Pre-training with Aspect-Content Text Mutual Prediction for Multi-Aspect Dense Retrieval

Xiaojie Sun
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
sunxiaojie21s@ict.ac.cn

Keping Bi
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
bikeping@ict.ac.cn

Jiafeng Guo
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
guojiafeng@ict.ac.cn

Xinyu Ma
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
xinyuma2016@gmail.com

Yixing Fan
CAS Key Lab of Network Data Science and Technology, ICT, CAS
University of Chinese Academy of Sciences
Beijing, China
fanyixing@ict.ac.cn

Hongyu Shan
Qishen Zhang
Zhongyi Liu
Ant Group
Beijing, China
{xinzong,qishen.zqs,zhongyi.lzy}@alibaba-inc.com

ABSTRACT

Grounded on pre-trained language models (PLMs), dense retrieval has been studied extensively on plain text. In contrast, there has been little research on retrieving data with multiple aspects using dense models. In the scenarios such as product search, the aspect information plays an essential role in relevance matching, e.g., category: *Electronics*, *Computers*, and *Pet Supplies*. A common way of leveraging aspect information for multi-aspect retrieval is to introduce an auxiliary classification objective, i.e., using item contents to predict the annotated value IDs of item aspects. However, by learning the value embeddings from scratch, this approach may not capture the various semantic similarities between the values sufficiently. To address this limitation, we leverage the aspect information as text strings rather than class IDs during pre-training so that their semantic similarities can be naturally captured in the PLMs. To facilitate effective retrieval with the aspect strings, we propose mutual prediction objectives between the text of the item aspect and content. In this way, our model makes more sufficient use of aspect information than conducting undifferentiated masked language modeling (MLM) on the concatenated text of aspects and content. Extensive experiments on two real-world datasets (product and mini-program search) show that our approach can outperform competitive baselines both treating aspect values as classes and conducting the same MLM for aspect and content strings. Code and related dataset will be available at the URL [1].

[1]https://github.com/sunxiaojie99/ATTEMPT

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615157

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Dense Retrieval, Multi-Aspect, Pre-training

ACM Reference Format:
Xiaojie Sun, Keping Bi, Jiafeng Guo, Xinyu Ma, Yixing Fan, Hongyu Shan, Qishen Zhang, and Zhongyi Liu. 2023. Pre-training with Aspect-Content Text Mutual Prediction for Multi-Aspect Dense Retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615157

1 INTRODUCTION

Dense retrieval models [9–11, 26, 28, 29] have achieved compelling performance with pre-trained language models (PLMs) [5, 24] as the backbone. Most studies on dense retrieval focus on unstructured data consisting of plain text, while little attention has been paid to structured item retrieval such as product and people search. In these scenarios, additional aspect information beyond the query or item content is critical for relevance matching, such as brand-*nike*, affiliation-*Stanford*. However, little work has explored how to use them effectively in dense retrieval models.

A typical way of leveraging aspect information for multi-aspect retrieval is to refine the item representations with an auxiliary aspect prediction objective [12]. Specifically, for each aspect of an item, the item content is used to predict its annotated value IDs during training. This approach has two major disadvantages: 1) It considers the values of an aspect as isolated classes and learns the embeddings of value IDs from scratch, ignoring their semantic relations. For example, among the category values, "Hunting & Fishing" is more related to "Sports & Outdoors" while unrelated to "Pet Supplies". However, such semantic relations may not be captured sufficiently if we treat them as independent classes. 2) It does not use query/item aspects such as category, brand, color,

etc. during test time, which limits the potential retrieval gains. Although it may be costly to obtain query aspects during online service, item aspects can be extracted offline and it is easy to also use them during inference if they are already used in training.

In this paper, we propose a method of pre-training with Aspect-contenT TExt Mutual PredicTion (ATTEMPT) to address the above limitations. Specifically, ATTEMPT leverages aspect values as text strings and concatenates them with the content using leading indicator tokens in between. For more effective retrieval, rather than simply conducting undifferentiated MLM on the concatenated aspect and content text, we specifically design an aspect-content mutual prediction objective. It keeps the entire aspect/content tokens and predicts the masked ones in the content/aspects. Also, to suit the scenario where the overhead of obtaining the query aspects online is high, we set the query aspect text to empty during inference. Our method has several advantages over the common approach: 1) In ATTEMPT, the text of an aspect value reuses the token embeddings from the powerful PLMs so the semantic relations between values can be naturally captured. 2) Being concatenated with the content, the item aspects can also take effect for relevance matching during test time. 3) The aspect-content mutual prediction objective promotes sufficient interactions between the aspect and content at the token level, producing better item representations for retrieval, which is confirmed by extensive experimental results.

As far as we know, there are no suitable large-scale public datasets for multi-aspect retrieval. We construct such a dataset by crawling the item categories from their pages to complement the aspects in the Amazon ESCI dataset [19]. Our experiments on this refined dataset and a real-world commercial mini-program dataset show that ATTEMPT can significantly outperform the competitive baselines both predicting the classes of aspect values and conducting the same MLM for aspect and content strings.

2 RELATED WORK

There are three threads of work related to our study. (1) **Multi-aspect Retrieval.** Some work has exploited multi-aspect information to rank products or entities before PLMs appear [1, 2, 21]. In the era of PLM [6], there has been limited research on multi-aspect retrieval until Kong et al. [12] first attempts to do so. They learn aspect embeddings by predicting their value IDs with item contents and fuse them to yield an item embedding. Later, Shan et al. [23] proposed a fine-tuning method that uses the local aspect-level matching signals to enhance the global query-item embedding matching. (2) **Multi-field Retrieval.** How to effectively leverage multiple fields (e.g., title, body, etc.) in a document has been a long-standing research topic. The most famous method is BM25F [22]. Methods leveraging multi-fields have also been proposed before and after PLMs appeared [3, 18, 27, 30]. The multi-fields are unstructured text in nature and the essential issue is how to weigh them differently during matching. Aspects, unlike fields, usually have a fixed value set which is much smaller than the space of field text. Thus, their core challenges are different. (3) **Pre-trained Models for Dense Retrieval.** Many studies have explored promoting the capabilities of PLMs for dense retrieval including introducing extra training objectives[4, 13, 15–17], special masking schemes [25], and model architecture changes [7], etc. Our method is grounded on the basic dual BERT encoders [5].

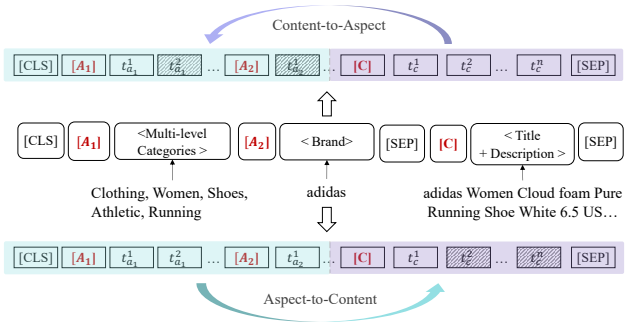

Figure 1: The mutual prediction MLM in ATTEMPT. The aspect and content texts are colored green and purple.

3 METHODOLOGY

3.1 Preliminary

For a query q or a candidate item i , we represent the content text (e.g., query string, title, description) as t_c and the aspect text (e.g., values for brand, color, and category) as t_a. Assuming q or i has k aspects, t_a is further denoted as $t_{a_1}, ..., t_{a_k}$. For each aspect a_j $(1 \le j \le k)$, it has a finite vocabulary of aspect values, denoted as V_{a_j}. Previous work [12] incorporates aspect information by predicting the IDs corresponding to the annotated values of each aspect a_j within the space V_{a_j}. In contrast, we propose to pre-train the encoder by conducting mutual prediction between text t_a and t_c.

3.2 ATTEMPT

To model the semantic relationship between various values of an aspect naturally, we treat the aspect values as text strings and concatenate them with the content text. For sufficient capture of the interactions between item aspects and contents, we introduce mutual prediction objectives as illustrated in Figure 1.

Encoder Input. To indicate different types of text segments, we prepend an indicator token $[A_j]$ $(1 \le j \le k)$ and $[C]$ to the aspect text t_{a_j} and the original content t_c, e.g., an encoder input is $[A_1]t_{a_1}[A_2]t_{a_2}[A_3]t_{a_3}[SEP][C]t_c[SEP]$. When a query/item does not have certain aspect information, the corresponding aspect text will be empty. In this case, the indicator tokens could still learn some implicit representations of the query/item content. Note that during relevance matching, we always keep the query aspect text empty to suit the practical retrieval scenarios where the overhead of obtaining query aspects is high and also avoid potential semantic drift. Table 3 will show that the query-side indicator tokens ($[A_j]$ $(1 \le j \le k)$, $[C]$) alone learned during pre-training are beneficial for retrieval. Since the other parts of ATTEMPT are exactly the same between q and i, we take i as an example for illustration.

Content Masked Language Modeling (MLM). To capture the interactions between the content tokens without any auxiliary information, ATTEMPT conducts MLM on the item content. It randomly masks tokens in the content text and predicts the masked tokens with the context-dependent representations encoded by Transformer layers [5]. The corresponding loss function is:

$$\mathcal{L}_{MLM}(\hat{\boldsymbol{t}_c}) = -\sum_{w \in m(\hat{\boldsymbol{t}_c})} logP(w|\hat{\boldsymbol{t}_c}_{\setminus m(\hat{\boldsymbol{t}_c})}), \quad (1)$$

where $\hat{\boldsymbol{t}_c}$ denotes the text produced by randomly masking some tokens in the text $\boldsymbol{t_c}$, $m(\hat{\boldsymbol{t}_c})$ denotes the masked tokens, and $\hat{\boldsymbol{t}_c}_{\setminus m(\hat{\boldsymbol{t}_c})}$ denotes the remaining tokens in $\hat{\boldsymbol{t}_c}$.

Aspect-to-Content MLM Prediction. We take the entire aspect text as context when predicting the masked tokens in the content text. Under this context, the prediction of masked content tokens has extra evidence for consideration and can act differently than content MLM alone. The aspect-to-content (a2c) loss $\mathcal{L}_{a2c}$ is:

$$\mathcal{L}_{a2c}(\boldsymbol{t_a} \oplus \hat{\boldsymbol{t_c}}) = -\sum_{w \in m(\hat{t_c})} logP(w|\boldsymbol{t_a} \oplus \hat{\boldsymbol{t_c}}_{\setminus m(\hat{t_c})}), \quad (2)$$

where $\oplus$ means concatenation. In particular, the leading tokens $[A_j]$ $(1 \le j \le k)$ and $[C]$ in the input will not be masked.

Content-to-Aspect MLM Prediction. The idea of content-to-aspect prediction is similar to the aspect classification in [12], both of which use the original content to predict the aspects. However, ATTEMPT predicts the masked words in the aspect text rather than the value classes (IDs), which encodes the aspect information in a softer manner. Specifically, the content-to-aspect (c2a) loss is:

$$\mathcal{L}_{c2a}(\hat{\boldsymbol{t_a}} \oplus \boldsymbol{t_c}) = -\sum_{w \in m(\hat{t_a})} logP(w|\hat{\boldsymbol{t_a}}_{\setminus m(\hat{t_a})} \oplus \boldsymbol{t_c}). \quad (3)$$

Overall Learning Objective. By introducing $\mathcal{L}_{a2c}$ and $\mathcal{L}_{c2a}$, ATTEMPT can incorporate the aspect information into the item representation sufficiently through bidirectional interactions. In summary, our overall pre-training objective is:

$$\mathcal{L}_{overall} = \mathcal{L}_{MLM}(\hat{\boldsymbol{t_c}}) + \lambda(\mathcal{L}_{a2c}(\boldsymbol{t_a} \oplus \hat{\boldsymbol{t_c}}) + \mathcal{L}_{c2a}(\hat{\boldsymbol{t_a}} \oplus \boldsymbol{t_c})), \quad (4)$$

where λ is a hyper-parameter.

4 EXPERIMENTAL SETUP

4.1 Datasets

We conduct model comparisons on two real-world datasets:

Multi-Aspect Amazon ESCI Dataset (MA-Amazon). Amazon ESCI Product Search [19] originally has multilingual real-world queries, product information such as brand, color, title, description, etc., and 4-level relevance labels: *Exact*, *Substitute*, *Complement*, and *Irrelevant*. We only use the English part and enrich the dataset by collecting multi-level product categories from the item pages. We merge all the items and get a corpus of 482K unique items, which is used for pre-training. For fine-tuning, we divide the original training set into training and validation sets by queries, and keep the test set, yielding 17K, 3.5K, and 8.9K queries respectively. As in [19], we treat *Exact* as relevant and the other labels as irrelevant during training and for recall calculation. MA-Amazon only has item aspect information, and the coverage of brand, color, and category of levels 1-2-3-4 are 94%, 67%, and 87%-87%-85%-71%, respectively.

Alipay Search Dataset. Alipay is a mini-program (app-like service) search dataset with binary manual relevance annotations. The pre-training query/item corpus has 1.3M/1.8M distinct queries/items with aspect information i.e., *brand* (44%/0.6% coverage on query/item) and *three-level categories* (91%-90%-56%/90%-90%-62% coverage for category 1-2-3 of query/item). The fine-tuning dataset consists of 60K/3.3K/3.3K unique queries in the training/validation/test set. Note that the queries for validation and testing do not appear in the pre-training query corpus.

4.2 Baselines

We compare ATTEMPT with the following pre-training methods (**-C** means that the input takes the same concatenation strategy for aspect and content text as ATTEMPT): (1) **BIBERT** [14, 20]: BIBERT, the backbone of ATTEMPT, is a prevalent dense retrieval method for plain text. It employs MLM [5] to pre-train the encoder using the content text of query/item. (2) **Condenser** [7]: It adds a short circuit between the tokens except CLS of the lower layer and the higher layer of BERT [5] to enhance the final CLS representation. (3) **BIBERT-C**: It only differs from BIBERT in the encoder input. It uses the aspect text in the same way as ATTEMPT during pre-training and fine-tuning. (4) **BIBERT-C(A)**: It refines BIBERT-C by assigning a higher mask ratio specifically for the aspect text, which is consistent with ATTEMPT. (5) **MTBERT [-C]** [12]: It conducts k additional aspect classification tasks on the CLS during pre-training. (6) **MADRAL [-C]** [12]: It initiates extra multiple aspect embeddings and learns them by predicting the value classes of each aspect and fuses them to produce the final item representation.

Table 1: Overall performance. The best results are in bold. † indicates significant differences between ATTEMPT and the best baselines in the first/second/third group.

Method	MA-Amazon			Alipay		
	r@100	r@500	ndcg@50	r@100	r@500	ndcg@50
BIBERT	0.6075	0.7795	0.3929	0.4464	0.6284	0.2033
Condenser	0.6091†	0.7801†	0.3960†	0.4520†	0.6423†	0.2072†
MTBERT	0.6137†	0.7852†	0.3969†	0.4498	0.6280	0.2064
MADRAL	0.6088	0.7815	0.3950	0.4506†	0.6383†	0.2057†
BIBERT-C	0.6137	0.7814	0.4005	0.4517	0.6291	0.2103
BIBERT-C(A)	0.6137	0.7841	0.4019	0.4611	0.6432†	0.2091
MTBERT-C	0.6142	0.7839	0.3997	0.4391	0.6189	0.2026
MADRAL-C	0.6169†	0.7850†	0.4041†	0.4376	0.6141	0.2044
ATTEMPT	**0.6233**	**0.7924**	**0.4097**	**0.4667**	**0.6592**	**0.2113**

4.3 Implementation and Evaluation Details

We implemented ATTEMPT and all the baselines by ourselves. For all the methods, the encoder is shared for both queries and items.

Pre-training. The maximum token length is 156, The learning rate and epoch for the MA-Amazon/Alipay dataset are set to 1e-4/5e-5 and 20/10, respectively. We initialize the BERT parameters with Google's public checkpoint and use Adam optimizer with a linear warm-up. For all -C baselines and ATTEMPT, the mask ratios are set to 0.15/0.3 for item/query content to account for the shorter query length. They all have the same mask ratio between aspect and content text except for BIBERT-C(A) and ATTEMPT, where the mask ratio for aspect text is 0.6. λ in Eq.4 is set as 1.0. We fine-tune the pre-trained model checkpoints every two epochs and select the best one on the validation dataset.

Fine-tuning. On both datasets, all models are trained for 20 epochs with the Tevatron toolkit[8]. We use a learning rate of 5e-6 and a batch size of 64. All methods are trained with softmax cross entropy loss with in-batch negatives and one hard negative. Note that we have not used auxiliary classification objective for MTBERT and MADRAL since no significant improvements are achieved.

Metrics. We report recall@100, recall@500 and ndcg@50. When calculating ndcg on MA-Amazon, following [19], we set the gains of E, S, C, and I to 1.0, 0.1, 0.01, and 0.0, respectively. We perform two-tailed t-tests (p-value ≤ 0.05) to see significant differences.

5 EXPERIMENTAL RESULTS

5.1 Main Results

The overall performance is shown in Table 1. We have the following observations: (1) Generally, methods using aspect information outperform those that don't, confirming the importance of aspects in

Table 2: Study of various component choices on MA-Amazon. † indicates significant improvements over BIBERT.

	r@100	r@500	ndcg@50
BIBERT	0.6075	0.7795	0.3929
ATTEMPT	**0.6233**†	**0.7924**†	**0.4097**†
only brand	0.5977	0.7710	0.3859
only color	0.5867	0.7626	0.3773
only cate1-4	0.6212†	0.7893†	0.4050†
brand+color+cate1	0.6192†	0.7863†	0.4040†
brand+color+cate1-2	0.6199†	0.7898†	0.4073†
brand+color+cate1-3	0.6223†	0.7910†	0.4092†
ATTEMPT $^{-\mathcal{L}_{c2a}}$	0.6211†	0.7882†	0.4068†
ATTEMPT $^{-\mathcal{L}_{a2c}}$	0.6127†	0.7846†	0.3997†
ATTEMPT $^{-\mathcal{L}_{mlm}}$	0.6145†	0.7851†	0.4013†
BIBERT^{+AGREE}	0.6246†	0.7913†	0.4112†
ATTEMPT $^{+AGREE}$	**0.6393**†	**0.8019**†	**0.4257**†

relevance matching. Notably, MADRAL performs worse than MT-BERT on MA-Amazon, possibly due to insufficient pre-training data to learn the aspect embeddings from scratch sufficiently. (2) Models treating aspect information as text strings (BIBERT-C/-C(A)) surpass those considering aspect values as discrete classes (MTBERT and MADRAL). When the aspect text has a larger mask ratio than the content (in BIBERT-C(A)), the retrieval performance will be boosted. This shows that the aspect text should be taken special care to encourage sufficient learning. (3) When aspect text concatenation is incorporated with methods using aspect values for classification (MTBERT and MADRAL), the retrieval performance does not always become better. This could be because the input aspect text becomes the shortcut for the models to predict its corresponding class ID. When the pre-training data is large (e.g., on Alipay), such relation is more likely grasped by models, deterring the learning of beneficial interactions. (4) More powerful pre-training method (Condenser) sometimes perform better than methods using aspects (MTBERT and MADRAL on Alipay). Note that the benefit from the advanced pre-training techniques is orthogonal to the aspect information and they can be combined for even better performance. We leave the study of this in future work. (5) Overall, our ATTEMPT achieves the best performance on both datasets, showing the efficacy of its pre-training objective specifically proposed for the concatenated text of aspect and content.

5.2 Further Analysis

We also probe ATTEMPT from various perspectives to verify its effectiveness. For reproducibility, our analysis is based on MA-Amazon. The only exception is the ablation study of query/item aspects since only Alipay has both of them.

Ablation Study of Aspects. We study the effects of various aspects in ATTEMPT (brand, color, and category from level 1 to 4) in Table 2. We find that: (1) When use each aspect alone, only the category information enhances model performance. This might because brand and color are often included in the item content already while the category is extra meta information. The observation that category matters the most is consistent with [12]. (2) Combining all aspects outperforms using category only, indicating that brand and color may take a better effect when interacting with the category. (3) More levels of category information will lead to better performance except that three and four levels have similar results. While adding more category levels provides richer information, the reduced coverage (refer to Section 4.1) might limit the benefits.

Table 3: Ablation study of query/item aspects on Alipay. † indicates significant improvements over BIBERT.

	r@100	r@500	ndcg@50
BIBERT	0.4464	0.6284	0.2033
ATTEMPT	**0.4667**†	**0.6592**†	0.2113†
ATTEMPT $^{only\ d}$	0.4563†	0.6437†	0.2105†
ATTEMPT $^{only\ q}$	0.4526	0.6366†	0.2059

Ablation Study of Loss Function. We remove each of the three losses from the overall loss to see how important it is. In Table 2, we find that: (1) The bidirectional prediction losses are beneficial to ATTEMPT, and excluding either leads to a performance drop. (2) The Aspect-to-Content(a2c) prediction is the most helpful, indicating that using aspects as context for content MLM prediction is a feasible way to infuse the aspect information into an item. (3) The performance also drops a lot when the vanilla MLM loss is eliminated, indicating the original content semantics without being affected by external information are also important.

Combination with Advanced Fine-tuning Techniques. AGREE [23] is a recently proposed fine-tuning method that incorporates a local aspect-query matching loss with the original global query-item matching loss. AGREE has not studied how to utilize query aspects, which suits MA-Amazon well since it does not have query aspects. Since AGREE concatenates the item aspects with content, it is easy to integrate AGREE during fine-tuning after pre-training with ATTEMPT. The last block in Table 2 shows the performance of AGREE alone and combining both. It shows that based on better fine-tuning techniques, ATTEMPT can achieve better performance. Notably, combining AGREE with methods that conduct aspect classification will not necessarily lead to better performance (Check MTBERT-C and MADRAL-C in Table 1).

Ablation Study of Query/Item Aspects. We examine the influence of the query and item aspects in Table 3. It shows that both query aspects and item aspects contribute to retrieval performance and the item aspects are more important. Since we only use item aspects during relevance matching, query aspects only take effect during pre-training and could have fewer contributions.

6 CONCLUSION

In this paper, we propose an effective pre-training method that uses aspects as text strings and conducts mutual prediction between the aspect and content text for multi-aspect retrieval. In contrast to previous approaches that treat aspect values as categorical IDs, ATTEMPT can capture the semantic relation between aspects by their text strings and perform finer-grained interactions between item aspect and content by mutual prediction. Our experiments on two real-world datasets show that ATTEMPT can outperform multiple competitive baselines significantly. Moreover, we release our enriched Multi-aspect Amazon Product Search dataset to encourage research on multi-aspect dense retrieval.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 61902381, the Youth Innovation Promotion Association CAS under Grants No. 2021100, the project under Grants No. JCKY2022130C039 and 2021QY1701, the Lenovo-CAS Joint Lab Youth Scientist Project. This work was also supported by Ant Group through Ant Innovative Research Program.

REFERENCES

[1] Qingyao Ai, Vahid Azizi, Xu Chen, and Yongfeng Zhang. 2018. Learning Heterogeneous Knowledge Base Embeddings for Explainable Recommendation. *Algorithms* 11, 9 (2018), 137. https://doi.org/10.3390/a11090137

[2] Qingyao Ai, Yongfeng Zhang, Keping Bi, and W. Bruce Croft. 2020. Explainable Product Search with a Dynamic Relation Embedding Model. *ACM Trans. Inf. Syst.* 38, 1 (2020), 4:1–4:29. https://doi.org/10.1145/3361738

[3] Saeid Balaneshinkordan, Alexander Kotov, and Fedor Nikolaev. 2018. Attentive Neural Architecture for Ad-hoc Structured Document Retrieval. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 1173–1182. https://doi.org/10.1145/3269206.3271801

[4] Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training Tasks for Embedding-based Large-scale Retrieval. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=rkg-mA4FDr

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[6] Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, and Jiafeng Guo. 2022. Pre-training Methods in Information Retrieval. *Found. Trends Inf. Retr.* 16, 3 (2022), 178–317. https://doi.org/10.1561/1500000100

[7] Luyu Gao and Jamie Callan. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 981–993. https://doi.org/10.18653/v1/2021.emnlp-main.75

[8] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Tevatron: An Efficient and Flexible Toolkit for Dense Retrieval. *CoRR* abs/2203.05765 (2022). https://doi.org/10.48550/arXiv.2203.05765 arXiv:2203.05765

[9] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 113–122. https://doi.org/10.1145/3404835.3462891

[10] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Real-time Inference in Multi-sentence Tasks with Deep Pretrained Transformers. *CoRR* abs/1905.01969 (2019). arXiv:1905.01969 http://arxiv.org/abs/1905.01969

[11] Omar Khattab and Matei Zaharia. 2020. 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, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 39–48. https://doi.org/10.1145/3397271.3401075

[12] Weize Kong, Swaraj Khadanga, Cheng Li, Shaleen Kumar Gupta, Mingyang Zhang, Wensong Xu, and Michael Bendersky. 2022. Multi-Aspect Dense Retrieval. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 3178–3186. https://doi.org/10.1145/3534678.3539137

[13] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 6086–6096. https://doi.org/10.18653/v1/p19-1612

[14] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool Publishers. https://doi.org/10.2200/S01123ED1V01Y202108HLT053

[15] Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is More: Pretrain a Strong Siamese Encoder for Dense Text Retrieval Using a Weak Decoder. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 2780–2791. https://doi.org/10.18653/v1/2021.emnlp-main.220

[16] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Pre-train a Discriminative Text Encoder for Dense Retrieval via Contrastive Span Prediction. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 848–858. https://doi.org/10.1145/3477495.3531772

[17] Xinyu Ma, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. A Contrastive Pre-training Approach to Discriminative Autoencoder for Dense Retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, Mohammad Al Hasan and Li Xiong (Eds.). ACM, 4314–4318. https://doi.org/10.1145/3511808.3557527

[18] Zhen Qin, Zhongliang Li, Michael Bendersky, and Donald Metzler. 2020. Matching Cross Network for Learning to Rank in Personal Search. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 2835–2841. https://doi.org/10.1145/3366423.3380046

[19] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. *CoRR* abs/2206.06588 (2022). https://doi.org/10.48550/arXiv.2206.06588 arXiv:2206.06588

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/v1/D19-1410

[21] Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2015. Mining, Ranking and Recommending Entity Aspects. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto (Eds.). ACM, 263–272. https://doi.org/10.1145/2766462.2767724

[22] Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004*, David A. Grossman, Luis Gravano, ChengXiang Zhai, Otthein Herzog, and David A. Evans (Eds.). ACM, 42–49. https://doi.org/10.1145/1031171.1031181

[23] Hongyu Shan, Qishen Zhang, Zhongyi Liu, Guannan Zhang, and Chenliang Li. 2023. Beyond Two-Tower: Attribute Guided Representation Learning for Candidate Retrieval. In *Proceedings of the ACM Web Conference 2023*. 3173–3181.

[24] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. *CoRR* abs/1904.09223 (2019). arXiv:1904.09223 http://arxiv.org/abs/1904.09223

[25] Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 538–548. https://aclanthology.org/2022.emnlp-main.35

[26] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=zeFrfgyZln

[27] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. 2018. Neural Ranking Models with Multiple Document Fields. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (Eds.). ACM, 700–708. https://doi.org/10.1145/3159652.3159730

[28] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1503–1512. https://doi.org/10.1145/3404835.3462880

[29] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. RepBERT: Contextualized Text Embeddings for First-Stage Retrieval. *CoRR* abs/2006.15498 (2020). arXiv:2006.15498 https://arxiv.org/abs/2006.15498

[30] Hongchun Zhang, Tianyi Wang, Xiaonan Meng, and Yi Hu. 2019. Improving Semantic Matching via Multi-Task Learning in E-Commerce. In *Proceedings of the SIGIR 2019 Workshop on eCommerce, co-located with the 42st International ACM SIGIR Conference on Research and Development in Information Retrieval, eCom@SIGIR 2019, Paris, France, July 25, 2019 (CEUR Workshop Proceedings, Vol. 2410)*, Jon Degenhardt, Surya Kallumadi, Utkarsh Porwal, and Andrew Trotman (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2410/paper2.pdf