Knowledge Distillation for Efficient and Effective Relevance Search on E-commerce[⋆]

Nguyen Vo[*], Hongwei Shang[*], Zhen Yang, Juexin Lin, Seyed Danial Mohseni Taheri and Changsung Kang

Walmart Global Tech
Sunnyvale, California, USA, 94086

Abstract

Ensuring the relevance of text between user queries and products is vital for e-commerce search engines to enhance user experience and facilitate finding desired products. Thanks to deep learning models' capabilities in semantic understanding, they have become the primary choice for relevance matching tasks. In real-time e-commerce scenarios, representation-based models are commonly used due to their efficiency. On the other hand, interaction-based models, while offering better effectiveness, are often time-consuming and challenging to deploy online. The emergence of large language model (LLM) has marked a significant advancement in relevance search, presenting both value and complexity when applied to e-commerce domain. To address these challenges, we propose a novel framework to distill a highly effective interaction-based LLM into a low latency representation-based architecture (i.e. student model). To further increase effectiveness of the LLM, we propose to use soft human labels and items' attributes. Our student model is trained to mimic the margin between a relevant document and a less relevant product outputted from the LLM. Experimental results showed that our model improves both relevancy and engagement metrics. Our model increased NDCG@5 by 1.30% and the number of sessions with clicks by 0.214% compared with a production system.

Keywords

cross encoders, dual encoders, knowledge distillation

1. Introduction

Major online shopping platforms such as Walmart, Ebay and Amazon cater to millions of users daily with a vast array of products. Search engines play a crucial role in helping users find what they are looking for, but in the realm of commercial e-commerce, search engines typically rely heavily on user engagement signals to understand query intent and provide the best possible search results [1, 2, 3]. Search queries from users are usually segmented into head, torso and tail queries. Head and torso queries generally provide enough user engagement data to train machine learning models for retrieving and reranking relevant items. However, it is difficult to effectively retrieve and rerank the most relevant products for tail queries due to the lack of engagement data. Ensuring that search results align closely with different types of queries from users is vital for maintaining customer satisfaction and trust over time.

eCom'24: ACM SIGIR Workshop on eCommerce, July 18, 2024, Washington, DC, USA

[*] Both authors contributed equally to this research.

nguyen.vo@walmart.com (N. Vo); hongwei.shang@walmart.com (H. Shang); zhen.yang@walmart.com (Z. Yang); juexin.lin@walmart.com (J. Lin); syeyddanial.mohseni@walmart.com (S. D. M. Taheri); changsung.kang@walmart.com (C. Kang)

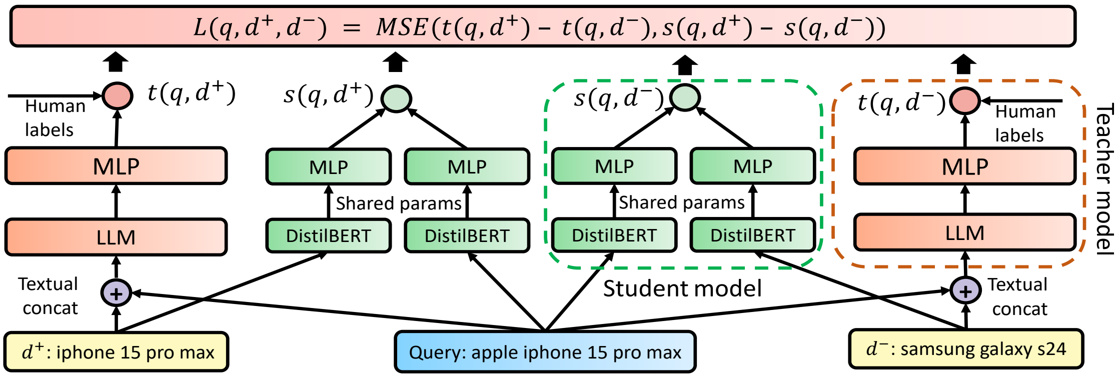

Figure 1: Our proposed knowledge distillation framework. Item d^+ is more relevant than item d^-. The LLM/teacher model is trained with human labels while the DistilBERT/student model is trained by using soft targets outputted from the LLM

Traditional methods of matching queries to products have limitations, particularly in bridging the vocabulary gap. To address this challenge, advanced neural network models have emerged as a powerful solution. These models, categorized into representation-based and interaction-based models, offer different approaches to text matching. Representation-based models encode queries and product titles into fixed-dimensional vectors separately, and compute cosine similarity as a semantic matching feature for reranking [4, 5, 6, 7, 8, 9], enabling efficient online computation but potentially sacrificing detailed matching information.

On the other hand, interaction-based models excel at capturing fine-grained matching details by analyzing different parts of queries and products at a low level before making a final decision based on aggregated evidence [10, 11, 12]. While these models outperform representation-based ones in many text matching scenarios, they face challenges in terms of online deployment due to their inability to pre-compute embeddings offline and consider context effectively.

Recent advancements like LLMs (e.g. BERT [13], Llamma [14], Mistral [15] and Gemma [16]) have revolutionized text matching tasks by combining the strengths of interaction-based and representation based models. Their multilayer architecture based on Transformer [17] allows for comprehensive interaction between queries and products at various semantic levels, addressing the shortcomings of previous models. Despite its effectiveness, LLM's computational intensity poses hurdles for practical online applications such as e-Commerce search engines.

In this work, our goal is to improve effectiveness of representation-based models used in production while still meeting strict latency requirements of e-Commerce search systems for tail queries segment. Toward this goal, we propose a novel knowledge distillation (KD) framework to distill an encoder-only LLM (i.e. BERT base [13]) into a representation-based student model (i.e. DistilBERT [18]), offering improved effectiveness of the student model while maintaining efficiency of the representation-based models. We firstly train a highly effective teacher model [1], followed by training the student network to mimic the LLM's behavior. To train the teacher model, we propose to use soft human labels converted from editorial feedback to make the model aware of differences between a perfect match item, an item with a mismatched attribute

[1] We use LLM and teacher model interchangeably

(e.g. brand, color, style etc) and completely irrelevant products, instead of simply using binarized labels [19] commonly adopted in literature. We show that using soft human labels improve effectiveness of our teacher model. We further incorporate items' attributes to our teacher model to enhance its performance. For our student model, we aim to mimic the margin between a relevant product d^+ and an irrelevant document d^- outputted by the teacher model. Intuitively, soft targets outputted by the LLM reduce noises and offer more informative knowledge about relevant differences between the two items [2]. The teacher model/LLM will be served offline while we can deploy the newly trained student model into production. The high level overview of our framework in shown in Figure 1. Our contributions are as follows:

- We propose a novel framework, consisting of a representation-based student model distilled from an LLM, to generate a semantic matching feature for a reranking system in a major e-Commerce search engine.
- We proposed to improve effectiveness of the teacher model by using soft human labels to distinguish a perfectly matched item from items with mismatched attributes and completely irrelevant items.
- We conducted extensive offline experiments on an in-house dataset and tested our framework with real-time production traffic. Online testing results showed significant gain of our model over an existing commercial production system.

2. Related work

In this section, we summarize related work about relevance search on e-Commerce, neural ranking models for text search and knowledge distillation methods.

2.1. Relevance e-Commerce Search and Ranking

The challenge of e-Commerce search surpasses that of traditional web search [20] owing to the shortness of user queries and the large number of potentially relevant items [21]. Researchers have suggested an iterative method involving multiple steps, starting with retrieving a set of candidate items, then iteratively reranking and reducing this set by selecting the top items [22]. In e-commerce, various signals are used to assess search result quality, with some studies [1, 2, 3, 23, 24] optimizing results based on user engagement metrics like click-through rate and conversion rate, best-selling products [25] and product result diversity [26]. However, sparseness of user engagement data may limit model performance on queries without engagement (e.g. tail queries). Recently, deep textual matching features based on deep neural-based models have been employed for retrieval and ranking, with enhancements such as incorporating different text representations and loss functions [27, 28, 29, 30, 31, 8, 32]. Additionally, some models have integrated interaction features between user queries and a product graph to capture relationships among similar products in the ranking process [33] and reinforcement learning for product search [34]. Our work develops a semantic matching feature based on our novel knowledge distillation framework, and is used among other engagement signals for reranking at a major e-Commerce search engine.

[2] We use items, products, documents interchangeably

2.2. Neural Ranking Models for Text Search

Neural ranking models for text search can be categorized into two groups: representation-based models and interaction-based methods. The former one seeks to learn representations of a query and a document, and measure their similarity [4, 5, 6, 7, 35, 36, 27], while the later one [37, 38, 39, 40, 41, 42, 43] aims to capture relevant matching signals between a query and a document based on word/tokens interactions. There are methods aiming to unified two categories within a single model such as Mitra et al. [44], Rao et al. [45]. Recent research has been centered around leveraging pretrained large language models, with BERT being a prominent example [13]. In the context of BERT-based relevance models, there are two common approaches in literature. The first one is about independently learning representations of queries and items/products using dual BERT encoders (e.g. siamese or two-tower structure) [8, 27, 28, 46, 9]. The second approach is to concatenate textual contents of a query-item pair and input the text into a BERT model [47, 48, 10, 11, 12, 49] which demonstrate state-of-the-art performance on various benchmarks. The former approach is known as representation-based learning method while the later one is an interaction-based approach. The e-commerce relevance task, akin to text matching, poses challenges for commercial search engines due to high traffic and low latency requirements. This makes deploying interaction-based LLMs online a significant hurdle. To address this issue, our work proposes distilling the interaction-based LLM (i.e. BERT base) into a representation-based architecture (i.e. DistilBERT), aiming to enhance ranking effectiveness while maintaining efficiency of online search systems.

2.3. Knowledge Distillation Methods

Online recommendation/search systems require strict latency in real time which hinders the deployment of LLMs (e.g BERT [50], LLamma [14], GPT [51]). Recently, researchers and practitioners utilize compression techniques to compress these models into smaller ones. One of the most widely used method is Knowledge Distillation [52]. It enables online systems to leverage sophisticated models like BERT effectively. The core concept of KD involves training a high-performance teacher model initially, followed by training a simpler student network to replicate the teacher's behavior. Knowledge distillation methods mainly fall into three groups: (1) response-based learning [53, 52, 48, 54, 55, 56, 57, 30], (2) representation-based methods [18, 58, 59, 60, 61] and (3) relation-based knowledge [62]. Our method can be viewed as a response-based one since our student model is optimized to learn from the soft targets generated by a large language model (LLM), which are more informative and less noisy. Our work is closest to [48, 30]. However, our teacher model is trained with products' attributes and soft ratings converted from editorial feedback to increase effectiveness.

3. Our Framework

Problem Formulation: Given a query q and an item d, where every item d has title and textual attributes such as product type, brand, color and gender, we aim to train a teacher model $t(q, d) \in \mathbb{R}$ and a student model $s(q, d) \in \mathbb{R}$. These two functions will determine relevancy of q and d. After training the LLM, we will train the student model by learning from soft-targets outputted by the LLM (i.e. knowledge distillation process). Our framework (Figure 1) consists of two main components: (1) the interaction-based LLM (i.e. BERT base) used as the teacher model, (2) the representation-based model (i.e. DistilBERT) which is the student model. Details of these components will be described in following subsections.

3.1. The teacher model

For each query-item pair (q, i), we utilize an LLM (i.e. BERT base) as encoder, and concatenate a query and title of an item as input to the BERT model. As the item title may not contain sufficient information to determine relevancy of the query and the item, we also concatenate the item's attributes (e.g. product type (PT), brand and so on) if they are available. The title and each of the attributes will have unique separator tokens as shown in Eq.1. The hidden state $\mathbf{E}_{(q,d)}([CLS])$ of $[CLS]$ token is taken as the query-item pair representation. To the best of our knowledge, our work is the first using items' attributes such as product types, brands, colors and genders to enhance effectiveness of an interaction-based LLM.

$$\mathbf{E}_{(q,d)} = BERT([CLS]\ q\ [SEP]\ [SEP_t]title[SEP_p]PT[SEP_b]brand) \tag{1}$$

To compute relevance score $t(q, d)$ of the teacher model, we input $\mathbf{E}_{(q,d)}([CLS])$ into MLP layers as follows:

$$t(q, d) = \mathbf{W}_2 \cdot layernorm(\mathbf{W}_1 \cdot \mathbf{E}_{(q,d)}([CLS])) \tag{2}$$

where $\mathbf{W}_1 \in \mathbb{R}^{768 \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times 1}$. We remove biases to avoid clutter. For each query-item pair (q, d), its rating can be Excellent (i.e. perfect match), Good (i.e. item with a mismatched attribute (e.g. brand, color, style etc)), Okay, Bad (i.e. irrelevant items) and so on. We can simply label excellent/good items as 1s and the rest as 0s similar to [19]. However, it is suboptimal since excellent items and good items are viewed as equal. To help our LLM distinguish these items, we propose to convert the editorial feedback into soft human labels by labelling an excellent item as 1, a good item as 0.5 and a completely irrelevant item as 0. The converted human labels are used in cross entropy loss to train our LLM as follows:

$$\mathcal{L}(t(q, d), y) = -y \cdot log(t(q, d)) - (1 - y) \cdot log(1 - t(q, d)) \tag{3}$$

where $y \in \{0, 1, 0.5\}$ converted from original editorial feedback.

3.2. The student model

As shown in Figure 1, our student model uses DistilBERT as encoder and has identical towers (Siamese network). For each query-item pair (q, i), we input the query to the DistillBERT as follows: $\mathbf{E}_q = DistilBERT([CLS]\ q\ [SEP])$ and use hidden state $\mathbf{E}_q([CLS])$ of the $[CLS]$ token as the query's representation. For the item, we concatenate its title and its available attributes, and input the concatenated text into DistilBERT as shown in 4. The hidden state $\mathbf{E}_d([CLS])$ of the $[CLS]$ token is used as the item's representation. The scoring function $t(q, d) = cosine\_sim(\mathbf{E}_q([CLS]), \mathbf{E}_d([CLS]))$.

$$\mathbf{E}_d = BERT([CLS]\ title\ [SEP_p]\ PT\ [SEP_b]\ brand) \tag{4}$$

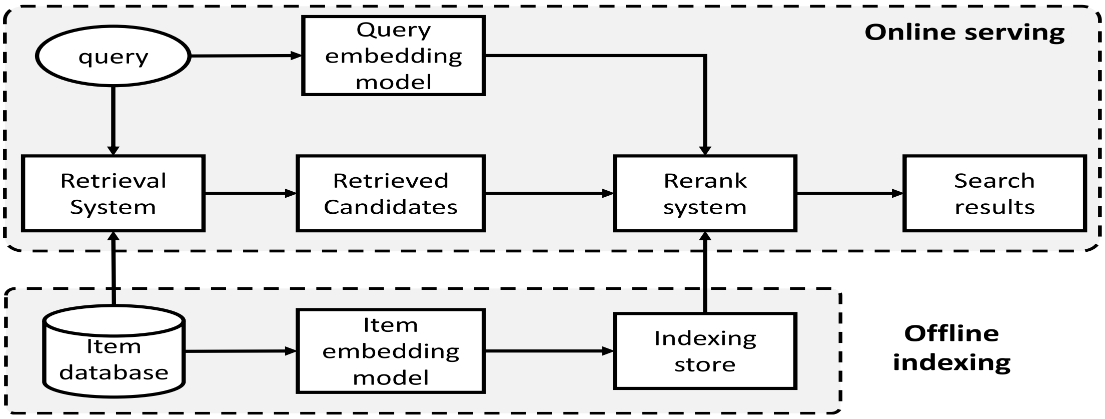

Figure 2: Overview of our online serving system

To train our student model, we use loss function similar as the margin MSE loss [48] to help the student model mimic the LLM's predicted margin. In [48] where simply binary labeling is used, triplets (q, d^+, d^-) are sampled where d^+ is relevant document and d^- is irrelevant document for the query q. The teacher's scores $t(q, d^+)$ and $t(q, d^-)$ viewed as soft targets, and the student scores $s(q, d^+)$ and $s(q, d^-)$ are computed. In [48], the margin MSE loss for a query q between a relevant document d^+ and a irrelevant document d^- is shown in Eq.5.

$$\mathscr{L}(q, d^+, d^-) = MSE(t(q, d^+) - t(q, d^-), s(q, d^+) - s(q, d^-)). \tag{5}$$

Extending Hofstätter et al. [48]'s work from binary classes to accommodate three distinct document classes, we use d^1, d^2, and d^3 to denote excellent document (label 1), good document (label 0.5) and irrelevant document (label 0) respectively for the query q. We sample triplets (q, d^i, d^j) where d^i is more relevant than document d^j for the query q, so there are three possible combinations (q, d^1, d^2), (q, d^1, d^3), and (q, d^2, d^3). We apply Eq.5 on the generated triplets to compute loss for the query q.

3.3. Online serving

After training our student model, we deploy it into production. The overview of our online serving system is in Figure 2. We index all products' embeddings with an offline pipeline. For every query q, we generate q's embedding online. From top-k retrieved candidates of a retrieval system, we compute a semantic matching feature based on the query's embedding and the retrieved items' embeddings. The feature will be used among other ranking features by a tree-based model to rank documents and return search results. The features used in a rerank system can be organized into three groups: (1) query features (e.g. query's attributes, length etc), (2) item features (e.g. item attributes, user reviews, ratings etc) and (3) query-item features (e.g. query-item engagement). Our semantic matching feature is a query-item feature.

4. Experiments

In this section, we discuss our strategies to collect data, performance of the teacher model and the student model, and our online tests.

4.1. Data Collection

To train text matching models, we can either use engagement information (e.g. click search logs) [43, 41] or human editorial feedback [8, 19]. While using engagement information to collect data may help us to generate large-scale data, we find that for tail queries, engagement information is usually limited and noisy, leading to poor effectiveness of our models. Therefore, we leverage human editorial labels, which may have smaller size but more reliable to capture textual relevancy between a query and an item, to train our models.

Over the years, our human editorial evaluation data is generated by manually assessing the top-ranked items for a set of sampled queries by a control ranking model and a variant model. The queries are sampled based on search traffic. Totally, we collected an in-house dataset where each query has a list of ~10-20 items with human editorial ratings similar to [43, 41, 8, 19]. Again, we did not use click-search logs to train our models in this paper. We convert the original ratings into soft human labels as discussed in Section 3.1. For each query-item pair (q, i), its rating can be Excellent (i.e. perfect match), Good (i.e. item with a mismatched attribute (e.g. brand, color, style etc)), Okay, Bad (i.e. irrelevant items) and so on. It should be noted that not all attributes hold equal importance. In this paper we will omit the specific details of our annotation guidelines. To further increase the number of query-item pairs, we have also included some hard negative items for each of the queries. While the addition of these hard negatives did not lead to significant relevance gains, we observed that including hard negatives resulted in the model yielding more consistent results than using random negative items.

4.2. Performance of the teacher model

We explored multiple methods to train our teacher models, with an emphasis on the labeling strategy and the loss function. Our current production model employs aggressive labeling where excellent items are labeled as positive 1, while all others are labeled as negative 0. Our analysis shows that subject mismatch accounts 20% of irrelevant search results, thus it is important to distinguish between good and irrelevant items for improving the relevance of top items. In Table 1, we compare the performance of our model trained with aggressive labeling to that trained with soft-labeling, where label is 1 for excellent match, 0.5 for good match, 0 for irrelevant match. We observe a relative gain of +0.47% in NDCG5 with the soft-labeling approach. Additionally, we explored other methods for distinguishing between good items and irrelevant items, including multi-class classification (**MCCE**) and Multivariate Ordinal Regression (**Ordinal**) [63], these approaches did not result in NDCG improvement. For knowledge distilling, using soft-labeling is also easier for knowledge distillation compared against **MCCE** and **Ordinal**. Soft-labeling approach generates a single logit output, simplifying the knowledge distillation process compared to the two-output approach of **MCCE** and **Ordinal**. Based on above, we adopt the soft-labeling method as our teacher model.

Teacher Model	NDCG@5	NDCG@10
BERT w/ aggressive labeling	0%	0%
BERT w/ soft-labeling w/o item attributes	+0.32%	+0.23%
BERT w/ soft-labeling	+0.47%	+0.39%

Table 1
Offline results of interaction-based teacher models

Model	NDCG@5 Lift	NDCG@10 Lift
DistilBERT w/o KD	0%	0%
Softmax CE loss [52]	+0.70%	+0.66%
Multi-Margin MSE [54]	+1.29%	+0.73%
KD-DistilBERT (our student model)	+1.81%	+1.46%
BERT with soft-labeling (our teacher model)	+3.98%	+2.85%

Table 2
Offline results of our KD-DistilBERT (the student model), our teacher model and baselines

We also conducted experiments both including and excluding item attributes in the model input. The results indicate that including item attributes improves the NDCG metrics (see Table 1).

4.3. Performance of the student model

We compare our student model (KD-DistilBERT) trained with margin MSE loss with state-of-the-art KD response-based method. We also include performance of our best teacher model. As shown in Table 2, all KD-based methods outperform distilBERT training without knowledge distillation significantly with p-value < 0.001 by using t-test, indicating the effectiveness of using soft-targets outputted by our teacher model. Our model (KD-DistilBERT) performs best among KD-based methods. We can see the teacher model outperforms all student models with large gaps. Note that, all student models have the same model architecture (DistilBERT) for fair comparisons. As the gap between our student model and our teacher model is considerable, we may consider using a bigger model as a student model to further improve effectiveness while latency increases modestly. We leave it for future work.

In terms of latency, we observe that the teacher model is much slower than our student model. In runtime, given a query (q, d), the teacher model needs to make inference for a concatenation of the query and the item, while for the student model, we can compute the item's embedding offline and as the content of the query is short, online inference for the query's representation is fast. Therefore, the student model is much more preferable for online applications. As our student model has same architecture with the existing production model, our student model does not incur any additional latency.

4.4. Online experiments

Our KD-DistilBERT performance was assessed by human evaluators who compared the top-10 results from our model with Walmart's production system which already has a semantic

Method	NDCG@5 Lift (P-value)	NDCG@10 Lift (P-value)
KD-DistilBERT	+1.30% (0.001)	+0.98% (0.00)

Table 3
Human evaluation on the top-10 ranking items on a sample of queries. KD-DistilBERT outperforms production system with statistical significance level p-value<0.01 by t-test.

Method	First-time buyer Lift (p-value)	Session Abandonment Rate Lift (p-value)	Sessions with Item Click Lift (p-value)
KD-DistilBERT	+2.55% (0.07)	-0.25% (0.00)	+0.214% (0.00)

Table 4
AB test results

matching feature by using siamese DistilBERT model [8]. As we use DistilBERT as encoder, our framework does not incur any additional latency. Queries were randomly sampled from search traffic at Walmart. As we can see in Table 3, our model outperforms the production system significantly on relevancy metrics (NDCG@5 and NDCG@10). Reported results were stastistically significance t-test. We also conducted A/B test to compare engagement metrics of our proposed framework and the production system. As reported in Table 4, our model increases first-time buyer by 2.55%, reduces abandonment search sessions by 0.25% and increase the number of sessions with click by 0.214%.

5. Discussion and Future Work

In this paper, we employ an encoder-only LLM (i.e. BERT) as the teacher model. We found that powerful decoder-only LLMs with more number of parameters (e.g. Llama [14], Mistral [15]) are more effective and can further improve effectiveness of the student model. We leave it for further work.

Currently, our student model is only trained on soft-targets outputted from the teacher model for query-item pairs in human editorial feedback dataset. It is suboptimal since we can apply the teacher model on unlabeled dataset to have a much larger dataset. Our preliminary results show that it is beneficial to generate soft-labels for unlabeled query-item pairs. We will further explore this direction in the future work. In addition to that, as a next step, we will explore the possibility of incorporating a multi-objective loss function that combines both relevance and engagement information.

As our model is served for tail queries only, we will expand it for head/torso segment and further include users' information to make search results more personalized. We leave it for further work as well.

6. Conclusion

In this paper, we proposed a novel knowledge distillation framework consisting of an LLM as the teacher model and a DistilBERT as the student model. We proposed to improve the effectiveness of LLM by using soft human labels and items' attributes. Our KD-DistilBERT outperformed baselines in offline and online experiments while maintaining efficiency of the existing production system. Our work opens the door for new industrial applications of other LLMs [15, 14, 16] in e-Commerce search.

References

[1] K. Bi, C. H. Teo, Y. Dattatreya, V. Mohan, W. B. Croft, Leverage implicit feedback for context-aware product search, arXiv preprint arXiv:1909.02065 (2019).

[2] X. Wu, A. Magnani, S. Chaidaroon, A. Puthenputhussery, C. Liao, Y. Fang, A multi-task learning framework for product ranking with Bert, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 493–501.

[3] S. K. Karmaker Santu, P. Sondhi, C. Zhai, On application of learning to rank for e-commerce search, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 475–484.

[4] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, L. Heck, Learning deep structured semantic models for web search using clickthrough data, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013, pp. 2333–2338.

[5] Y. Shen, X. He, J. Gao, L. Deng, G. Mesnil, 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, 2014, pp. 101–110.

[6] A. Severyn, A. Moschitti, 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, 2015, pp. 373–382.

[7] Y. Nie, H. Chen, M. Bansal, Combining fact extraction and verification with neural semantic matching networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6859–6866.

[8] A. Magnani, F. Liu, S. Chaidaroon, S. Yadav, P. Reddy Suram, A. Puthenputhussery, S. Chen, M. Xie, A. Kashi, T. Lee, et al., Semantic retrieval at walmart, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 3495–3503.

[9] H. Shan, Q. Zhang, Z. Liu, G. Zhang, C. Li, Beyond two-tower: Attribute guided representation learning for candidate retrieval, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 3173–3181.

[10] Z. Dai, J. Callan, Deeper text understanding for ir with contextual neural language modeling, in: Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 985–988.

[11] N. Yadav, N. Monath, M. Zaheer, A. Mccallum, Efficient k-nn search with cross-encoders using adaptive multi-round cur decomposition, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 8088–8103.

[12] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over Bert, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 39–48.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional

transformers for language understanding, in: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2018.

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[15] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[16] J. Banks, T. Warkentin, Gemma: Introducing new state-of-the-art open models, 2024. URL: https://blog.google/technology/developers/gemma-open-models/.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[18] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of Bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[19] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, J. Lin, Ms marco: Benchmarking ranking models in the large-data regime, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1566–1576.

[20] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 129–136.

[21] F. Sarvi, N. Voskarides, L. Mooiman, S. Schelter, M. de Rijke, A comparison of supervised learning to match methods for product search, arXiv preprint arXiv:2007.10296 (2020).

[22] A. Trotman, J. Degenhardt, S. Kallumadi, The architecture of ebay search., in: eCOM@ SIGIR, 2017.

[23] P. Li, R. Li, Q. Da, A.-X. Zeng, L. Zhang, Improving multi-scenario learning to rank in e-commerce by exploiting task relationships in the label space, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 2605–2612.

[24] S. Yao, J. Tan, X. Chen, K. Yang, R. Xiao, H. Deng, X. Wan, Learning a product relevance model from click-through data in e-commerce, in: Proceedings of the Web Conference 2021, 2021, pp. 2890–2899.

[25] B. Long, J. Bian, A. Dong, Y. Chang, Enhancing product search by best-selling prediction in e-commerce, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 2479–2482.

[26] N. Parikh, N. Sundaresan, Beyond relevance in marketplace search, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 2109–2112.

[27] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, O. Frieder, Efficient document re-ranking for transformers by precomputing term representations, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 49–58.

[28] J.-T. Huang, A. Sharma, S. Sun, L. Xia, D. Zhang, P. Pronin, J. Padmanabhan, G. Ottaviano, L. Yang, Embedding-based retrieval in facebook search, in: Proceedings of the 26th ACM

SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2553–2561.

[29] W. Kong, S. Khadanga, C. Li, S. K. Gupta, M. Zhang, W. Xu, M. Bendersky, Multi-aspect dense retrieval, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 3178–3186.

[30] L. Kumar, S. Sarkar, Listbert: Learning to rank e-commerce products with listwise Bert, arXiv preprint arXiv:2206.15198 (2022).

[31] P. Pobrotyn, T. Bartczak, M. Synowiec, R. Białobrzeski, J. Bojar, Context-aware learning to rank with self-attention, arXiv preprint arXiv:2005.10084 (2020).

[32] E. P. Brenner, J. Zhao, A. Kutiyanawala, Z. Yan, End-to-end neural ranking for ecommerce product search, Proceedings of SIGIR eCom 18 (2018) 7.

[33] Y. Zhang, D. Wang, Y. Zhang, Neural ir meets graph embedding: A ranking model for product search, arXiv preprint arXiv:1901.08286 (2019).

[34] Y. Hu, Q. Da, A. Zeng, Y. Yu, Y. Xu, Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application, in: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 368–377.

[35] M. Zhu, A. Ahuja, W. Wei, C. K. Reddy, A hierarchical attention retrieval model for healthcare question answering, in: The World Wide Web Conference, 2019, pp. 2472–2482.

[36] L. Gao, J. Callan, Condenser: a pre-training architecture for dense retrieval, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 981–993.

[37] H. Chen, F. X. Han, D. Niu, D. Liu, K. Lai, C. Wu, Y. Xu, Mix: Multi-channel information crossing for text matching, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 110–119.

[38] K. Hui, A. Yates, K. Berberich, G. de Melo, Pacrr: A position-aware neural ir model for relevance matching, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1049–1058.

[39] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, X. Cheng, Text matching as image recognition, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[40] J. Guo, Y. Fan, Q. Ai, W. B. Croft, A deep relevance matching model for ad-hoc retrieval, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 2016, pp. 55–64.

[41] C. Xiong, Z. Dai, J. Callan, Z. Liu, R. Power, End-to-end neural ad-hoc ranking with kernel pooling, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 55–64.

[42] K. Hui, A. Yates, K. Berberich, G. De Melo, Co-pacrr: A context-aware neural ir model for ad-hoc retrieval, in: Proceedings of the eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 279–287.

[43] Z. Dai, C. Xiong, J. Callan, Z. Liu, Convolutional neural networks for soft-matching n-grams in ad-hoc search, in: Proceedings of the eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 126–134.

[44] B. Mitra, F. Diaz, N. Craswell, Learning to match using local and distributed representations of text for web search, in: Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 1291–1299.

[45] J. Rao, L. Liu, Y. Tay, W. Yang, P. Shi, J. Lin, Bridging the gap between relevance matching and semantic matching for short text similarity modeling, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5373–5384.

[46] J. Ni, C. Qu, J. Lu, Z. Dai, G. H. Abrego, J. Ma, V. Zhao, Y. Luan, K. Hall, M.-W. Chang, et al., Large dual encoders are generalizable retrievers, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9844–9855.

[47] R. Nogueira, W. Yang, K. Cho, J. Lin, Multi-stage document ranking with Bert, arXiv preprint arXiv:1910.14424 (2019).

[48] S. Hofstätter, S. Althammer, M. Schröder, M. Sertkan, A. Hanbury, Improving efficient neural ranking models with cross-architecture knowledge distillation, arXiv preprint arXiv:2010.02666 (2020).

[49] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, M. Zaharia, Colbertv2: Effective and efficient retrieval via lightweight late interaction, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 3715–3734.

[50] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, 2019, p. 2.

[51] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[52] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[53] W. Sun, L. Yan, X. Ma, P. Ren, D. Yin, Z. Ren, Is chatgpt good at search? investigating large language models as re-ranking agent, arXiv preprint arXiv:2304.09542 (2023).

[54] A. Menon, S. Jayasumana, A. S. Rawat, S. Kim, S. Reddi, S. Kumar, In defense of dual-encoders for neural ranking, in: International Conference on Machine Learning, PMLR, 2022, pp. 15376–15400.

[55] Y. Jiang, Y. Shang, Z. Liu, H. Shen, Y. Xiao, S. Xu, W. Xiong, W. Yan, D. Jin, Bert2dnn: Bert distillation with massive unlabeled data for online e-commerce search, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 212–221.

[56] A. Muhamed, I. Keivanloo, S. Perera, J. Mracek, Y. Xu, Q. Cui, S. Rajagopalan, B. Zeng, T. Chilimbi, Ctr-bert: Cost-effective knowledge distillation for billion-parameter teacher models, in: NeurIPS Efficient Natural Language and Speech Processing Workshop, 2021.

[57] Z. Liu, C. Wang, H. Feng, L. Wu, L. Yang, Knowledge distillation based contextual relevance matching for e-commerce product search, arXiv preprint arXiv:2210.01701 (2022).

[58] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, Q. Liu, Tinybert: Distilling Bert for natural language understanding, arXiv preprint arXiv:1909.10351 (2019).

[59] S. Yao, J. Tan, X. Chen, J. Zhang, X. Zeng, K. Yang, Reprbert: Distilling bert to an efficient representation-based relevance model for e-commerce, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4363–4371.

[60] S. Liu, L. Li, J. Song, Y. Yang, X. Zeng, Multimodal pre-training with self-distillation for product understanding in e-commerce, in: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, 2023, pp. 1039–1047.

[61] K. Howell, J. Wang, A. Hazare, J. Bradley, C. Brew, X. Chen, M. Dunn, B. A. Hockey, A. Maurer, D. Widdows, Domain-specific knowledge distillation yields smaller and better models for conversational commerce, in: Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5), 2022, pp. 151–160.

[62] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, Y. Duan, Knowledge distillation via instance relationship graph, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7096–7104.

[63] L. Yan, Z. Qin, X. Wang, G. Shamir, M. Bendersky, Learning to rank when grades matter, arXiv preprint arXiv:2306.08650 (2023).