A REVIEW OF RESEARCH IN FIRST-STAGE RETRIEVAL

Mengxue Du, Shasha Li, Jie Yu, Jun Ma, Huijun Liu and Miaomiao Li

College of Computer, National University of Defense Technology,
Changsha, China

ABSTRACT

In this paper, the first-stage retrieval technology is studied from four aspects: the development background, the frontier technology, the current challenges, and the future directions. Our contribution consists of two main parts. On the one hand, this paper reviewed some retrieval techniques proposed by researchers and drew targeted conclusions through comparative analysis. On the other hand, different research directions are discussed, and the impact of the combination of different techniques on first-stage retrieval is studied and compared. In this way, this survey provides a comprehensive overview of the field and will hopefully be used by researchers and practitioners in the first-stage retrieval domain, inspiring new ideas and further developments.

KEYWORDS

first-stage retrieval, information retrieval, natural language processing, artificial intelligence, review

1. INTRODUCTION

First-stage retrieval technology is the process of matching and retrieving text information by using computers [2]. Today, Internet users are generally facing the problem of information overload. First-stage retrieval technology is an important auxiliary means to solve the problem of information overload, which can help humans obtain important information more quickly and accurately. As one of the research hotspots in the field of natural language processing, first-stage retrieval technology effectively improves the efficiency of browsing and processing information.

In recent years, the development of deep learning technology has greatly affected the field of natural language processing. Researchers have used deep learning technology to improve traditional first-stage retrieval technology and meet the limitations of term-based retrieval systems. Traditional first-stage retrieval relies on term-matching techniques. However, term-based retrieval systems have natural defects in dealing with polysemy, synonymy, and other issues between queries and candidate texts [83]. The information used by the traditional methods is mainly based on the correlation measure reflected by the statistical frequency of terms, and generally compares the surface level, but ignores the use of the discourse information and semantic information of the document. Deep models with sufficient model capacity have greater potential to learn these complex tasks compared to traditional shallow models. Because of these potential benefits and the expectation that deep learning might achieve similar success in first-stage retrieval [14], there has been substantial growth in recent years in both academic and industrial work on applying neural networks to build retrieval models.

David C. Wyld et al. (Eds): ARIN, CSITA, ISPR, NBIOT, DMAP, MLCL, CONEDU -2023
pp. 45-64, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130705

As deep learning technology has achieved a series of breakthroughs in distributed semantics, machine translation, and other tasks, the application of related methods in first-stage retrieval tasks has also attracted wide attention. The semantic representation of the document itself has a strong structure, and the semantic units are closely related to each other. The architecture based on a bi-decoder is currently the most popular because it can avoid tedious manual feature extraction and is simple to start training [4]. However, these methods require a larger training corpus than traditional methods. In addition, although deep learning methods have to a certain extent solved some problems previously existing in first-stage retrieval, the quality of retrieval candidate lists still falls short of the relevant requirements of human subjective evaluation.

As an important research direction in the field of natural language processing, first-stage retrieval still faces a lot of problems. Half a century of continuous research has made significant progress in some first-stage retrieval tasks, but there are still many key problems to be solved to improve its application value and expand its application scope. Typically, the retrieval process of current first-stage retrieval systems is faced with the problems of poor ability to measure the relevance of query and candidate text and low retrieval efficiency. This leads to low correlation and poor real-time performance of the candidate sequences generated by retrieval.

Query：what is the interaction between surface water and groundwater in a watershed?
Ranked 1: Groundwater and surface water are essentially one resource, physically connected by the hydrologic cycle. Although water law and water policy often consider groundwater and surface water as separate resources, groundwater and surface water are functionally inter-dependent.
Ranked 2: There is more of an interaction between the water in lakes and rivers and groundwater than most people think. Some, and often a great deal, of the water flowing in rivers comes from seepage of groundwater into the streambed. Groundwater contributes to streams in most physiographic and climatic settings.
.....
Ranked 14: Ground-water interaction with surface water is a natural phenomenon dictated by the fact that the two water media are critical components of one system that some workers in this field have described as a three-dimensional watershed.

Figure 1. Retrieval results of a query in MSMARCO

The real example shown in Figure 1 is used here to illustrate the problem, and the example is taken from the MSMARCO [134] dataset. The orange part is the information that the human writes the query to focus on, and the blue part is the information that the retrieval model gets from the "relevant" paragraph to focus on. We can see that the main purpose of the query is to ask for the definition and content of the "interaction between surface water and groundwater." However, among the top two results generated by the machine, the first paragraph only focuses on the term information of "groundwater" and "surface water," which does not capture the overall tendency and purpose of the query semantics. It can be said that the retrieval list generated by the machine and the real user query goal are inconsistent. And in this case, the positive candidate that meets the purpose of the user query has been ranked after 10. It can be said that the retrieval list generated by the machine is not ideal yet.

We hope that in the future, computers can judge the relevance of queries and candidate texts like humans and can compile and retrieve a high-quality list of candidate texts. However, the poor ability of the first-stage retrieval system to measure the relevance of query and candidate text

often leads to the low ranking of the ideal item in the generated candidate list. These problems seriously damage the quality of retrieval. Low-quality candidate lists also create a lot of obstacles to the application of retrieval systems. Therefore, we urgently need to solve this critical problem. Our survey paper aims to provide a comprehensive understanding of the field to benefit researchers and practitioners in the field of first-stage retrieval.

2. BACKGROUND

The goal of information retrieval is to provide the most relevant information to the user's query [94]. As Figure 2 shows, this task mainly consists of two stages: 1) The first-stage retrieval stage: filtering the text that is relevant to the query from the candidate set. 2) The ranking stage: re-ranking the text produced in the previous stage according to the relevance score.

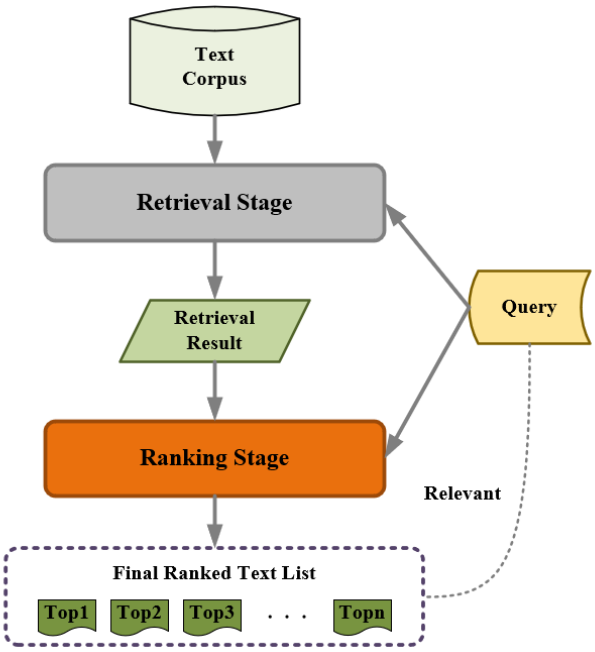

Figure 2. The architecture of a modern information retrieval system

In the first-stage retrieval stage, the retriever often needs to provide a set of documents related to the user's query from a large number of candidate data, so time efficiency and recall rate should be considered simultaneously. The techniques commonly used in this stage are vector space models [5], probabilistic models [6], learned ranking (LTR) models [7, 8], and pre-trained models such as Latent Semantic Indexing (LSI) [85], Latent Dirichlet Allocation (LDA) [86], and BERT [87].

In the ranking phase, the ranker adjusts the ranking obtained from the retrieval in the first phase based on the relevance score. Since the candidate text size is greatly reduced after filtering in the first stage, this stage focuses on improving the effectiveness of the results rather than efficiency. Traditional techniques at this stage include neural models such as RankNet [88], learned ranking [7, 8, 89], DRMM [40], and Duet [90]. These models leverage recent techniques such as reinforcement learning [91], contextual embedding [92], and attention mechanisms [93] to learn how to rank relevant texts based on a user's query terms [1].

Note that this work only focuses on the first stage of the retrieval system.

3. TERM-BASED RETRIEVAL

Term-based retrieval directly matches terms in the query and the candidate text [102]. Its advantage is that it is simple and fast, but it cannot solve problems such as synonymy and polysemy in first-stage retrieval [103, 135]. Important concepts in the field of retrieval, such as TF-IDF [105, 106], are proposed in this field. In addition, the Query Likelihood model [107] and the divergence from randomness [108] proposed in recent years use probability and language modeling techniques to improve retrieval performance, which have attracted wide interest. Other methods, such as the TDM proposed by [104], have been proposed to improve the accuracy and efficiency of retrieval systems.

3.1. Expansion Technologies

Expansion technology is a traditional improvement technology in the field of term-based retrieval, which includes query expansion and candidate text expansion. Query expansion typically identifies related terms using the semantic relations of the lexicon [109]. The main contributions in the field of query expansion include representative methods such as using local models to incorporate the obtained relevance feedback into first-stage retrieval systems [110–114] and using query context information to improve retrieval performance [81]. Correspondingly, candidate text expansion improves retrieval performance by expanding the representation of candidate texts [115, 116]. Representative methods of candidate text expansion include extending candidate text representation with relevant terms [117], extending candidate text representation with large lexical databases [118], and extending text representation with semantically relevant terms or external collections [119, 120], etc.

3.2. Topic Model

Topic models are often used by researchers to improve text representation and indexing in first-stage retrieval systems [1, 97, 100, 101]. For example, [95] introduces the concept of a generalized vector space model. This year, various topic models are often applied to first-stage retrieval. For example, researchers have proposed a new method for AD hoc retrieval that applies Latent Dirichlet Allocation to the indexing and language modeling stages [98]. Work [99] compares the performance of various topic models commonly used in first-stage retrieval, such as LDA and PLSA. Researchers in their work [96] describe the use of corpus statistics to improve ad-hoc first-stage retrieval systems.

The topic model is a kind of generative model based on the Bayesian network that reduces the dimension of high-dimensional document data to latent topic space, defines each topic as a probability distribution specific to a given dictionary, and assumes that each document is composed of all latent topics and that the proportion of the mixture follows a multinomial distribution. In recent years, topic models have been widely used in document classification, social network text analysis, and other fields due to their excellent data dimensionality reduction ability and latent semantic mining ability.

Based on different research purposes and methods, topic models are divided into supervised and unsupervised topic models. 1) Representative studies of unsupervised topic models include Latent Dirichlet Allocation (LDA) proposed by [20], in which models document words based on bags of words. In order to describe the correlation between topics, [21] proposed the Relevant Topic Model (CTM), which replaced the unreasonable Dirichlet distribution in the LDA model with the logical normal distribution and other works. 2) The most representative research in the supervised topic model includes [22]'s proposed supervised latent Dirichlet distribution model for

regression problems, but it can only deal with single-label documents. Based on the polynomial inverse regression model [23], Rabinovich et al. [24] proposed the Inverse Regression Topic Model (IRTM). Other related work will not be described.

One idea [82] is to combine a topic model and clustering algorithm to generate topic-related candidate text clusters and construct topic-aware contrastive learning samples, so as to bridge the inconsistency between retrieval matching relevance and semantic similarity and improve the relevance measurement ability of the retrieval system.

3.3. Other Term-based Technologies

Other technologies related to term-based retrieval include lexical dependency, multilingual retrieval, and other models. The lexical dependency model mainly studies the relationship between terms in the text and their influence on retrieval performance [121]. Representative work includes proposing new term dependency weighting schemes [122–125] to improve model retrieval performance. In addition, [126] improved retrieval efficiency by capturing term relationships through a novel dependency language model. The method of combining kernel function with BM25 retrieval was proposed by [127] for first-stage retrieval relevance ranking. Multilingual retrieval models are another research hotspot in the field of information retrieval, which treats the retrieval task as a statistical translation problem [128]. Some representative works estimate statistical translation models using temporary information from mutual information or click volume data [129, 130]. In addition, theoretical analysis work in this field [131–133] has demonstrated the potential of using statistical translation models to improve information retrieval systems.

4. DENSE RETRIEVAL

With the application of deep learning in the field of retrieval, Dense vector Retrieval (DR) is used to improve retrieval performance in the first stage [3]. The DR model can be divided into the representation-based retrieval model and the interactive retrieval model.

4.1. Representation-based Retrieval Model

The basic assumption of the representation-based retrieval model is that the relevance between query and candidate text depends on the semantic composition of the text. Therefore, such models usually use deep neural networks as the representation functions φ and ψ of the query and the candidate text to obtain a high-level representation of the input query q and the candidate text t, and use some simple evaluation function g (such as Euclidean distance, cosine function or multilayer perceptron) to produce the final relevance score.

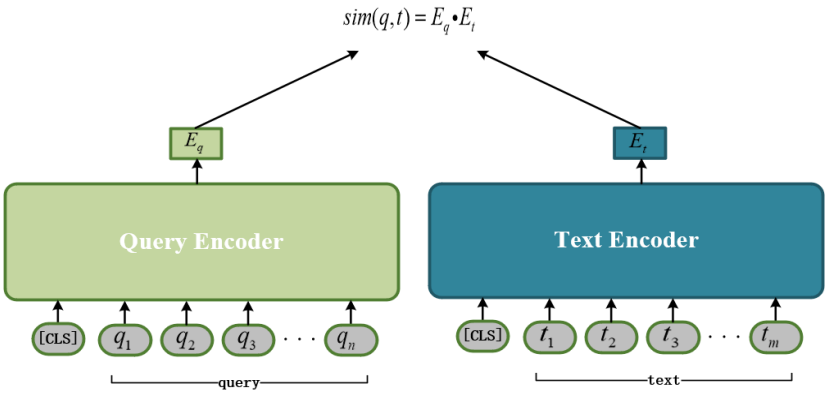

Figure 3. Schematic of the representation-based retrieval model

Early applications of deep network structures in φ and ψ mainly include fully connected networks, convolutional networks, and recurrent networks. To the best knowledge of the authors, DSSM [15] is the first work to implement the functions of the representation functions φ and ψ using fully connected networks. This was followed by works such as Arc-I [16], CNTN [17], and CLSM [18], in which both φ and ψ used convolutional networks. Taking Arc-I as an example, the stacked convolutional layer and the Max pooling layer are applied to the input query q and the candidate text d respectively, to generate their high-level representations, and then the two representations are concatenated and an MLP is applied as the evaluation function g. CNTN adopts a similar structure to Arc-I, with the difference that a different evaluation function is used. However, recurrent networks have been used to implement φ and ψ in works such as LSTM-RNN [19] and MV-LSTM [20].

With the extensive research and application of pre-trained deep language model technology, deep bidirectional language model Bert [21] and other models have become mainstream text representation encoders. Recent advances in natural language understanding have driven rapid advances in first-stage retrieval, and the pretraining-fine-tuning paradigm has been highly effective in learning text encodings. Recent works employ BERT-based models as deep bidirectional encoders to encode input text sequences into vector representations [22–28], and such models are also known as dense encoders or dual encoders. By fine-tuning the actual semantic vector similarity association, vectors can realize text comparison and retrieval through inner product methods.

As shown in Figure 3, the representational retrieval model uses a dual encoder to map the query and the candidate text encoding into the same semantic vector space and constructs a matching function to score the matching degree of the query and the candidate text. The advantage of this method is that the representation vectors of all candidate texts can be calculated offline, and only the representation vectors of the query need to be calculated and then matched online, which has a small amount of calculation and a fast running speed [22, 24–26]. The disadvantage is that the representation vectors learned by the query and candidate text are static, which creates a certain semantic drift and makes the final matching accuracy low [22, 23, 25, 28].

Representational models can be further divided into two categories: 1) Single-vector representational models; 2) Multi-vector representational models.

The single vector representation model uses a single vector to represent the query and the candidate text, respectively, which is the mainstream architecture in the current industry and the

point of academic focus. Representative works include DPR [29], RAG [30], RepBERT [31], CoRT [32], RocketQA [33], etc.

A multi-vector representation retrieval model uses multiple vectors to represent text, mainly including representation based on different language granularities (words, short phrases, sentences, etc.) and vector expansion generation. Representative works include MUPPET [34], ME-BERT [35], ColBert [36], DensePhrases [37], etc.

The advantage of this method is that the representation vectors of all candidate texts can be computed offline, while only the representation vectors of the query need to be computed and matched online, so the computational complexity is small and the operation speed is fast. The disadvantage is that the interaction between the query and the candidate text occurs only when the final match score is computed. The vectors learned for query and candidate texts are static and represented the same in different contexts, leading to a certain semantic drift and a lower final matching accuracy.

4.2. Interactive Retrieval Model

The underlying assumption of interactive retrieval models is that relevance is essentially about the relationship between a query and a candidate text, so learning directly from interactions is more effective than learning from individual representations. These models calculate correlation scores by defining interaction functions rather than representation functions and abstracting the interactions with a complex evaluation function g (i.e., a deep neural network).

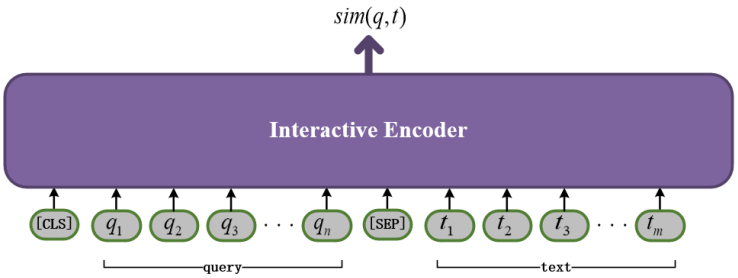

Figure 4. Schematic of the interactive retrieval model

As shown in Figure 4, the interactive retrieval model adopts a cross-encoder. The query and candidate texts will be mutually calculated during encoding so that the encoding can better integrate contextual information and highlight key semantic content. The advantage of the interaction model is that it can achieve more fine-grained matching and effectively improve the matching effect. The disadvantage is that the candidate text cannot be encoded offline, and all the computations are done online, which is time-consuming to run. When comparing different interaction functions, they can be divided into two categories, namely, nonparametric interaction functions and parametric interaction functions.

Non-parametric interaction functions refer to functions that reflect the distance between inputs without learnable parameters, such as binary indicator functions, cosine similarity functions, dot product functions, and radial basis functions [38]. Another part of the work defines the interaction between a word vector and a set of word vectors, such as the matching histogram mapping in DRMM [39] and the kernel pooling layer in K-NRM [4, 40].

Parametric interaction functions use parametric interaction functions to learn a similarity function from data. The parameterized interaction function is usually adopted when the training data is sufficient because it can improve the flexibility of the model while reducing its complexity. For example, Arc-II [41] uses convolutional layers to enable the interaction between two phrases. Match-SRNN [42] introduces neural tensor layers to model complex interactions between input words. State-of-the-art work on interactive models is BERT-based models [43] that use attention as an interaction function, learning interaction vectors (i.e., the [CLS] vector) between inputs.

This method uses a cross-encoder to interactively calculate the query and candidate text during the encoding process so that the encoding can better fuse the context information and highlight the key semantic content. The specific interaction scheme includes the early multi-level similarity calculation and the current different granularity interaction calculation based on the attention structure. The advantage of this interaction model is that it can achieve more fine-grained matching and effectively improve the matching effect. The disadvantage is that the candidate text cannot be encoded offline and all computations are done online, which is poor in real-time.

5. CURRENT CHALLENGES

5.1. Deficiencies in Text Representation

Recent advances in natural language understanding have driven rapid advances in first-stage retrieval, and the pretraining-fine-tuning paradigm has been highly effective in learning text encodings. In recent years, most of the state-of-the-art BERT-based series models have been used as deep bidirectional encoders to encode the input text sequence into a vector representation, which is also known as dense encoders or dual encoders. By fine-tuning the actual semantic vector similarity association, vectors can realize text comparison and retrieval by inner product and other methods.

Current first-stage retrieval systems that balance quality and efficiency generally encode text sequences such as sentences and paragraphs into a single dense vector representation by fine-tuning deep language models to achieve efficient text comparison and retrieval. In addition to the large amount of data and complex techniques required for effective training in the fine-tuning stage, another key problem is that the internal attention structure of advanced bidirectional language encoding models in the NLU field cannot directly meet the output requirements of dense retrieval encoders, and further aggregation of text information into a single dense representation is required. Existing works solve this problem in a mild way, which can be summarized into two categories: 1) Selecting the fine-tuned CLS vector (BERT preset) as the representation of the entire text; 2) The final output sequences of standard pre-trained bidirectional language encoding models are simply fused into a single vector: sequence averaging, Max pooling, etc.

It is worth noting that the above two ways of obtaining a single vector representation of the text have certain defects: 1) For the method of using Bert's preset CLS vector as a single dense representation of the text, the underlying assumption is that CLS can well integrate the context information of the text token sequence, thus providing a global representation of the text. However, this method lacks sufficient theoretical basis and experimental support. From the perspective of theoretical origin, CLS is a preset vector added to the first place of the input text by BERT, which is applied to the NSP task, one of the two pre-training tasks of BERT. In the NSP task, the final hidden state of CLS is passed through the MLP as the sentence pair coherence score. Due to the lack of consistency between the pre-training task and the downstream text

representation task serving retrieval, CLS does not have the natural advantages of text representation. From the perspective of experimental analysis, the recent Sentence-BERT experiment shows that the CLS method is significantly worse than simple average pooling in the sentence vector generation task for similarity evaluation. 2) For the method of simply fusing the output sequences of the final layer of the standard pre-trained bidirectional language encoding model, there are the following defects: the spatial anisotropy of the representation vector is aggravated, and the semantic distribution of the representation space is fuzzy.

5.2. The Asymmetry of Query-Text

The application scenarios of first-stage retrieval mainly include ad-hoc retrieval, question answering, community question answering, and other applications such as product search, sponsored search, etc. In ad-hoc retrieval and question answering, which is the most widely used scene, there is a key problem that hinders the development of semantic matching: the query and the candidate text often have obvious differences in length, semantic information asymmetry, and syntactic structure dissimilarity. Taking the MSMARCO dataset in the field of reading comprehension released by Microsoft as an example, the average length of the query sentence is 10, and the average length of the candidate paragraph is 120, which is an order of magnitude different. The above asymmetry brings difficulties to the retrieval semantic matching task.

In the early technology of the bag-of-words model in the field of retrieval, the relevance measure was based on the common term frequency. At the same time, due to the short query sentence and few key terms, the matching terms are too restrictive. Based on this, query expansion technology emerges as the need arises, which alleviates this problem to a certain extent. Query expansion mainly includes expansion based on lexical relations [78, 79], expansion based on concepts [80], and relevance feedback methods [81]. With the development of deep learning, the combination of traditional query expansion methods and deep learning is an idea for this project.

6. FURTHER DIRECTIONS

6.1. Joint Research on Contrastive Learning and First-Stage Retrieval

Contrastive learning is a discriminative representation learning framework based on the idea of contrast, which is mainly used for unsupervised representation learning. Similar to the Masked Language Model in deep language models, contrastive learning, as a framework for representation learning using unlabeled data, is not limited to a specific model category.

In the past two years, contrastive learning has started to set off a wave in the field of computer vision. MoCo [44], SimCLR [45], BYOL [46], SimSiam [47], and other model methods based on contrastive learning ideas have emerged one after another. As an unsupervised representational learning method, contrastive learning has outperformed supervised learning on some tasks in the field of computer vision. Since then, there have been some follow-up works in the field of natural language processing, such as ConSERT [48], SimCSE [49], etc., which use the idea of contrastive learning to learn sentence representation, and reach a level beyond SOTA in the evaluation of the Semantic Text Similarity Matching (STS) task.

The application method of contrastive learning in the field of text representation usually compares an example with its semantically similar example (a "positive example") and its semantically dissimilar example (a "negative example"). By designing the model structure and contrastive loss, the representations corresponding to semantically similar examples are closer in

the representation space, and the representations corresponding to semantically dissimilar examples are farther away in order to achieve the effect of class-like clustering.

The goal of contrastive learning applied to text representation is to learn a high-quality semantic representation space from data. In recent years, BERT-flow [76] and BERT-whitening [77] and other works have analyzed the problems of space collapse in text representation space and proposed improved methods. Recent work [48] has proposed two indicators to measure the quality of representation space: alignment and uniformity. Alignment requires that representations of similar examples be as close as possible in space, while uniformity requires that representations of dissimilar examples be uniformly distributed on the hypersphere. Since the uniform distribution has the highest information entropy, the more uniform the distribution is, the more information is retained, and the representation space satisfying alignment and uniformity is considered to be more ideal.

Some important works on contrastive learning in the field of text representation, such as ConSERT and SimCSE, focus on sentence representation and are evaluated on the STS task of text similarity matching. We have directly grafted the above works into the first-stage retrieval task, and the evaluation results show that the above works are not improving as expected. There is a certain difference between the "relevance" of semantic matching in the retrieval domain and the "similarity" between sentence pairs. In the widely used ad-hoc retrieval and traditional question-answering tasks, the length and semantics of the query sentence and the candidate text are asymmetric, and the grammatical structure is different. The positive query-candidate text pair does not strongly depend on the global semantic and syntactic structure similarity. On the contrary, in the STS series of tasks, the target score is strongly related to the semantic, grammatical, and length similarity of the sentence, which may be the main reason for the failure of the direct grafting of contrastive learning in the field of sentence similarity to the retrieval field.

6.2. Joint Research on Knowledge Distillation and First-Stage Retrieval

With the development of deep learning, the performance of many NLP tasks has reached unprecedented levels. Researchers believe that complex models can significantly improve the learning performance of deep learning tasks, but at the same time they consume a lot of storage space and computing resources. To solve this problem, the method of model compression greatly alleviates the problem of insufficient computing resources and storage space, and the knowledge distillation method is a specific method under model compression.

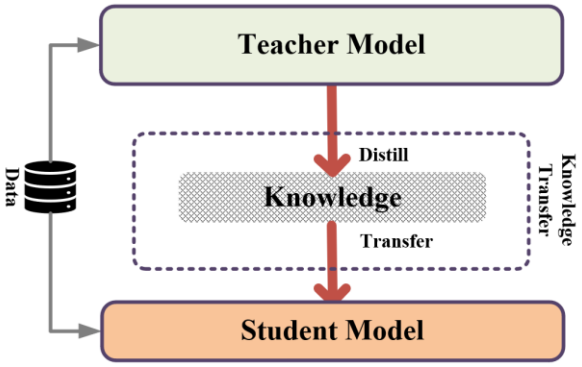

Figure 5. Schematic of the general framework for knowledge distillation

The knowledge distillation model adopts the method of transfer learning, as shown in Figure 5. By using the output of the pre-trained Teacher model as a supervision signal to train another simple Student model, the complex model can be viewed as a large-scale network structure obtained under strong constraints. Distillation aims to extract a small application-side model from a complex, large-scale model. The difficulty is to reduce the structure of the network while retaining the knowledge in the network. Hinton proposed a way to preserve generalization ability when converting complex models to small-scale models [50]: introduce soft targets to maximize the amount of information in complex models, i.e., entropy. The entropy value of discrete data is the highest in the case of equal probability, which is close to the case of equal probability so that the soft target can obtain more information and a smaller gradient variance in the training process to facilitate the small-scale model's ability to complete the training with fewer data and a smaller learning rate.

Work on knowledge distillation for model compression inspired by human teacher-student interactions has recently been extended to ideas such as teacher-student learning [51], mutual learning [52], assisted instruction [53], lifelong learning [54], and self-learning [55]. Extensions to knowledge distillation have focused on the compression of deep neural networks, and the resulting lightweight student networks can be easily deployed in applications such as visual recognition, speech recognition, and natural language processing. Furthermore, knowledge transfer from one model to another in knowledge distillation can be extended to other tasks such as adversarial attacks [56], data augmentation [57, 58], data privacy and security [59], etc.

The representative works of knowledge distillation applied to first-stage retrieval include TCT-ColBERT [60], TRMD [61], Margin-MSE loss [62], RocketQA [33], etc. The application in the RocketQA model helped it obtain the SOTA of the representation architecture-based question-answering retrieval model.

The current representative applications of knowledge distillation in the field of first-stage retrieval are as follows: 1) Training a high-performance cross-encoder MC (teacher encoder) from the original training set of retrieval tasks. Combined with the negative example mining training strategy, for each query, the target dual encoder MD(student encoder) randomly extracts difficult negative examples from the top k relevant texts retrieved from MC. This design is to adapt the cross-encoder to the distribution of results retrieved by the dual encoder, thus prompting the cross-encoder to be used in a subsequent step to optimize the dual encoder. 2) The general framework of knowledge distillation is directly applied to text encoders with similar frameworks to achieve model compression and efficiency optimization. For example, ColBert, a Bert-based dual-encoding retrieval model with a 12-layer Transformer block with advanced retrieval performance, is compressed to a 6-layer DistilBERT-based model through knowledge distillation technology.

First-stage retrieval models mainly include pointwise, pairwise, and listwise in terms of learning objectives. Among them, pointwise reduces the problem of ranking relevance in retrieval to a set of classification or regression problems. Specifically, given a set of query and candidate texts and their corresponding associated annotations, the pointwise learning objective tries to optimize the ranking model. In contrast to pointwise, where the final ranking loss is the sum of the losses for each document, pairwise is calculated based on the permutation of all possible document pairs. Listwise builds a loss function that directly reflects the final ranking performance of the retrieval model, and instead of comparing two documents at a time, it computes a ranking loss based on each query and its list of candidate texts.

In the work of combining knowledge distillation techniques to train first-stage retrieval models, there is a lack of targeted work on how to reconstruct the supervised signal, which needs further research.

6.3. Applications of Asymmetric Semantic Models

6.3.1. Semantic Expansion of Query

There are two methods in the semantic expansion of query: (1 Global approach: query expansion and refactoring without considering the initial return of the original query, such as optimization based on synonyms. (2 Local methods: The query is modified according to the initial results of the original query matching, such as relevance feedback and pseudo relevance feedback, and indirect relevance feedback.

6.3.2. Text Segmentation Technologies for Candidate Texts

Using text segmentation techniques, multiple-vector representations are generated for candidate texts. A long candidate text often involves several topics or multiple aspects of the same topic. If the semantic structure of the text can be automatically divided, and the semantic segment can be used as the basic processing unit of retrieval, it will greatly improve the phenomenon that the mainstream text retrieval technology takes the whole candidate text as the basic processing unit, and the text retrieval can be detailed from the original document level to the semantic segment level. Text segmentation technology can automatically identify a text as several semantic paragraphs with independent meaning according to semantic relations and distinguish them with tags for further analysis.

We believe that if we want to apply the text segmentation model to the first-stage retrieval framework, the most fundamental issues to be solved are the topic similarity measure and boundary search strategy: considering text similarity, region similarity, semantic paragraph length, a similarity weighting strategy based on sentence pair distance, and other clues to characterize the topic similarity, the appropriate boundary search strategy is selected on this basis to obtain performance advantages.

7. STRATEGIES & METRICS

7.1. Training Strategies

First-stage retrieval models can be divided into three categories from the aspect of training strategy: supervised, semi-supervised and weakly supervised learning.

Supervised learning is the most commonly used learning strategy, where query and candidate text are labeled. This data can be used as feedback through expert evaluation, crowdsourcing, or gathering from user interactions with search engines. With supervised training strategies, we can train a model using any learning objective currently in the domain, such as pointwise or pairwise. However, due to the limited labeled data, researchers can only learn models with a limited parameter space under this training paradigm, which motivates the work on learning first-stage retrieval from limited data [63, 64].

Weakly supervised learning refers to a learning strategy where the labels of the query and candidate text are automatically generated using traditional term retrieval models such as BM25 [4]. Pseudo-labels for training first-stage retrieval models were first proposed by Asadi et al. [65].

Recently, Dehghani et al. proposed using weak supervision to train neural first-stage retrieval models and observed a 35% improvement over BM25, which plays the role of weak labeling.

Semi-supervised learning has been widely studied in the field of first-stage retrieval. In the field of neural models, fine-tuning a weakly supervised model using a small amount of labeled data [66] and learning to control the learning rate [67] is an example of a semi-supervised ranking method. Recently, Li et al. [68] proposed a neural model that combines supervised and unsupervised loss functions, where the supervised loss controls query candidate text mismatch errors, while the unsupervised loss calculates candidate text reconstruction errors.

7.2. Loss Function

First-stage retrieval models can be divided into three categories in terms of loss function: pointwise, pairwise, and listwise.

7.2.1. Pointwise

The idea of pointwise is to reduce the problem of ranking relevance in retrieval to a set of classification or regression problems. The advantage of pointwise is twofold. 1) Compute a pointwise ranking objective based on each query and candidate text separately, making it simple and easy to scale. 2) The outputs of neural models learned with pointwise loss functions are often of practical significance and value in practice [4, 69]. For example, in sponsored search, by learning a model of cross-entropy loss and click-through rate, we can directly predict the probability that a user will click on a search ad, which in some use cases is more important than creating a good result list. However, the goal of pointwise is considered less effective for standard retrieval, mainly because there is no guarantee that an optimal ranking table will be generated when the model loss reaches a global minimum [70].

7.2.2. Pairwise

Pairwise focuses on optimizing the relationship between query and text. In contrast to pointwise, pairwise is calculated based on the permutation of all possible document pairs [71]. Since the performance of most retrieval tasks is evaluated based on the ranking of relevant documents, pairwise is very effective in many tasks. However, in practice, pairwise document preference does not always lead to an improvement in the final ranking metric for two reasons: 1) It is impossible to build a ranking model that can predict document preference correctly in all cases; 2) When computing most existing ranking metrics, not all document pairs are equally important [4]. This means that the performance of pairwise prediction is not equal to the performance of the final retrieval result as a list. Given this problem, research [72] further proposed the goal of learning to rank.

7.2.3. Listwise

The idea of listwise is to build loss functions that directly reflect the final ranking performance of the retrieval model. Instead of comparing two documents at a time, the listwise loss function computes the ranking loss based on each query and its list of candidate texts. While column-listwise is generally more efficient than pairwise, its higher computational cost often limits applications. Listwise is suitable for the stage of re-ranking (fine-tuning) a small set of candidate texts. Listwise targets have become increasingly popular since many current practical search systems employ neural models to rerank documents [73 – 75].

8. CONCLUSIONS

This survey provides a comprehensive overview of first-stage retrieval. We cover a wide range of topics, from earlier term-based retrieval methods to recent dense vector retrieval methods, and discuss the connections between them. In terms of structure, we focus on the current hot spot in the field: deep retrieval model learning, and summarize the model training strategies and loss functions commonly used in this field. In addition, the survey highlights current difficulties in the field and points to promising directions for future research. Overall, we hope this survey can motivate new ideas by reviewing past representative work, and we are expecting significant breakthroughs will be achieved for first-stage retrieval in the near future.

REFERENCES

[1] Hambarde, K. A., & Proenca, H. (2023). Information Retrieval: Recent Advances and Beyond. arXiv preprint arXiv:2301.08801.

[2] Zhuo Wang, Longlong Tian, Dianjie Guo and Xiaoming Jiang. “Optimization and analysis of large scale data sorting algorithm based on Hadoop..”Cornell University - arXiv(2015): n. pag.

[3] Migliori D T. Database-driven entity framework for internet of things: U.S. Patent 9,495,401[P]. 2016-11-15.

[4] Guo J, Fan Y, Pang L, et al. A deep look into neural ranking models for information retrieval[J]. Information Processing & Management, 2020, 57(6): 102067.

[5] Nallapati, Ramesh, et al. “Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond.” Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016, pp. 280–290.

[6] G. Salton, A. Wong, C. S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (11) (1975) 613–620.

[7] S. E. Robertson, K. S. Jones, Relevance weighting of search terms, Journal of the American Society for Information science 27 (3) (1976) 129–146.

[8] T.-Y. Liu, Learning to rank for information retrieval, Found. Trends Inf. Retr. 3 (3) (2009) 225–331.

[9] H. Li, Learning to Rank for Information Retrieval and Natural Language Processing, Morgan & Claypool Publishers, 2011.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Se nior, V. Vanhoucke, P. Nguyen, B. Kingsbury, T. Sainath, Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Pro cessing Magazine 29 (2012) 82–97.

[11] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.

[12] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436 EP –.

[13] Y. Goldberg, Neural network methods for natural language processing, Synthesis Lectures on Human Language Technologies 10 (1) (2017) 1–309.

[14] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.

[15] L. Pang, Y. Lan, J. Guo, J. Xu, J. Xu, X. Cheng, Deeprank: A new deep architecture for relevance ranking in information retrieval, in: Pro ceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM ’17, ACM, New York, NY, USA, 2017, pp. 257–266.

[16] 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, CIKM ’13, ACM, New York, NY, USA, 2013, pp. 2333–2338.

[17] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Advances in Neural Informa tion Processing Systems 27, Curran Associates, Inc., 2014, pp. 2042–2050.

[18] X. Qiu, X. Huang, Convolutional neural tensor network architecture for community-based question answering, in: Proceedings of the 24th Interna tional Conference on Artificial Intelligence, IJCAI’15, AAAI Press, 2015, pp. 1305–1311.

[19] 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, CIKM '14, ACM, New York, NY, USA, 2014, pp. 101–110.

[20] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, R. Ward, Deep sentence embedding using long short-term memory networks: Anal ysis and application to information retrieval, IEEE/ACM Trans. Audio, Speech and Lang. Proc. 24 (4) (2016) 694–707.

[21] S. Wan, Y. Lan, J. Guo, J. Xu, L. Pang, X. Cheng, A deep architecture for semantic matching with multiple positional sentence representations, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, AAAI Press, 2016, pp. 2835–2841.

[22] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[23] Seo M, Kwiatkowski T, Parikh A P, et al. Phrase-indexed question answering: A new challenge for scalable document comprehension[J]. arXiv preprint arXiv:1804.07726, 2018.

[24] Lee J, Sung M, Kang J, et al. Learning dense representations of phrases at scale[J]. arXiv preprint arXiv:2012.12624, 2020.

[25] Li Y, Liu Z, Xiong C, et al. More Robust Dense Retrieval with Contrastive Dual Learning[C]//Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. 2021: 287-296.

[26] Ren R, Qu Y, Liu J, et al. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking[J]. arXiv preprint arXiv:2110.07367, 2021.

[27] Zhang Y, Nie P, Geng X, et al. DC-BERT: Decoupling Question and Document for Efficient Contextual Encoding[J]. arXiv preprint arXiv:2002.12591, 2020.

[28] Yang Y, Jin N, Lin K, et al. Neural Retrieval for Question Answering with Cross-Attention Supervised Data Augmentation[J]. arXiv preprint arXiv:2009.13815, 2020.

[29] Singh D, Reddy S, Hamilton W, et al. End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering[J]. Advances in Neural Information Processing Systems, 2021, 34.

[30] Karpukhin V, Oğuz B, Min S, et al. Dense passage retrieval for open-domain question answering[J]. arXiv preprint arXiv:2004.04906, 2020.

[31] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. arXiv preprint arXiv:2005.11401, 2020.

[32] Zhan J, Mao J, Liu Y, et al. RepBERT: Contextualized text embeddings for first-stage retrieval[J]. arXiv preprint arXiv:2006.15498, 2020.

[33] Wrzalik M, Krechel D. CoRT: Complementary Rankings from Transformers[J]. arXiv preprint arXiv:2010.10252, 2020.

[34] Qu Y, Ding Y, Liu J, et al. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 5835-5847.

[35] Feldman Y, El-Yaniv R. Multi-hop paragraph retrieval for open-domain question answering[J]. arXiv preprint arXiv:1906.06606, 2019.

[36] Luan Y, Eisenstein J, Toutanova K, et al. Sparse, Dense, and Attentional Representations for Text Retrieval[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 329-345.

[37] Khattab O, Zaharia M. Colbert: Efficient and effective passage search via contextualized late interaction over bert[C]//Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020: 39-48.

[38] Lee J, Wettig A, Chen D. Phrase retrieval learns passage retrieval, too[J]. arXiv preprint arXiv:2109.08133, 2021.

[39] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, X. Cheng, Text matching as im age 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 Con ference on Information and Knowledge Management, CIKM '16, ACM, New York, NY, USA, 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, SIGIR '17, ACM, New York, NY, USA, 2017, pp. 55–64.

[42] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Advances in Neural Informa tion Processing Systems 27, Curran Associates, Inc., 2014, pp. 2042–2050.
[43] S. Wan, Y. Lan, J. Xu, J. Guo, L. Pang, X. Cheng, Match-srnn: Model ing the recursive matching structure with spatial rnn, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, AAAI Press, 2016, pp. 2922–2928.
[44] W. Yang, H. Zhang, J. Lin, Simple applications of bert for ad hoc docu ment retrieval, arXiv preprint arXiv:1903.10972
[45] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9729-9738.
[46] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.
[47] Grill J B, Strub F, Altché F, et al. Bootstrap your own latent: A new approach to self-supervised learning[J]. arXiv preprint arXiv:2006.07733, 2020.
[48] Chen X, He K. Exploring simple siamese representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15750-15758.
[49] Yan Y, Li R, Wang S, et al. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer[J]. arXiv preprint arXiv:2105.11741, 2021.
[50] Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[J]. arXiv preprint arXiv:2104.08821, 2021.
[51] Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
[52] Heo, B., Lee, M., Yun, S. & Choi, J. Y. (2019b). Knowledge distillation with adversarial samples supporting decision boundary. In: AAAI.
[53] Zhang, Y., Xiang, T., Hospedales, T. M. & Lu, H.(2018b). Deep mutual learning. In: CVPR.
[54] Mirzadeh, S. I., Farajtabar, M., Li, A. & Ghasemzadeh, H. (2020). Improved knowledge distillation via teacher assistant. In: AAAI.
[55] Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M. & Mori, G. (2019). Lifelong gan: Continual learning for conditional image generation. In: ICCV.
[56] Yuan, L., Tay, F. E., Li, G., Wang, T. & Feng, J.(2020). Revisit knowledge distillation: a teacher-free framework. In: CVPR.
[57] Papernot, N., McDaniel, P., Wu, X., Jha, S. & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In: IEEE SP.
[58] Lee, H., Hwang, S. J. & Shin, J. (2019a). Rethink ing data augmentation: Self-supervision and self distillation. arXiv preprint arXiv:1910.05872
[59] Gordon, M. A. & Duh, K. (2019). Explaining sequence level knowledge distillation as data-augmentation for neural machine translation. arXiv preprint arXiv:1912.03334.
[60] Wang, J., Bao, W., Sun, L., Zhu, X., Cao, B. & Philip, SY. (2019a). Private model compression via knowledge distillation. In: AAAI.
[61] Lin S C, Yang J H, Lin J. Distilling dense representations for ranking using tightly-coupled teachers[J]. arXiv preprint arXiv:2010.11386, 2020.
[62] Choi J, Jung E, Suh J, et al. Improving Bi-encoder Document Ranking Models with Two Rankers and Multi-teacher Distillation[J]. arXiv preprint arXiv:2103.06523, 2021.
[63] Hofstätter S, Althammer S, Schröder M, et al. Improving efficient neural ranking models with cross-architecture knowledge distillation[J]. arXiv preprint arXiv:2010.02666, 2020.
[64] H. Zamani, M. Dehghani, F. Diaz, H. Li, N. Craswell, Sigir 2018 workshop on learning from limited or noisy data for information retrieval, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, ACM, New York, NY, USA, 2018, pp. 1439–1440.
[65] D. Cohen, B. Mitra, K. Hofmann, W. B. Croft, Cross domain regulariza tion for neural ranking models using adversarial learning, in: The 41st International ACM SIGIR Conference on Research & Development in In formation Retrieval, ACM, 2018, pp. 1025–1028.
[66] N. Asadi, D. Metzler, T. Elsayed, J. Lin, Pseudo test collections for learn ing web search ranking functions, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Re□trieval, SIGIR '11, ACM, New York, NY, USA, 2011, pp. 1073–1082.

[67] X. Zhang, B. He, T. Luo, Training query filtering for semi-supervised learning to rank with pseudo labels, World Wide Web 19 (5) (2016) 833–864.
[68] M. Dehghani, A. Severyn, S. Rothe, J. Kamps, Avoiding your teacher's mistakes: Training neural networks with controlled weak supervision, CoRR abs/1711.00313. arXiv:1711.00313
[69] B. Li, P. Cheng, L. Jia, Joint learning from labeled and unlabeled data for information retrieval, in: Proceedings of the 27th International Con□ference on Computational Linguistics, COLING'18, Association for Com putational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 293–302
[70] A. Severyn, A. Moschitti, Learning to rank short text pairs with convo lutional deep neural networks, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, ACM, New York, NY, USA, 2015, pp. 373–382.
[71] K. D. Onal, Y. Zhang, I. S. Altingovde, M. M. Rahman, P. Karagoz, A. Braylan, B. Dang, H.-L. Chang, H. Kim, Q. Mcnamara, A. Angert, E. Banner, V. Khetan, T. Mcdonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. Rijke, M. Lease, Neural information retrieval: At the end of the early years, Inf. Retr. 21 (2-3) (2018) 111–182.
[72] W. Chen, T.-Y. Liu, Y. Lan, Z.-M. Ma, H. Li, Ranking measures and loss functions in learning to rank, in: Advances in Neural Information Processing Systems, 2009, pp. 315–323.
[73] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. N. Hullender, Learning to rank using gradient descent, in: Proceedings of the 22nd International Conference on Machine learning (ICML-05), 2005, pp. 89–96.
[74] Q. Ai, K. Bi, J. Guo, W. B. Croft, Learning a deep listwise context model for ranking refinement, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2018, pp. 135–144.
[75] Q. Ai, X. Wang, N. Golbandi, M. Bendersky, M. Najork, Learning groupwise scoring functions using deep neural networks, arXiv preprint arXiv:1811.04415.
[76] Q. Ai, J. Mao, Y. Liu, W. B. Croft, Unbiased learning to rank: Theory and practice, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ACM, 2018, pp. 2305–2306.
[77] Li B, Zhou H, He J, et al. On the sentence embeddings from BERT for semantic textual similarity[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 9119-9130.
[78] Su J, Cao J, Liu W, et al. Whitening sentence representations for better semantics and faster retrieval[J]. arXiv preprint arXiv:2103.15316, 2021.
[79] Lesk M E. Word-word associations in document retrieval systems[J]. American documentation, 1969, 20(1): 27-38.
[80] Voorhees E M. Query expansion using lexical-semantic relations[C]//SIGIR'94. Springer, London, 1994: 61-69.
[81] Qiu Y, Frei H P. Concept based query expansion[C]//Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. 1993: 160-169.
[82] Zamani H, Dadashkarimi J, Shakery A, et al. Pseudo-relevance feedback based on matrix factorization[C]//Proceedings of the 25th ACM international on conference on information and knowledge management. 2016: 1483-1492.
[83] Du, M., Li, S., Yu, J., Ma, J., Ji, B., Liu, H., ... & Yi, Z. (2022, September). Topic-Grained Text Representation-Based Model for Document Retrieval. In Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part III (pp. 776-788).
[84] Croft, W.B.; Metzler, D.; Strohman, T. Search engines: Information retrieval in practice; Vol. 520, Addison-Wesley Reading, 2010
[85] Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. Journal of the American society for information science 1990, 41, 391–407.
[86] Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. Journal of machine Learning research 2003, 3, 993–1022.
[87] Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[88] Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; Hullender, G. Learning to rank using gradient descent. Proceedings of the 22nd international conference on Machine learning, 2005, pp. 89–96.
[89] Burges, C.; Ragno, R.; Le, Q. Learning to rank with nonsmooth cost functions. Advances in neural information processing systems 2006,19.
[90] Mitra, B.; Diaz, F.; Craswell, N. Learning to match using local and distributed representations of text for web search. Proceedings of the 26th international conference on world wide web, 2017, pp. 1291–1299.
[91] Zhou, J.; Agichtein, E. Rlirank: Learning to rank with reinforcement learning for dynamic search. Proceedings of The Web Conference 2020, 2020, pp. 2842–2848.
[92] MacAvaney, S.; Yates, A.; Cohan, A.; Goharian, N. CEDR: Contextualized embeddings for document ranking. Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, 2019, pp. 1101–1104.
[93] Li, J.; Zeng, H.; Peng, L.; Zhu, J.; Liu, Z. Learning to rank method combining multi-head self-attention with conditional generative adversarial nets. Array 2022, 15, 100205.
[94] Yinqiong Cai, Yixing Fan, Jiafeng Guo, Fei Sun, Ruqing Zhang and Xueqi Cheng. “Semantic Models for the First-stage Retrieval: A Comprehensive Review..” ACM Transactions on Information Systems(2021): n. pag.
[95] Wong, S.M.; Ziarko, W.; Wong, P.C. Generalized vector spaces model in information retrieval. Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, 1985, pp. 18–25.
[96] Kurland, O.; Lee, L. Corpus structure, language models, and ad hoc information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, pp. 194–201.
[97] Diaz, F. Regularizing ad hoc retrieval scores. Proceedings of the 14th ACM international conference on Information and knowledge management, 2005, pp. 672–679.
[98] Wei, X.; Croft, W.B. LDA-based document models for ad-hoc retrieval. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 178–185.
[99] Yi, X.; Allan, J. A comparative study of utilizing topic models for information retrieval.European conference on information retrieval. Springer, 2009, pp. 29–41.
[100] Lu, Y.; Mei, Q.; Zhai, C. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Information Retrieval 2011, 14, 178–203.
[101] Atreya, A.; Elkan, C. Latent semantic indexing (LSI) fails for TREC collections. ACM SIGKDD Explorations Newsletter 2011, 12, 5–10.
[102] Manning, C.D.; Raghavan, P.; Sch"utze, H. Introduction to Information Retrieval; Cambridge University Press, 2008.
[103] Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval; Pearson Education, 2011.
[104] Van Rijsbergen, C.J. A theoretical basis for the use of co-occurrence data in information retrieval. Journal of documentation 1977.
[105] Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. Information processing & management 1988, 24, 513–523.
[106] Salton, G. Developments in automatic text retrieval. science 1991, pp. 974–980.
[107] Ponte, J.M.; Croft, W.B. A language modeling approach to information retrieval. ACM SIGIR Forum. ACM New York, NY, USA, 1998, Vol. 51, pp. 202–208.
[108] Amati, G.; Van Rijsbergen, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness.ACM Transactions on Information Systems (TOIS) 2002, 20, 357–389.
[109] Voorhees, E.M. Query expansion using lexical-semantic relations. SIGIR'94. Springer, 1994, pp. 61–69.
[110] Rocchio, J. Relevance feedback in information retrieval. The Smart retrieval system-experiments in automatic document processing 1971, pp. 313–323.
[111] Cao, G.; Nie, J.Y.; Gao, J.; Robertson, S. Selecting good expansion terms for pseudo-relevance feedback. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 243–250.

[112] Zhai, C.; Lafferty, J. Model-based feedback in the language modeling approach to information retrieval. Proceedings of the tenth international conference on Information and knowledge management, 2001, pp. 403–410.
[113] Lv, Y.; Zhai, C. A comparative study of methods for estimating query language models with pseudo feedback. Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 1895–1898.
[114] Zamani, H.; Dadashkarimi, J.; Shakery, A.; Croft, W.B. Pseudo-relevance feedback based on matrix factorization. Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 1483–1492.
[115] Liu, X.; Croft, W.B. Cluster-based retrieval using language models. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, pp. 186–193.
[116] Billerbeck, B.; Zobel, J. Document expansion versus query expansion for ad-hoc retrieval. Proceedings of the 10th Australasian Document Computing Symposium. Citeseer, 2005, pp. 34–41.
[117] Tao, T.; Wang, X.; Mei, Q.; Zhai, C. Language model information retrieval with document expansion. Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 2006, pp. 407–414.
[118] Agirre, E.; Arregi, X.; Otegi, A. Document expansion based on WordNet for robust IR. Coling 2010: Posters, 2010, pp. 9–17.
[119] Efron, M.; Organisciak, P.; Fenlon, K. Improving retrieval of short texts through document expansion. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 911–920.
[120] Sherman, G.; Efron, M. Document expansion using external collections. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 1045–1048.
[121] Fagan, J.L. Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and nonsyntactic methods; Cornell University, 1988.
[122] Mitra, M.; Buckley, C.; Singhal, A.; Cardie, C.; others. An analysis of statistical and syntactic phrases. RIAO, 1997, Vol. 97, pp. 200–214.
[123] Song, F.; Croft, W.B. A general language model for information retrieval. Proceedings of the eighth international conference on Information and knowledge management, 1999, pp. 316–321.
[124] Jones, K.S.; Walker, S.; Robertson, S.E. A probabilistic model of information retrieval: development and comparative experiments: Part 2. Information processing & management 2000, 36, 809–840.
[125] Nallapati, R.; Allan, J. Capturing term dependencies using a language model based on sentence trees. Proceedings of the eleventh international conference on Information and knowledge management, 2002, pp. 383–390.
[126] Gao, J.; Nie, J.Y.; Wu, G.; Cao, G. Dependence language model for information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, pp. 170–177.
[127] Xu, J.; Li, H.; Zhong, C. Relevance ranking using kernels. Asia Information Retrieval Symposium. Springer, 2010, pp. 1–12.
[128] Berger, A.; Lafferty, J. Information retrieval as statistical translation. ACM SIGIR Forum. ACM New York, NY, USA, 1999, Vol. 51, pp. 219–226.
[129] Karimzadehgan, M.; Zhai, C. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 323–330.
[130] Gao, J.; He, X.; Nie, J.Y. Clickthrough-based translation models for web search: from word models to phrase models. Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 1139–1148.
[131] Karimzadehgan, M.; Zhai, C. Axiomatic analysis of translation language model for information retrieval. European Conference on Information Retrieval. Springer, 2012, pp. 268–280.
[132] Riezler, S.; Liu, Y. Query rewriting using monolingual statistical machine translation. Computational Linguistics 2010, 36, 569–582.
[133] Gao, J.; Nie, J.Y. Towards concept-based translation models using search logs for query expansion. Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 1–10.

[134] Nguyen, T., Rosenberg, M.: Ms marco: A human generated machine reading comprehension dataset. In: CoCo@ NIPS (2016).
[135] Zhao Zhang, Biao Li, Junchao Chen and Qi Guo. “Construction of Higher-Order Smooth Positons and Breather Positons via Hirota's Bilinear Method” Nonlinear Dynamics(2021): n. pag.