GOING BEYOND TOKEN-LEVEL PRE-TRAINING FOR EMBEDDING-BASED LARGE-SCALE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the large-scale query-document retrieval problem: given a query (e.g., a question), return the set of relevant documents (e.g., paragraphs containing the answer) from a large document corpus. This problem is often solved in two steps. The retrieval phase first reduces the solution space, returning a subset of candidate documents. The scoring phase then scores and re-ranks the documents. The algorithm used in the retrieval phase is critical. On the one hand, it needs to have high recall – otherwise some relevant documents won't even be considered in the scoring phase. On the other hand, it needs to be highly efficient, returning the candidate documents in time sublinear to the total number of documents. Unlike the scoring phase which witnessed significant advances recently due to the BERT-style cross-attention models, the retrieval phase remains less well studied: most previous works rely on the classic Information Retrieval (IR) methods such as BM-25 (token matching + TF-IDF weights). In this paper, we conduct a comprehensive study on different retrieval algorithms and show that the two-tower Transformer models with properly designed pre-training tasks can largely improve over the widely used BM-25 algorithm. The pre-training tasks we studied are Inverse Cloze Task (ICT), Body First Selection (BFS), Wiki Link Prediction (WLP) and the combination of them.

1 INTRODUCTION

We consider the large-scale retrieval problem: given a query, return the most relevant documents from a large corpus, where the size of the corpus can be hundreds of thousands or more. One can view this problem as learning a scoring function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, that maps a pair of query and document $(\boldsymbol{q}, \boldsymbol{d}) \in \mathcal{X} \times \mathcal{Y}$ to a score $f(\boldsymbol{q}, \boldsymbol{d})$. The function should be designed such that the relevant $(\boldsymbol{q}, \boldsymbol{d})$ pairs have high scores, whereas the irrelevant ones have low scores. Many real-world applications besides query-document retrieval can be cast in this form. For example, in recommendation systems, $\boldsymbol{q}$ represents a user query and $\boldsymbol{d}$ represents a candidate item to recommend (Krichene et al., 2019). In extreme multi-label classification, $\boldsymbol{q}$ represents a web-page document and $\boldsymbol{d}$ represents the categories or hashtags of interests (Jain et al., 2019; Chang et al., 2019). In open-domain question answering, $\boldsymbol{q}$ represents a question and $\boldsymbol{d}$ represents an evidence passage containing the answer (Chen et al., 2017; Hu et al., 2019; Lee et al., 2019).

Central to the above is designing the scoring function f. Recently BERT (Devlin et al., 2019), along with its successors XLNet (Yang et al., 2019b) and RoBERTa (Liu et al., 2019), leads to significant improvements to many NLP tasks such as sentence pairs classification and question-answering. In BERT, the scoring function f is a pre-trained deep bidirectional Transformer model. While BERT-style models are very successful, it cannot be directly applied to large-scale retrieval problems because computing $f(\boldsymbol{q}, \boldsymbol{d})$ for every possible document can be prohibitively expensive. Thus, one typically first uses a less powerful but more efficient algorithm (another scoring function f) to reduce the solution space (the "retrieval phase"), and then use the BERT-style model to re-rank the retrieved documents (the "scoring phase").

The retrieval phase is critical. Ideally speaking, the algorithm should have a high recall; otherwise, many relevance documents won't even be considered in the scoring phase. The algorithm also needs to be very fast: it should return a small subset of relevant documents in time sublinear to the number of all documents. Although significant developments are advancing the scoring algorithms, the retrieval algorithms remain to be less well studied, and this is the focus of this paper.

The retrieval algorithm can be put into two categories. The first type is classic information retrieval (IR) algorithms relying on token-based matching with TF-IDF weights. One example is BM-25 (Robertson et al., 2009), which remains to be the most widely used (Nguyen et al., 2016; Yang et al., 2017; 2019a) and hard to beat (Chapelle & Chang, 2011; Lee et al., 2019) algorithm. Here the scoring function f is based on token-matching between the two high-dimensional sparse vectors with TF-IDF token weights, and retrieval can be done in sublinear time using inverted index.

The second option is to jointly embed queries and documents in the same embedding space and use an inner product or cosine distance to measure the similarity between queries and documents. Let the query embedding model be $\phi(\cdot)$ and the doc embedding model be $\psi(\cdot)$, The scoring function is

$$f(\boldsymbol{q}, \boldsymbol{d}) = \langle \phi(\boldsymbol{q}), \psi(\boldsymbol{d}) \rangle.$$

In the inference stage, retrieving relevant documents then becomes finding the nearest neighbors of a query in the embedding space. Since the embeddings of all candidate documents can be pre-computed and indexed, the inference can be done efficiently with approximate nearest neighbor search algorithms in the dense embedding space (Shrivastava & Li, 2014; Guo et al., 2016).

In this paper, we refer the above as the *two-tower retrieval model*, because the query and document embeddings are coming from two separate "towers" of neural networks. In other recent works, it is also known as the Siamese network (Das et al., 2016) or dual-encoder model (Cer et al., 2018). Compared to the sparse token-based models such as BM-25, the two-tower model can capture deeper semantic relationships within queries and documents.

In the heart of two-tower models is to design the embedding function $\phi(\cdot)$ and $\psi(\cdot)$. A modern choice is using Transformers to model the attention within queries and within documents, rather than the cross-attention between them as in the BERT model. The token-level Masked-LM pre-training task is crucial to the success of BERT-style cross-attention models. Nevertheless, what pre-training tasks are useful for improving two-tower Transformer models in large-scale retrieval, remains a crucial yet unsolved research problem. In this paper, we aim to answer this question by studying different pre-training tasks for the two-tower Transformer models. We contribute the following insight:

- The two-tower Transformer models with proper pre-training can significantly outperform the widely used BM-25 algorithm;
- Pre-training tasks such as the sentence-level Inverse Cloze Task (ICT), the paragraph-level Body First Selection (BFS), and document-level Wiki Link Prediction (WLP) hugely improve the retrieval quality, whereas the most widely used pre-training task (the token-level masked-LM) gives only marginal gains.

To the best of our knowledge, this is the first comprehensive study on pre-training tasks for efficient large-scale retrieval algorithms. The rest of the paper is organized as follows. We start by introducing the two-tower retrieval model in Section 2. The pre-training tasks are presented in 3, and the experiments and analysis are presented in Section 4. Finally, we conclude this work in Section 5.

2 The two-tower retrieval model

We begin with introducing some backgrounds of the two-tower models. Given a query $\boldsymbol{q} \in \mathcal{X}$ and a document $\boldsymbol{d} \in \mathcal{Y}$, we consider two-tower retrieval models that consist of two encoder functions, $\phi : \mathcal{X} \to \mathbb{R}^k$ and $\psi : \mathcal{Y} \to \mathbb{R}^k$ which map a sequence of tokens in $\mathcal{X}$ and $\mathcal{Y}$ to their associated embeddings $\phi(\boldsymbol{q})$ and $\psi(\boldsymbol{d})$, respectively. The scoring function $f : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ is then defined to be the inner product[1] of the embeddings

$$f(\boldsymbol{q}, \boldsymbol{d}) = \langle \phi(\boldsymbol{q}), \psi(\boldsymbol{d}) \rangle, \tag{1}$$

In this paper, we are interested in parameterizing the encoders ϕ, ψ as deep Transformer models (Vaswani et al., 2017) due to its expressive power in modeling natural language.

[1] This also includes cosine similarity scoring functions when the embeddings $\phi(\boldsymbol{q}), \psi(\boldsymbol{d})$ are normalized.

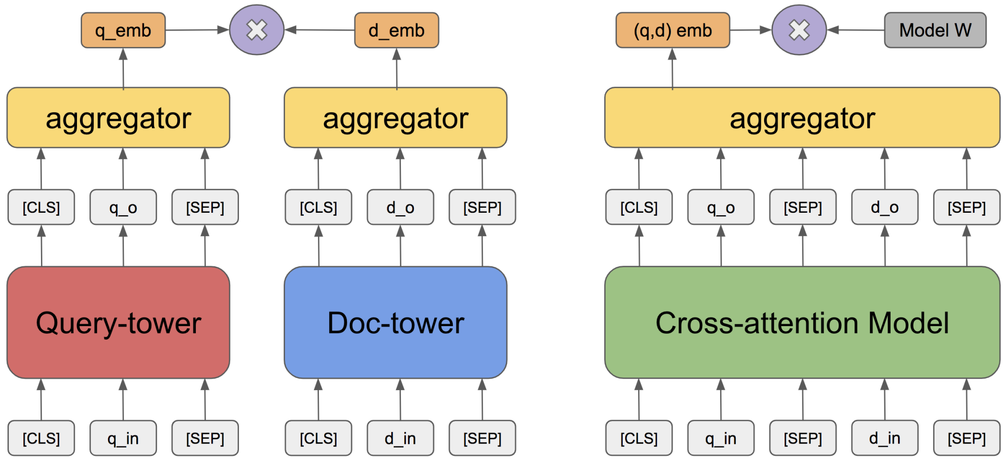

Figure 1: Difference between two-tower models and cross-attention models. Following previous works, we consider [CLS] embedding and average pooling as the aggregator's output for the two-tower Transformer model and the two-tower MLP model, respectively.

In the rest of this section, we illustrate the advantage of two-tower models in the inference phase, then discuss the pros and cons of two-tower models in comparison with BERT-like cross-attention models; present the learning procedure of estimating model parameters under maximum likelihood principle; and review the related works.

Inference The difference between two-tower models and cross-attention models is shown in Figure 1. The advantage of the two-tower model is its efficiency in large-scale retrieval in the inference time. First, all the document embeddings can be pre-computed. Then, given an unseen query $\boldsymbol{q}$, we only need to rank the document based on its inner product with the query embedding. This is way more efficient than running inference on a cross-attention BERT-style model (often used in the scoring stage). To see this, the scoring function of BERT-style model is with the form

$$f_{\theta,\boldsymbol{w}}(\boldsymbol{q},\boldsymbol{d}) = \psi_\theta(\boldsymbol{q} \oplus \boldsymbol{d})^T \boldsymbol{w}, \tag{2}$$

where $\oplus$ denotes the concatenate operation of the query and the document sequence and $\boldsymbol{w} \in \mathbb{R}^k$ is an additional model parameters. In BERT, for each query, one has to do the above expensive inference on all documents. For example, with the 128-dimensional embedding space, inner product between 1000 query embeddings with 1 million document embeddings only takes hundreds of milliseconds on CPUs, while computing the same scores with cross-attention models takes hours if not more even on GPUs.

Furthermore, retrieving the documents with the maximum inner product (MIPS) is, in fact, a well-studied problem: it can be further efficiently approximated by hashing and other indexing algorithms in sublinear time, with almost no loss in the recall (Shrivastava & Li, 2014; Guo et al., 2016).

Learning In this paper, we assume that the training data is presented as relevant "positive" query-document pairs $\mathcal{T} = \{(\boldsymbol{q}_i, \boldsymbol{d}_i)\}_{i=1}^{|\mathcal{T}|}$. Let θ be the model parameters. We estimate the model parameters by maximizing the log likelihood $\max_\theta \sum_{(\boldsymbol{q},\boldsymbol{d})\sim\mathcal{T}} \log p_\theta(\boldsymbol{d}|\boldsymbol{q})$ where the conditional probability is defined over the Softmax distribution:

$$p_\theta(\boldsymbol{d}|\boldsymbol{q}) = \frac{\exp\big(f_\theta(\boldsymbol{q},\boldsymbol{d})\big)}{\sum_{\boldsymbol{d}'\sim\mathcal{T}} \exp\big(f_\theta(\boldsymbol{q},\boldsymbol{d}')\big)}. \tag{3}$$

The Softmax involves the expensive partition function in the denominator of equation 3 that scales linearly to the number of documents. In practice, we consider Sampled Softmax, an approximation of the full-Softmax where pairs of query and document are uniformly sampled as mini-batches $\mathcal{B} = \{(\boldsymbol{q}_i, \boldsymbol{d}_i)\}_{i=1}^{B}$:

$$\hat{p}_\theta(\boldsymbol{d}|\boldsymbol{q}) = \frac{\exp\big(f_\theta(\boldsymbol{q},\boldsymbol{d})\big)\big)}{\sum_{\boldsymbol{d}'\sim\mathcal{B}} \exp\big(f_\theta(\boldsymbol{q},\boldsymbol{d}')\big)}. \tag{4}$$

Sampled Softmax has been widely used in language modeling in NLP (Chen et al., 2016; Grave et al., 2017), recommendation system (Yu et al., 2017; Krichene et al., 2019) and extreme classifications (Blanc & Rendle, 2018; Reddi et al., 2019). We then optimize the Sampled Softmax objective with variants of stochastic gradient methods such as Adam with weight decay (Kingma & Ba, 2014).

Here, the positive pairs $\mathcal{T}$ can be either from the downstream task or the pre-training tasks. Since we often have a limited amount of supervised data from the downstream task, the model is first trained with pre-training tasks and then fine-tuned with the downstream task. We will present the set of pre-training tasks we study in Section 3.

Related Works Cer et al. (2018) study the two-tower Transformer model as a universal sentence encoder. The model is learned with multiple tasks including the unsupervised Skip-Thought task, the supervised conversation input-response task (Henderson et al., 2017), and the supervised sentence classification SNLI task (Bowman et al., 2015). Humeau et al. (2019) propose the Poly-encoders architecture to balance the computation/expressiveness tradeoff between two-tower models and cross-attention models. Reimers & Gurevych (2019) fine-tune the deep two-tower models on two supervised datasets, SNLI and MNLI (Williams et al., 2018), then apply it in solving other downstream tasks. Unlike all the above works that consider training the two-tower Transformer models on a limited amount of supervised corpus for the sentence classification tasks, we study different pre-training tasks and their contributions in the large-scale retrieval settings.

Another closely related topic is open-domain question answering. Previous works consider using BM25 or other lexical matching methods to efficiently retrieve the top-k relevant passages and then deploy the more expensive cross-attention scoring function to find the answer (Chen et al., 2017; Yang et al., 2017; 2019a). Das et al. (2019) encode query and document separately with LSTM encoders. They employ a training procedure different from ours and do not consider pre-training. Very recently, Lee et al. (2019) propose to pre-train two-tower Transformer models with the Inverse Cloze Task (ICT) to replace BM25 in the passage retrieval phase. The advantage is that the retriever can be trained jointly with the reader/scorer. Nevertheless, their pre-trained two-tower models do not outperform BM25 on the SQuAD dataset, potentially because the fine-tuning is only performed on the query-tower.

Model distillation (Hinton et al., 2015) can be used to compress expensive BERT-like cross-attention models into efficient two-tower Transformer models for large-scale retrieval problems. For example, Tang et al. (2019) demonstrate initial success in distilling the BERT model into a two-tower model with BiLSTM as encoders. The pre-training tasks we study in this paper can be used as additional supervision in the distillation process, and therefore complementary to model distillation.

3 Pre-training tasks of different semantic granularities

As mentioned in Section 2, due to the limited amount of supervised data from downstream tasks, a crucial step of learning deep retrieval models is to pre-train the model with a set of pre-training tasks (we will verify this in Section 4). Sentence-level pre-training tasks have been studied before. One example is reconstructing the surface form of surrounding sentences given the encoded sentence (Le & Mikolov, 2014; Kiros et al., 2015), and another one is discriminating the next sentence from random candidates (Jernite et al., 2017; Logeswaran & Lee, 2018).

In this paper, we assume that the pre-training data is defined as positive query-document $(\boldsymbol{q}, \boldsymbol{d})$ pairs. A good pre-training task should have the following two properties. **1)** It should be relevant to the downstream task. For example, when solving the question-answering retrieval problem, the model should capture different granularities of semantics between the query and document. The semantics can be the local context within a paragraph, global consistency within a document, and even semantic relation between two documents. **2)** It should be cost-efficient to collect the pre-training data, ideally not requiring additional human supervision.

Next, we present three pre-training tasks that emphasize on different aspects of semantics between queries and documents: Inverse Cloze Task (ICT), Body First Selection (BFS), and Wiki Link Prediction (WLP). In specific, BFS and WLP are newly proposed in this paper. The training data of all these tasks can be freely obtained based on Wikipedia without additional manual labeling process. Figure 2 provides illustrative examples of these tasks.

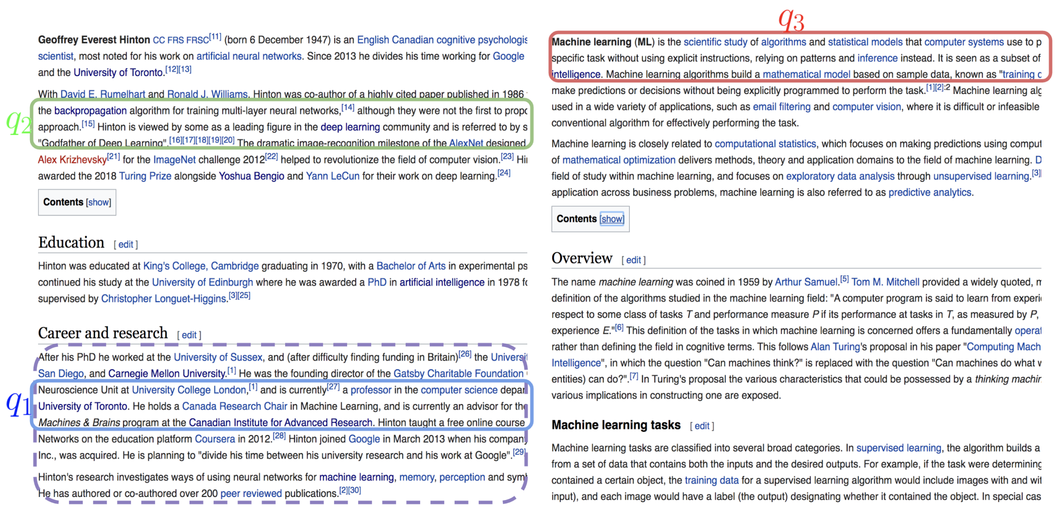

Figure 2: An illustrative example of the three pre-training tasks where each query $\boldsymbol{q}$ is highlighted in different colors. All queries are paired with the same text block $\boldsymbol{d}$. Concretely, ($\boldsymbol{q}_1$,$\boldsymbol{d}$) of ICT is defined locally within a paragraph; ($\boldsymbol{q}_2$,$\boldsymbol{d}$) of BFS is defined globally within an article; ($\boldsymbol{q}_3$,$\boldsymbol{d}$) of WLP is defined distantly across two related articles hyper-linked by the Wikipedia entity.

Inverse Cloze Task (ICT) Given a passage $\boldsymbol{p}$ consisting of n sentences $\boldsymbol{p} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, the query $\boldsymbol{q}$ is a sentence randomly drawn from the passage $\boldsymbol{q} = \boldsymbol{s}_i, i \sim [1, n]$ and the document is others sentences $\boldsymbol{d} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{i-1}, \boldsymbol{s}_{i+1}, \ldots, \boldsymbol{s}_n\}$. See the ($\boldsymbol{q}_1$,$\boldsymbol{d}$) in Figure 2 for an example. This task captures the semantic context of a sentence and is originally proposed by Lee et al. (2019).

Body First Selection (BFS) We propose BFS to capture semantic relationship outside of the local passage. Here, the query $\boldsymbol{q}_2$ is a random sentence in the first section of a Wikipedia page, and the document $\boldsymbol{d}$ is a random passage from the same page (Figure 2). Since the first section of a Wikipedia article is often the description or summary of the whole page, we expect it contains information central to the topic.

Wiki Link Prediction (WLP) We propose WLP to capture inter-page semantic relation. The query $\boldsymbol{q}_3$ is a random sentence in the first section of a Wikipedia page, and the document $\boldsymbol{d}$ is a passage from another page where there is a hyperlink link to the page of $\boldsymbol{q}_3$ (Figure 2). Intuitively, a hyperlink link indicates relationship between the two Wikipedia pages. Again, we take a sentence from the first section because it is often the description or summary of the topic.

Masked LM (MLM) In addition to the above tasks, we also consider the classic masked language model (MLM) pre-training task as a baseline: predict the randomly masked tokens in a sentence. MLM is the primary pre-training task used in BERT (Devlin et al., 2019).

4 Experiments

4.1 Experimental Setting

The two-tower retrieval model Each tower of the retrieval model follows the architecture and hyper-parameters of the 12 layers BERT-base model. For both towers, the final embedding is generated by applying a linear layer on the hidden state of the [CLS] token. The embedding dimension is 512. The sequence length for the query encoder and document encoder are set to be 64 and 288, respectively. We pre-train the model on 32 TPU v3 chips for 100K steps with an Adam optimizer and batch size of 8192. This process takes about 2.5 days. The Adam optimizer has an initial learning rate 1e-4 with the warm-up ratio 0.1, followed by a linear learning rate decay. For fine-tuning, the learning rate of Adam is set to 3e-5 with 2000 training steps and batch size 512.

Pre-training tasks	#tokens	#pairs	Query length	Document length
ICT	11.2B	50.2M	30.41	193.89
BFS	3.3B	17.5M	28.02	160.46
WLP	2.7B	24.9M	29.42	82.14

Table 1: Data statistics of three pre-training tasks.

ReQA Dataset	#query	#documents	#pair	Query length	Document length
SQuAD	97,888	101,951	99,024	11.55	291.35
Natural Questions	74,097	239,008	74,097	9.29	352.67

Table 2: Data statistics of ReQA benchmark. Query denotes the question while document represents the (answer sentence, surrounding context) pair.

Pre-training tasks The pre-training tasks we consider are the token-level task MLM, the sentence-level task ICT, the paragraph-level task BFS and the document-level task WLP. The data of ICT, BFS and WLP are generated with the Wikipedia data. The data statistics are reported in Table 1. Note that #tokens represents number of sub-words tokenized by WordPiece (Wu et al., 2016). The pre-training tasks define the positive $(\boldsymbol{q}, \boldsymbol{d})$ pair for learning the two-tower Transformer models. For ICT, the $\boldsymbol{d}$ is a pair of article title and passage separated by [SEP] symbol as input to the doc-tower.

We propose to pre-train the two-tower Transformer models jointly with all three sentence-level pre-training tasks, hence the name ICT+BFS+WLP. Specifically, the model is pre-trained on one combined set of $(\boldsymbol{q}, \boldsymbol{d})$ pairs, where each pair is uniformly sampled from the three pre-training tasks in Table 1. See Section 4.2 and 4.3 for its outstanding performance over other baselines.

Downstream tasks We consider the sentence-level retrieval task of question answering. The evaluation is based on the Retrieval Question-Answering (ReQA) version of SQuAD and Natural Questions datasets (Ahmad et al., 2019)[2]. In original QA datasets, each entry is a tuple $(\boldsymbol{q}, \boldsymbol{a}, \boldsymbol{p})$, where $\boldsymbol{q}$ is the question, $\boldsymbol{a}$ is the answer span, and $\boldsymbol{p}$ is the evidence passage containing answer $\boldsymbol{a}$. In our retrieval task, a passage is split into sentences, $\boldsymbol{p} = \boldsymbol{s}_1 \boldsymbol{s}_2 \ldots \boldsymbol{s}_n$. In the converted dataset, each entry is transformed to a tuple of $(\boldsymbol{q}, \boldsymbol{s}_i, \boldsymbol{p})$ where $\boldsymbol{s}_i$ is the sentence contains the answer span $\boldsymbol{a}$.

The retrieval problem is that given a question $\boldsymbol{q}$, retrieve the correct $(\boldsymbol{s}, \boldsymbol{p})$ pair from the document corpus. For each passage $\boldsymbol{p}$ we create a set of candidates document $(\boldsymbol{s}_i, \boldsymbol{p})$ where $i = 1 \ldots n$, and the document corpus is built by combining such pairs for all passages. This problem is more challenging than retrieving the evidence passage only since the larger number of candidates to be retrieved.

The data statistics of SQuAD (Rajpurkar et al., 2016) and Natural Questions (Kwiatkowski et al., 2019) as the ReQA benchmark are shown in Table 2.

For each dataset, we consider different training/test split of the data ($1\%/99\%$, 5%/95%, $10\%/90\%$ and, $80\%/20\%$) in the fine-tuning stage. The split is created assuming a cold-start retrieval scenario where the queries in the test (query, document) pairs are not seen in training.

4.2 Main Results

Table 3 and Table 4 compare the proposed combination of pre-training methods, namely ICT+BFS+WLP, to various baselines on SQuAD and Natural Questions, respectively. In both benchmarks, ICT+BFS+WLP notably outperforms all other methods. This suggests that *one should use a two-tower Transformer model with properly designed pre-training in the retrieval stage to replace the widely used* BM-25 *algorithm.* We present some of the detailed findings below.

[2]Different from (Ahmad et al., 2019), whose goal is to use other large-scale weakly-supervised query-answer pair datasets (e.g. reddit data) to improve the model, the goal of this paper is to study different unsupervised pre-training tasks not identical to the downstream task. Therefore our approaches are not directly comparable to the results presented in their paper.

train/test ratio	Method	R@1	R@5	R@10	R@50	R@100
1%/99%	BM-25	**41.86**	58.00	63.64	74.15	77.91
	MLP	0.00	0.01	0.02	0.12	0.25
	No-SSL	0.02	0.06	0.08	0.31	0.54
	MLM	0.18	0.51	0.82	2.46	3.93
	ICT +BFS +WLP	37.43	**61.48**	**70.18**	**85.37**	**89.85**
5%/95%	BM-25	41.87	57.98	63.63	74.17	77.91
	MLP	0.01	0.03	0.09	0.49	1.03
	No-SSL	0.17	0.36	0.54	1.43	2.17
	MLM	1.19	3.59	5.40	12.52	17.41
	ICT +BFS +WLP	**45.90**	**70.89**	**78.47**	**90.49**	**93.64**
80%/20%	BM-25	41.77	57.95	63.55	73.94	77.49
	MLP	4.66	22.36	35.51	63.17	71.66
	No-SSL	12.32	26.88	34.46	53.74	61.53
	MLM	27.34	49.59	58.17	74.89	80.33
	ICT +BFS +WLP	**58.35**	**82.76**	**88.44**	**95.87**	**97.49**

Table 3: Recall@k on SQuAD. Numbers are in percentage (%).

The BM-25 baseline BM-25 is the state-of-the-art unsupervised retrieval method based on token-matching with TF-IDF weights. The performance of BM-25 is competitive on SQuAD because SQuAD is biased towards question-answer pairs with overlapping tokens. On both datasets, BM-25 outperforms the two-tower Transformer model without pre-training (No-SSL) and with the token-level masked language model pre-training MLM in most cases. This verifies that BM-25 is a robust retrieval model and therefore widely used in recent works (Yang et al., 2017; Lee et al., 2019).

Encoder architecture We also justify the use of Transformer as encoders by comparing it with the simple MLP model. MLP lookups the uni-gram representations from the embedding table (we empirically found that adding bi-grams does not further improve the performance on these tasks possibly due to over-fitting), aggregate the embeddings with average pooling, and pass them through a shallow two-layer MLP network with tanh activation to generate the final 512-dimensional query/document embeddings.

For the two-tower models, we compare the MLP encoder vs. the Transformer encoder (No-SSL). Although without pre-training tasks both of them are not performing well, we see that the quality of MLP is much worse than Transformer in most cases, confirming that modeling attentions within queries and documents is important.

Pre-training tasks When pre-training the two-tower Transformer model, we compare the pre-training tasks to two baselines: No-SSL and MLM. No-SSL represents no pre-training, and MLM is the token-level Masked-LM task introduced in Section 3.

On both datasets, the token-level pre-training task MLM only marginally improves over the no-pretraining baseline No-SSL, whereas combining the sentence/paragraph/document-level pretraining tasks ICT+BFS+WLP provides huge boost on the performance. This verifies our assumption that the design of task-related pre-training tasks is crucial. The performance of adding individual pre-training tasks will be presented in the next section.

4.3 Ablation Study

We conduct a more thorough ablation study on Natural Questions involving (1) the number of layers in Transformer; (2) different pre-training tasks; and (3) dimension of the embedding space. The result is presented in Table 5.

Index 1, 2, and 3 show the individual performance of three pre-training tasks. All of these tasks are much more effective than MLM. Among them, ICT has the best performance, followed by BFS, and then WLP. This suggests that the (query,document) pairs defined by local context within passage are suitable for the ReQA task.

train/test ratio	Method	R@1	R@5	R@10	R@50	R@100
1%/99%	BM-25	4.99	11.91	15.41	24.00	27.97
	MLP	0.00	0.00	0.01	0.04	0.09
	No-SSL	0.04	0.13	0.19	0.45	0.72
	MLM	0.18	0.56	0.81	1.95	2.98
	ICT +BFS +WLP	**17.31**	**43.62**	**55.00**	**76.59**	**82.84**
5%/95%	BM-25	5.03	11.96	15.47	24.04	28.00
	MLP	0.00	0.02	0.04	0.35	1.19
	No-SSL	0.38	1.10	1.46	2.90	3.88
	MLM	1.10	3.42	4.89	10.49	14.37
	ICT +BFS +WLP	**21.46**	**51.03**	**62.99**	**83.04**	**88.05**
80%/20%	BM-25	4.93	11.52	14.96	23.64	27.77
	MLP	0.10	0.64	2.41	30.38	39.66
	No-SSL	7.49	20.11	25.40	38.26	43.75
	MLM	16.74	40.48	49.53	67.91	73.91
	ICT +BFS +WLP	**30.27**	**63.97**	**75.85**	**91.84**	**94.60**

Table 4: Recall@k on Natural Questions. Numbers are in percentage (%).

Index	Ablation Configuration			R@100 on different train/test ratio			
	#layer	SSL	emb-dim	1%	5%	10%	80%
1	4	ICT	128	77.13	82.03	84.22	91.88
2	4	BFS	128	72.99	78.34	80.47	89.82
3	4	WLP	128	56.94	68.08	72.51	86.15
4	12	No-SSL	128	0.72	3.88	6.94	38.94
5	12	MLM	128	2.99	12.21	22.97	71.12
6	12	ICT	128	79.80	85.97	88.13	93.91
7	12	ICT+BFS+WLP	128	81.31	87.08	89.06	94.37
8	12	ICT+BFS+WLP	256	81.48	87.74	89.54	94.73
9	12	ICT+BFS+WLP	512	82.84	88.05	90.03	94.60

Table 5: Ablation study on Natural Questions based on Recall@100. Index 9 represents the proposed method appeared in Table 4.

The advantage of increasing number of layers is manifest at Index 4, 5, 6, and 7. Index 6 and 7 show that ICT+BFS+WLP pre-training is better than ICT. This finding reflects the challenging nature of Natural Questions: scoring (question, answer) pair of Natural Questions needs multi-hop reasoning such as different passages within the same article or even going beyond different articles. Finally, Index 8 and 9 show the benefit of increasing the dimension of embedding space.

5 Conclusion

We conducted a comprehensive study on how different pre-training tasks help in the large-scale retrieval problem such as evidence retrieval for question-answering. We showed that the two-tower Transformer models with random initialization (No-SSL) or the unsuitable token-level pre-training task (MLM) are no better than the robust IR baseline BM-25 in most cases. With properly designed sentence-level (ICT), paragraph-level (BFS), and document-level (WLP) pre-training tasks, the two-tower Transformer models can considerably improve over the widely used BM-25 algorithm.

For future works, we plan to study how the pre-training tasks apply to other types of encoders architectures, generating the pre-training data from corpora other than Wikipedia, and how pre-training compares with different types of regularizations.

References

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. ReQA: An evaluation for end-to-end answer retrieval models. *arXiv preprint arXiv:1907.04780*, 2019.

Guy Blanc and Steffen Rendle. Adaptive sampled softmax with kernel based sampling. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 590–599, 2018.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 632–642, 2015.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. X-BERT: extreme multi-label text with bert. *arXiv preprint arXiv:1905.02331*, 2019.

Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the learning to rank challenge*, pp. 1–24, 2011.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pp. 1870–1879, 2017.

Welin Chen, David Grangier, and Michael Auli. Strategies for training large vocabulary neural language models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1975–1985, 2016.

Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. Together we stand: Siamese networks for similar question retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pp. 378–387, 2016.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. Multi-step retriever-reader interaction for scalable open-domain question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Edouard Grave, Armand Joulin, Moustapha Cissé, Hervé Jégou, et al. Efficient softmax approximation for gpus. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1302–1310. JMLR. org, 2017.

Ruiqi Guo, Sanjiv Kumar, Krzysztof Choromanski, and David Simcha. Quantization based fast inner product search. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 482–490, 2016.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. Retrieve, read, rerank: Towards end-to-end multi-document reading comprehension. In *Proceedings of InProceedings of the 57th Annual Meeting of the Association for Computa-tional Linguistics (ACL)*, 2019.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Real-time inference in multi-sentence tasks with deep pretrained transformers. *arXiv preprint arXiv:1905.01969*, 2019.

Himanshu Jain, Venkatesh Balasubramanian, Bhanu Chunduri, and Manik Varma. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 528–536. ACM, 2019.

Yacine Jernite, Samuel R Bowman, and David Sontag. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*, 2017.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3294–3302, 2015.

Walid Krichene, Nicolas Mayoraz, Steffen Rendle, Li Zhang, Xinyang Yi, Lichan Hong, Ed Chi, and John Anderson. Efficient training on very large corpora via gramian estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7:453–466, 2019.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning (ICML)*, pp. 1188–1196, 2014.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, 2016.

Sashank J Reddi, Satyen Kale, Felix Yu, Dan Holtmann-Rice, Jiecao Chen, and Sanjiv Kumar. Stochastic negative mining for learning with large output spaces. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Anshumali Shrivastava and Ping Li. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (mips). In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2321–2329, 2014.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from BERT into simple neural networks. *arXiv preprint arXiv:1903.12136*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT 2018), Volume 1 (Long Papers)*, June 2018.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1253–1256. ACM, 2017.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with BERTserini. In *InProceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT 2019): Demonstrations*, 2019a.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019b.

Hsiang-Fu Yu, Mikhail Bilenko, and Chih-Jen Lin. Selection of negative samples for one-class matrix factorization. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 363–371. SIAM, 2017.