ambiguous and do not clearly reflect the user's intention. From the same example above, for a query like *'iPhone 13'*, the user is most likely looking to buy the base variant or to check out other device variants. However, this intent is not clear from the query, and the system can even rank *'iPhone 13 cover'* among the top retrieved products. Hence, a major challenge faced by search systems is to retrieve titles that are likely to be relevant to the user intent at high ranks, and push down negative titles such as *'iPhone 13 cover'* which have semantic proximity to positive titles within the embedding space of the computational model but may not reflect users underlying objective.

Repetition Similar to the example above, the repetition of the exact string of words from a user's query, such as *'iPhone 13'*, in both relevant and irrelevant titles often renders embeddings-based similarity approaches futile as the proximity of positive and negative titles in the embedding space may not be reflective of their relevancy. In such cases, human annotation towards user intent for a query-title pair is needed to establish a clear ranking among products retrieved by the model.

Alphanumeric Queries Queries such as *'S2716DG'* consist of alphanumeric characters where a letter or number can signify important

detail for the product/model. For example, based on the naming convention of PC monitors, a single letter defines the type of panel in the product. In this case, the Dell S2716DG is a 27-inch monitor with a TN panel, and changing the last letter to P would refer to a monitor with an IPS panel. Similarly, product colour or a specific spare part can be identified from such queries. Unless the product title contains this alphanumeric sequence of characters, the semantic similarity between the query and a non-intended product can be high, thus misleading the system.

In this paper, we investigate the challenges listed above and take a two-step approach to improve product retrieval and ranking. We curate samples from existing internal datasets at eBay consisting of user search queries paired with retrieved product titles on their platform. These datasets are human-annotated based on detailed guidelines to produce two buyer-centric relevance annotations. First, a widely used relevance ranking schema where query-title pairs are provided a ranked class from among Bad (1), Fair (2), Good (3), Excellent (4) and Perfect (5), where ‘*perfect*’ reflects an exact query-title pair match, *i.e.*, the annotator is very confident that the user found precisely what they were looking for, while ‘*bad*’ reflects no match between the product and the need expressed in the query (Jiang et al., 2019; Kang et al., 2016). Second, query-title pairs are annotated with a binary *centrality score*, obtained from majority voting over multiple human annotations, *i.e.*, indicating whether the item reflects the need expressed in the query. The difference between centrality and relevance scoring is that the latter detects whether an item is an outlier, a surprising addition to the recall set, or the item centrally matches the expectations. Figures 1 and 2 show two examples of the centrality annotation for the same query, “Thomas Sabo charm”. Figure 1 shows a product central to the query since, based on purchase data, this query typically reflects the user’s need for a charm (a small ornament worn on a necklace or bracelet). On the other hand, the product in Figure 2 is not central to the user’s intent as it is a *Thomas Sabo charm* attached to a bracelet; the user intent is a charm, not a bracelet. Although both titles are semantically related to the query, based on the *degree of specificity* expressed in the query, the product in Figure 2 becomes less central to the user’s intent and gets annotated with 0 as its centrality score whereas the product in Figure 1 receives 1. We use an internal human-annotated



Figure 1: Central Title: Thomas sabo charms with 18k Rose gold pearl



Figure 2: Non-central title: Thomas Sabo charm club bracelet with detachable dragonfly charm

dataset for this task. Henceforth, we refer to it as Internal Graded Relevance or IGR dataset.

We extract challenging evaluation sets from the IGR dataset based on the challenges discussed above. Our objective is to increase the retrieval and ranking efficiency of product search by training a model for query-title pairs that integrates the user intent in the similarity algorithm. Given the search query, we propose using a user-intent centrality optimisation (UCO) step for existing models which cater to the ranking of relevant products. Further, we propose utilising a dual-loss based optimisation to address the query-title pairs which constitute hard negatives, *i.e.*, query-title pairs where the product title is semantically relevant to the user’s query but is annotated as non-central to the user intent, or has *Bad* or only *Fair* annotated relevancy.

We hypothesise that there is an unwanted semantic proximity of such negative titles to their search queries in the model embeddings space. To improve search, we optimise the existing ranking model with our dual-loss-based optimisation approach, ensuring that the retrieval algorithm should have the most “typical” titles for a query ranked highly than other titles which *may be relevant but are not typical*. Our contributions are 1) curating challenging evaluation sets that cater to this problem and 2) user-intent centrality optimisation (UCO), which results in a stark improvement on all the evaluation sets.

2 Related Work

Our work is based on a two-step approach to improve product ranking given a search query for retrieving items. Existing literature on traditional candidate retrieval research focused on learning query rewrites (Bai et al., 2018; Guo et al., 2008) as an indirect approach to bridge the vocabulary gap between queries and documents/titles. Some approaches, including latent semantic indexing with matrix factorization (Deerwester et al., 1990), and with probabilistic models (Hoffman, 1990), and semantic hashing with an auto-encoder (Salakhutdinov and Hinton, 2009), have been proposed. Most of these are unsupervised models based on word co-occurrence in documents/product titles.

Modern IR systems deploy semantic retrieval models as bi-encoders (Muennighoff, 2022) or Siamese networks (Chiang and Chen, 2021) comprising two encoders. Most existing studies focus on designing or pre-training encoders with different representation learning approaches (Gao et al., 2011; Salakhutdinov and Hinton, 2009; Yih et al., 2011; Huang et al., 2020; Liu et al., 2020). Representative works, namely, the Deep Semantic Similarity Model (DSSM) (Huang et al., 2013), and CDSSM (Shen et al., 2014b), are some of the earliest methods which utilise a deep neural network (DNN) using clickthrough data. Subsequently, CNNs (Gao et al., 2014; Shen et al., 2014a,b; Severyn and Moschitti, 2015) and RNNs (Palangi et al., 2014, 2016) have been utilised for semantic retrieval. Recently, new models, including DRRM (Guo et al., 2016) and Duet (Mitra et al., 2017) were developed to include traditional IR lexical matching (*e.g.*, exact matching, term importance) within semantic retrieval performed by DNNs. However, (Mitra et al., 2018) argues that most works proposed in this direction focus on the ranking stage, where the optimisation objectives differ from candidate title retrieval. To further improve the performance of semantic retrieval, Transformer-based Pre-trained Language models (PTLMs) like BERT (Devlin et al., 2018) and ERNIE (Zhang et al., 2019) have been leveraged (Fuchs et al., 2020; Wang et al., 2024; Liu et al., 2021). Using larger pre-trained models, semantic retrieval has observed a significant performance improvement and generalisation for retrieval but without a specific focus on ambiguous or alphanumeric queries, which is what we essentially address in this paper.

Further, interaction-based approaches (Moe, 2003; Long et al., 2012; Gu et al., 2020; Yates et al., 2021; Zou et al., 2020; Dai et al., 2023) have also been widely used for IR systems, which further go into semantic matching to model for query-document/title interaction using DNNs (Lu and Li, 2013; Mitra et al., 2017; Wan et al., 2016; Zhao et al., 2020; Kabir et al., 2022). Most of these approaches focus on user personalisation needs, and often rely on hand-crafted rules. Often, such approaches cannot cache the document embeddings offline for faster retrieval, and may be inefficient for retrieval (Liu et al., 2021). (Su et al., 2018) use the results of an online survey and search logs from a commercial product search engine to show that product search falls into categories like Target Finding, Decision Making and Exploration. (Yao et al., 2021) propose Personal Word-embeddings for Personalized Search (PEPS) which uses as additional layer trained on user embeddings and personal logs.

While personalised embeddings and interaction-based approaches improve ranking performance for ambiguous user queries, our work focuses on dealing with similar challenges using a different approach *infusing centrality-awareness*. To be considered an impactful solution for the challenges at hand, we believe that product ranking approaches can be more generalised compared to personalised embeddings, improving the base retrieval with a focus on user intent. Our approach utilises two existing loss functions that cater to the task and optimise the retrieval model, which can be used at both stages, retrieval and ranking.

3 Methodology

3.1 Baseline Model: eBERT

For training our system, we employ the in-house multilingual eBERT¹ model. eBERT is trained on item/product data from eBay and general domain (Wikipedia and RefinedWeb) text. The item data used to train this model consists of approximately 3 billion item titles. We also test another eBERT variant, eBERT-siam, which is fine-tuned to generate similar embeddings for item titles using a Siamese network. This model is designed specifically for tasks related to similarity search on query and product titles. Both models are used offline to perform experiments and are optimised with UCO to note performance changes for retrieval and ranking.

¹eBERT Language Model

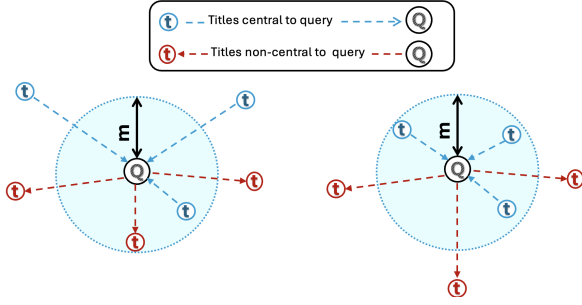


Figure 3: The figure shows how the loss function algorithm works with hard negatives. The algorithm targets those non-central titles (red) that are inside the margin.²

3.2 User-intent Centrality Optimization (UCO)

We perform UCO as an optimisation step to overcome the problem of top-ranked, hard negative query-title pairs that are semantically relevant but not central to user intent. Thus, we fine-tune the baseline model with a supervised binary classification task on product centrality. Then, based on our hypothesis for transfer learning capabilities, we employ the knowledge learned from the domain information of centrality optimisation as an inductive bias to boost the ranking capability of a retrieval model, thereby, optimising the ranking task for our challenging evaluation sets. We employ dual-loss optimisation, as explained in the next section.

3.3 Dual-Loss Based Optimisation

Multiple Negative Ranking Loss (MNRL) (Henderson et al., 2017) is the first loss function we employ. MNRL quantifies the difference between positive and negative samples for a query. MNRL is used to create a clear distinction between relevant (positive) and irrelevant (negative) data points, achieved by minimising the distance between the query and positive samples while maximising it for multiple negative samples. Multiple negatives provide more context, enabling the optimisation to discriminate between varying degrees of irrelevance. Mathematically, it can be represented as follows:

$$\text{MNRL} = \sum_{i=1}^P \sum_{j=1}^N \max(0, f(q, p_i) - f(q, n_j) + \text{margin}) \quad (1)$$

where P is the number of positive titles, N is the number of negative titles, q is the query, f is our similarity function, which is cosine similarity, and margin is a hyperparameter defining the optimum distance between positive and neg-

ative titles defined by the centrality of the user-intent. The MNRL minimises the distance between (q, p_i) while it simultaneously maximises the distance (q, n_j) for all P and N titles.

Online Contrastive Loss (OCL) is a variant of Contrastive Loss (CL) (Carlsson et al., 2020). OCL attends to negative pairs that have a lower distance than the positive pairs with the largest distance, as well as, the positive pairs that have a higher distance than the lowest distance of negative pairs, *i.e.*, the hard cases in a batch, and computes the loss only for these cases. It selects hard positive (positives that are far apart) and hard negative pairs (negatives that are close), and backpropagates only for such pairs. OCL can be represented as follows:

$$\text{OCL} = Y * D + (1 - Y) * \max(\text{margin} - D, 0)^2 \quad (2)$$

where Y is our centrality score between the query and title, it will be 1 if the title is central to the user intent and 0 if it is not. The D variable is the function that returns the distance between the query and title embeddings, which is the cosine similarity in our case. The \max function takes the largest value of 0 and the margin minus the distance. The negative samples (centrality = 0) should have a distance of at least the margin value which we empirically set during training. This means that if we define some radius/margin, all the central titles should fall inside this margin, and all the non-central ones should fall outside.

MNRL primarily reduces the distance between positive pairs out of a large set of possible candidates and hence works particularly well when the dataset has a significant number of positives, which caters to the dataset skew in our case. However, MNRL does not push dissimilar pairs away. Therefore, we combine both losses for better optimisation (see below for ablation results).

Figure 3 explains how our approach proposes this dual loss optimisation. We address query-title pairs where semantic distance is not proportional to the centrality specifications defined by previously annotated data. As can be seen from the figure, dual loss optimisation ensures that for each query (Q), the maximum intra-class distance (blue arrows) is smaller than the minimum inter-class distance (the red arrow). We define a radius/margin m , for all the central product titles, while all the non-central product titles fall outside the margin. Please note that the loss penalises the model for non-central titles having a distance to Q less than m .

²Adopted from (Hadsell et al., 2006) with modifications.

Eval Split	# Corpus	# Dev-Q	# Test-Q
<i>CQ</i>	187469	5776	17325
<i>CQ-balanced</i>	46561	5776	17325
<i>CQ-common-str</i>	12508	2117	6351
<i>CQ-alphanum</i>	162115	4111	12333

Table 1: Data Distribution in each split. Q -> queries

4 Experiment Setup

4.1 Dataset Curation

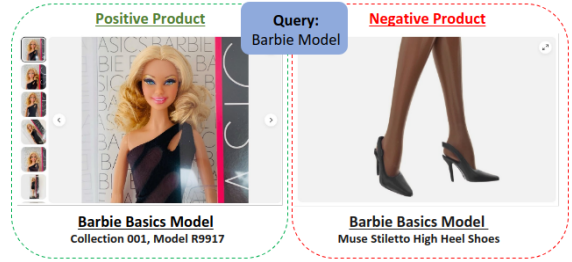
We preprocess all query-title pairs from the IGR dataset by filtering out non-English pairs to ensure linguistic consistency and relevance. Once preprocessed, we select queries that have both the corresponding positive titles (relevancy > 3) and negative titles (relevancy < 3) from the IGR dataset. This selection forms our initial split, referred to as *Common Queries (CQ)*. We observed a notable imbalance towards positive query-title pairs in *CQ*, stemming from the inherent nature of e-commerce product listings and the data collection strategy highlighted in Section 1, which emphasises capturing relevant matches. To address this imbalance and ensure a fair comparison, we introduce a balanced version of *CQ*, where the number of positive and negative query-product title pairs is approximately equal, referred to as *CQ-balanced*.

Upon examining the query-title pairs, as also discussed in Section 1, we found that often, the exact string of a query appears in both positive and negative product titles. We isolate these query-title pairs to form our third split, named *CQ-common-str* (see Figure 4). This task necessitates considering both, user centrality and semantic connections between the query and product titles. We conduct a correlation test, and observe that Pearson, Kendall and Spearman correlations between the *graded relevance score* and the *binary centrality score* are 0.78, 0.73 and 0.77, respectively, validating our assumption that both types of scores are highly correlated and hence the ranked results are expected to conform with the overall pattern of the dataset.

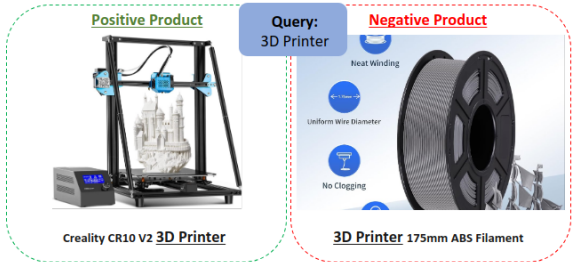
Lastly, to facilitate the evaluation of our proposed methodology specifically on alphanumeric query-title pairs, we create a separate split containing only queries and titles with alphanumeric characters, referred as *CQ-alphanum*. For each evaluation split, all the positive and negative titles constitute the retrieval corpus, while we create distinct development and test query sets in an 80:20 ratio. Table 1 shows the number of entries in the corpus and query sets for each split. The

development query set assists in selecting the best-performing UCO model (*i.e.*, during optimisation on user-intent centrality), while the unseen test query set validates the ranking capability of UCO.

Visual Samples We believe that the split, *CQ-common-str*, presents the most demanding evaluation scenario, requiring the model to simultaneously differentiate the semantic relationships of the strings in both positive and negative product titles.



(a) The sub-string “Barbie Model” is a part of both positive and negative product titles.



(b) The sub-string “3D Printer” is a part of both positive and negative product titles.

Figure 4: Examples of query-title pairs from the *CQ-common-str* split. Both, positive and negative product titles have high semantic correlation to the user query, however only the positive product title exhibits a central idea/intent.

Figure 4a shows common query string in positive and negative titles: query “barbie model”, positive title “Barbie Top Model Summer Doll 2008 Ginger Hair” (the real doll), negative title “Barbie Model Pointed Toe Fashion High Heel Shoes” (only the shoes). Similarly, Figure 4b shows the query “3d printer”, positive title “Creality CR10 V2 3D Printer”, and the negative title “3D Printer 175mm ABS Filament Made in UAE Premium Quality”; where the negative title is just the printer filaments. Note that query-title pairs such as these are challenging for traditional IR methods too, which use lexical matching.

Encoder	UCO	Precision@ k (\uparrow)			Recall@ k (\uparrow)			NDCG@ k (\uparrow)			MRR (\uparrow)
		3	5	10	3	5	10	3	5	10	@10
CQ test											
BERT	\times	16.20	13.03	8.93	11.31	14.41	18.83	0.1912	0.1818	0.1833	0.2771
eBERT	\times	20.71	17.25	12.54	14.46	19.19	26.26	0.2392	0.2330	0.2430	0.3415
	\checkmark	64.76	55.74	39.22	49.63	63.92	79.65	0.7439	0.7488	0.7672	0.8189
eBERT (siam)	\times	55.25	48.33	34.90	42.36	56.09	72.22	0.6315	0.6428	0.6704	0.7263
	\checkmark	66.25	57.16	40.20	51.18	65.79	81.66	0.7635	0.7698	0.7886	0.8347
CQ-balanced test											
BERT	\times	7.13	4.94	2.95	21.26	24.58	29.33	0.1824	0.1961	0.2115	0.1862
eBERT	\times	9.72	6.94	4.22	29.02	34.58	42.07	0.2428	0.2657	0.2899	0.2495
	\checkmark	28.57	18.15	9.50	85.40	90.42	94.62	0.7851	0.8059	0.8197	0.7789
eBERT (siam)	\times	25.99	16.68	8.89	77.66	83.08	88.59	0.6888	0.7112	0.7291	0.6784
	\checkmark	29.19	18.39	9.58	87.26	91.58	95.43	0.8046	0.8225	0.8351	0.7965
CQ-common-str test											
BERT	\times	9.41	6.31	3.65	28.15	31.47	36.35	0.2532	0.2669	0.2828	0.2579
eBERT	\times	12.62	8.64	5.00	37.79	43.10	49.92	0.3272	0.3491	0.3714	0.3315
	\checkmark	32.03	19.58	9.92	95.84	97.65	98.87	0.9091	0.9166	0.9206	0.8979
eBERT (siam)	\times	29.93	18.76	9.68	89.57	93.58	96.50	0.8194	0.8361	0.8456	0.8063
	\checkmark	32.12	19.64	9.92	96.11	97.94	98.93	0.9117	0.9193	0.9226	0.9003
CQ-alphanum test											
BERT	\times	20.54	16.65	11.47	13.45	17.32	22.82	0.2333	0.2176	0.2226	0.3350
eBERT	\times	23.35	19.54	13.77	15.53	20.76	27.85	0.2630	0.2516	0.2617	0.3739
	\checkmark	64.58	57.27	40.35	44.05	59.97	77.00	0.7119	0.7094	0.7344	0.8018
eBERT (siam)	\times	60.67	54.10	38.54	41.32	57.10	74.20	0.6652	0.6654	0.6951	0.7618
	\checkmark	67.10	59.70	41.81	46.07	62.72	79.76	0.7375	0.7371	0.7609	0.8171

Table 2: Evaluating the efficacy of the proposed UCO on the all test sets, using different encoder backbones. Precision and Recall values are shown in (%); higher values are preferred.

4.2 Implementation Details

We optimise both the encoder backbones on the centrality score classification-train split for a maximum of 10 epochs. During training, we run two sequential evaluators on both the centrality scores and the retrieval ranking in the curated IGR datasets. First, an evaluator that will compute the embeddings for both query and title and use them to calculate the cosine similarity. If the similarity is above a threshold, we have a central title. Second, given a query and the corpus of all titles, the evaluator finds the most relevant product title to the query (top 3, 5 and 10 titles). During optimisation, we save the checkpoint that performs best on the second evaluator. For all experiments, we use a batch size of 32, with the Adam optimiser and $2e - 05$ as the learning rate, and 0.01 as weight decay. Optimising one encoder backbone using the above parameters takes 30 hours on a single NVIDIA V100 GPU.

For evaluation, we use cosine similarity as scoring function.

Evaluation Metrics We use different existing evaluation metrics to measure the overall model performance. Precision@ k measures the proportion of relevant products in the top- k recommendations (considering their relevance), while Recall@ k measures the proportion of relevant products that were retrieved among all relevant products (irrespective of their rank). NDCG (Järvelin and Kekäläinen, 2002) measures the ranking quality by comparing the recommended items’ order against an ideal ranking. As a result, NDCG considers both the relevance and rank of the recommended products. Mean Reciprocal Rank (MRR) evaluates the average rank of the first relevant item across all queries. A high MRR is an indication of being able to provide users with relevant products ranked as high as possible.

5 Results and Discussion

Considering various aspects like retrieval and ranking quality, we analyse model performance using a diverse set of metrics (explained in §4.1). We also perform an ablation test on the eBERT model to identify the contribution of both loss functions, MNRL and OCL, and discuss the qualitative analysis below. Table 2 displays the results for each of the evaluation splits, *CQ*, *CQ-balanced*, *CQ-common-str* and *CQ-alphanumeric*. Across each split, a consistent pattern emerges: the incorporation of UCO leads to a substantial improvement in product retrieval performance across all metrics. This improvement is evident regardless of whether the backbone encoder employed is eBERT or eBERT-siam. This highlights UCO’s capability to enhance an existing model’s embedding space, enabling it to capture semantic relationships between user queries and product titles attuned to the user intent, thus retrieving products with high user centrality. It is evident that BERT, a publicly available model, was unable to capture query-title relations given it was not pre-trained on internal data. Even with internal models, the results without UCO show the challenge posed by these evaluation splits curated for this work. For alphanumeric queries, the NDCG performance improvement ranges from 7% points for the base model to 47% points, including the model fine-tuned with the Siamese approach, demonstrating the efficacy of UCO. For query-titles with common strings, it ranges from 8% to 58% points. We also see similar improvements in all metrics, for the other two evaluation sets.

Loss Ablation We conducted a quick ablation test over the *CQ* evaluation split. For this test, we fine-tuned the eBERT and eBERT-siam models using individual loss functions and their combination, which is our finalised approach. From Table 3, it is clear that the combination of both loss functions helps improve performance for both models. We evaluate this using both NDCG and MRR evaluation metrics. When employed individually, MNRL seems to outperform OCL in both metrics. Overall, dual-loss based optimisation emerges as a clear winning strategy.

Qualitative Analysis We discuss the performance improvement shown by UCO with two examples in Figures 5 and 6, *shown in the Appendix below*. We use the eBERT-siamese model to rank retrieved products with and without UCO optimisa-

Loss	eBERT		eBERT-siam	
	NDCG@5	MRR@10	NDCG@5	MRR@10
MNRL	0.7139	0.7899	0.7254	0.8016
OCL	0.5497	0.6559	0.5812	0.6978
MNRL + OCL	0.7488	0.8189	0.7698	0.8347

Table 3: Ablation experiment to study the efficacy of MNRL and OCL losses when taken individually; higher values are preferred.

tion. In Figure 5, search query ‘1080’ from the test set retrieves more ‘central’ products when UCO optimised model is used, *i.e.*, graphics card variants. Similarly, on the use of the alphanumeric search query in Figure 6, most relevant products are ranked on top, *i.e.*, keyboard with the same product identifier, showing how UCO model optimisation helps rank relevant products on top.

6 Conclusion and Future Work

This work addresses product search queries that represent an important challenge for e-commerce platforms. The main challenge occurs when the retrieved titles are semantically relevant, but not *central* to the user-intent as is reflected by the specificity of the query. The challenge is even greater with ambiguous queries where the same query string is present in both relevant and irrelevant titles as well as when queries are alphanumeric. We address the semantic complexity of these challenging query-title pairs by fine-tuning existing internal models with a user-intent centrality optimisation (UCO) step to infuse information about the typicality of query-title pairs. The retrieval model performance showed significant improvement with several hard example datasets with a dual-loss based optimisation approach, which pays attention to negative pairs that have a lower distance than the positive pairs with the largest distance. The dual-loss based optimisation helps in separating the irrelevant pairs of queries and titles while keeping the distance smaller for relevant query-title pairs. The improvement in ranking performance demonstrated by our approach helps identify and categorise what users intend to find online when they search the platform.

In future, we aim to restructure queries in our hard-negative pairs to be less ambiguous. Leveraging GenAI-based prompt engineering and explainability using approaches like chain-of-thought, we can investigate titles that indicate typical queries, *aligning them closer to the user intent*, and moving towards *explainable product retrieval*.

References

- Xiao Bai, Erik Ordentlich, Yuanyuan Zhang, Andy Feng, Adwait Ratnaparkhi, Reena Somvanshi, and Aldi Tjahjadi. 2018. Scalable query n-gram embedding for improving matching and relevance in sponsored search. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 52–61.
- Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International conference on learning representations*.
- Hung-Yun Chiang and Kuan-Yu Chen. 2021. [A BERT-based Siamese-structured retrieval model](#). In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 163–172, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).
- Shitong Dai, Jiongnan Liu, Zhicheng Dou, Haonan Wang, Lin Liu, Bo Long, and Ji-Rong Wen. 2023. Contrastive learning for user sequence representation in personalized product search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 380–389.
- S Deerwester, Susan T Dumais, and R Harshman. 1990. Indexing by latent semantic analysis. 1990. *Journal of the American Society for Information Science*, 41(6).
- 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*.
- Gilad Fuchs, Yoni Acriche, Idan Hasson, and Pavel Petrov. 2020. Intent-driven similarity in e-commerce listings. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2437–2444.
- Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng. 2014. Modeling interestingness with deep neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2–13.
- Jianfeng Gao, Kristina Toutanova, and Wen-tau Yih. 2011. Clickthrough-based latent semantic models for web search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 675–684.
- Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, Lixin Zou, Yiding Liu, and Dawei Yin. 2020. Deep multifaceted transformers for multi-objective ranking in large-scale e-commerce recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2493–2500.
- Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 55–64.
- Jiafeng Guo, Gu Xu, Hang Li, and Xueqi Cheng. 2008. A unified and discriminative model for query refinement. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 379–386.
- Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.
- Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.
- Thomas Hoffman. 1990. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, 1990*, pages 50–57.
- Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.
- Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. [Cumulated gain-based evaluation of ir techniques](#). *ACM Trans. Inf. Syst.*, 20(4):422446.
- Yunjiang Jiang, Yue Shang, Rui Li, Wen-Yun Yang, Guoyu Tang, Chaoyi Ma, Yun Xiao, and Eric Zhao. 2019. [A unified neural network approach to e-commerce relevance learning](#). In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data, DLP-KDD '19*, New York, NY, USA. Association for Computing Machinery.
- Md Ahsanul Kabir, Mohammad Al Hasan, Aritra Mandal, Daniel Tunkelang, and Zhe Wu. 2022. Ordsim: Ordinal regression for e-commerce query similarity prediction. *arXiv preprint arXiv:2203.06591*.
- Daekook Kang, Wooseok Jang, and Yongtae Park. 2016. [Evaluation of e-commerce websites using fuzzy hierarchical topsis based on e-s-qual](#). *Applied Soft Computing*, 42:53–65.

- Yiding Liu, Yulong Gu, Zhuoye Ding, Junchao Gao, Ziyi Guo, Yongjun Bao, and Weipeng Yan. 2020. Decoupled graph convolution network for inferring substitutable and complementary items. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2621–2628.
- Yiding Liu, Weixue Lu, Suqi Cheng, Daiting Shi, Shuaiqiang Wang, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained language model for web-scale retrieval in baidu search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3365–3375.
- Bo Long, Jiang Bian, Anlei Dong, and Yi Chang. 2012. Enhancing product search by best-selling prediction in e-commerce. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2479–2482.
- Zhengdong Lu and Hang Li. 2013. A deep architecture for matching short texts. *Advances in neural information processing systems*, 26.
- Bhaskar Mitra, Nick Craswell, et al. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.
- Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th international conference on world wide web*, pages 1291–1299.
- Wendy W Moe. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology*, 13(1-2):29–39.
- Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.
- Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and R Ward. 2014. Semantic modelling with long-short-term memory for information retrieval. *arXiv preprint arXiv:1412.6629*.
- Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707.
- Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978.
- Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382.
- Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014a. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 101–110.
- Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014b. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web*, pages 373–374.
- Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User intent, behaviour, and perceived satisfaction in product search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 547–555.
- Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. 2016. A deep architecture for semantic matching with multiple positional sentence representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing bert for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys*, 56(7):1–33.
- Jing Yao, Zhicheng Dou, and Ji-Rong Wen. 2021. Clarifying ambiguous keywords with personal word embeddings for personalized search. *ACM Trans. Inf. Syst.*, 40(3).
- Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.
- Wen-tau Yih, Kristina Toutanova, John C Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 247–256.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.
- Xiangyu Zhao, Haochen Liu, Hui Liu, Jiliang Tang, Weiwei Guo, Jun Shi, Sida Wang, Huiji Gao, and Bo Long. 2020. Memory-efficient embedding for recommendations. *arXiv preprint arXiv:2006.14827*.
- Lixin Zou, Long Xia, Yulong Gu, Xiangyu Zhao, Weidong Liu, Jimmy Xiangji Huang, and Dawei Yin. 2020. Neural interactive collaborative filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 749–758.

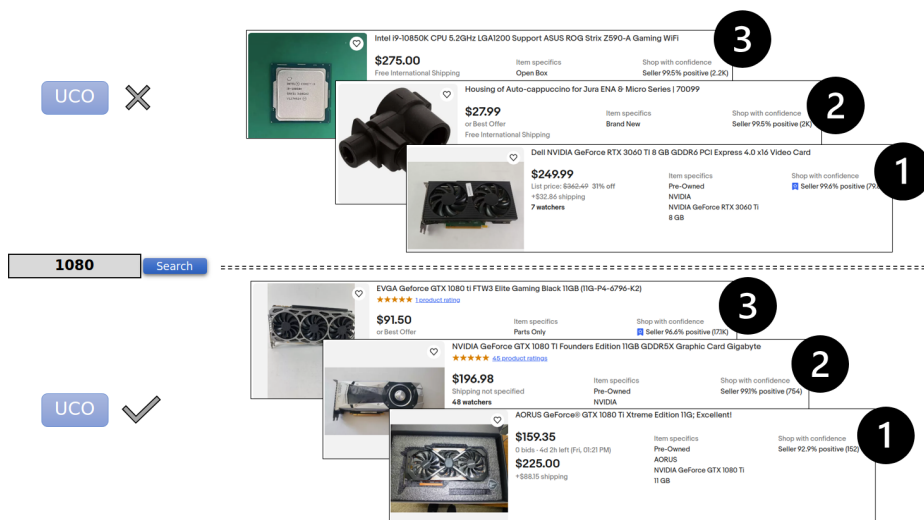


Figure 5: Qualitative comparison of the proposed UCO on a sample from the *CQ-common-str* test set, when using the eBERT (siam) as the encoder backbone. We showcase the top-3 retrieved product titles for both encoders.

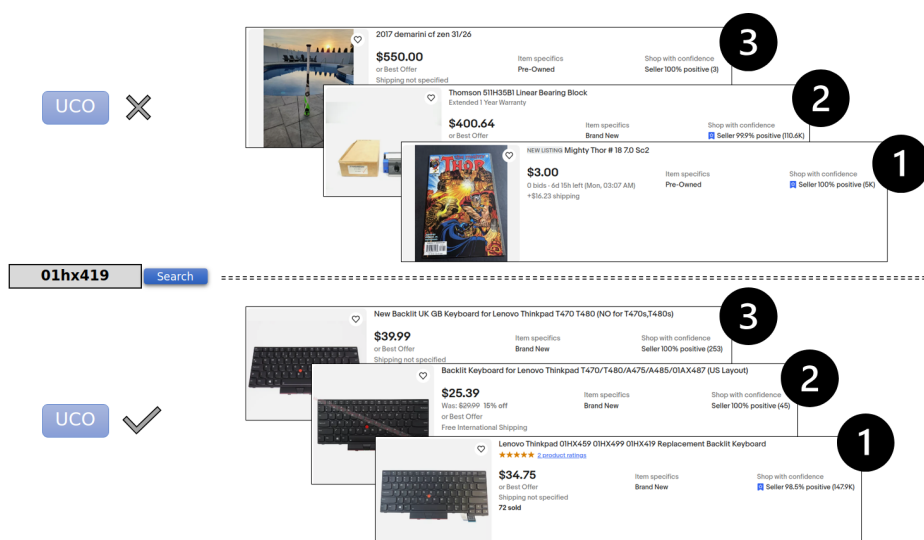


Figure 6: Qualitative comparison of the proposed UCO on a sample from the *CQ-alphanum* test set, when using the eBERT (siam) as the encoder backbone. We showcase the top-3 retrieved product titles for both encoders.