Neural Ranking with Weak Supervision for Open-Domain Question Answering : A Survey

Xiaoyu Shen[†1], **Svitlana Vakulenko**[1], **Marco del Tredici**[1], **Gianni Barlacchi**[1]
Bill Byrne[1,2] **and Adrià de Gispert**[1]
[1]Amazon Alexa AI
[2]Univeristy of Cambridge
†gyouu@amazon.com

Abstract

Neural ranking (NR) has become a key component for open-domain question-answering in order to access external knowledge. However, training a good NR model requires substantial amounts of relevance annotations, which is very costly to scale. To address this, a growing body of research works have been proposed to reduce the annotation cost by training the NR model with weak supervision (WS) instead. These works differ in what resources they require and employ a diverse set of WS signals to train the model. Understanding such differences is crucial for choosing the right WS technique. To facilitate this understanding, we provide a structured overview of standard WS signals used for training a NR model. Based on their required resources, we divide them into three main categories: (1) only documents are needed; (2) documents and questions are needed; and (3) documents and question-answer pairs are needed. For every WS signal, we review its general idea and choices. Promising directions are outlined for future research.

Resource	**Weak-Supervision Signal**
Documents (§3)	Self Contrastive Learning (§3.1)
	Question Generation (§3.2)
Documents +Questions (§4)	Sparse Retriever (§4)
	Pre-trained Language Model (§4)
	Supervised Teacher Model (§4)
Documents +QA Pairs (§5)	Answer as Document (§5.1)
	Answer-Document Mapping (§5.2)
	Latent-Variable Model (§5.3)

Table 1: Overview of different weak-supervision signals, together with their required resources, that we can apply to train NR models for open-domain question-answering.

1 Introduction

Open-Domain Question Answering (ODQA) aims to provide precise answers in response to the user's questions by drawing on a large collection of documents (Voorhees et al., 1999). The majority of modern ODQA models follow the retrieve(-rerank)-read architecture: 1) given a question, a set of relevant documents are selected from a large document collection, and 2) the reader model produces an answer given this selected set and the question (Chen et al., 2017; Verga et al., 2021; Lee et al., 2021). Compared with parametric models without access to external knowledge, this architecture can better adapt to updated knowledge, offer easier interpretation and reduce hallucination (Zhu et al., 2021a; Shuster et al., 2021; Guo et al., 2022).

Conventional methods use sparse retrievers (SRs) such as TF-IDF and BM-25 in the first stage to match questions and documents via lexical overlap (Robertson and Walker, 1994), a process that may overlook semantically relevant documents with low lexical overlap with the question. Neural ranking (NR) models resolve this issue by encoding the questions and documents into dense vectors so that synonyms and paraphrases can be mapped to similar vectors through task-specific fine-tuning (Das et al., 2019; Karpukhin et al., 2020). However, training good NR models requires substantial amount of relevance annotations to perform competitively and NR models have been found to generalize poorly across domains (Thakur et al., 2021; Ren et al., 2022). In practice, collecting question-document relevance annotations is time-consuming. For a given question, an extensive annotation effort may be required to find the relevant documents. Repeating this annotation for every language and domain is not feasible (Shen et al., 2022c).

To reduce annotation costs, many techniques have been proposed to train NR models with *weak supervision (WS) signals* instead. This survey aims to provide a clear taxonomy to characterize these WS signals based on their required resources. There are three common resources that can be leveraged: (1) **Document** set, which is a bare minimum for building a ODQA system; (2) **Questions** without ground-truth relevance or answer annotations;

Findings of the Association for Computational Linguistics: EACL 2023, pages 1736–1750
May 2-6, 2023 ©2023 Association for Computational Linguistics

(3) **Question-Answer (QA) Pairs**, which are a set of already answered questions. Different applications have different levels of resource availability. For example, common domains such as e-commerce normally already have large amounts of question-answer pairs from customer services while smaller domains or low-resource languages can only have a document set without any existing questions. For each level of resource availability, we review the applicable WS signals. An overview can be seen in Table 1.

While there have been surveys that describe general neural information retrieval (IR) approaches (Mitra et al., 2018; Guo et al., 2020; Lin et al., 2021a; Zhu et al., 2021a; Guo et al., 2022), we focus specifically on the low-resource scenarios, which makes our contribution unique in this respect. The closest to our work is the BEIR benchmark for zero-shot cross-domain evaluation of IR models (Thakur et al., 2021) and its multiple related studies (Mokrii et al., 2021; Reddy et al., 2021; Wang et al., 2022; Ren et al., 2022). Nonetheless, these studies test specific algorithms but do not provide a holistic overview of how they are related. Our survey can be useful in that: (1) future WS research can use it as a reference book to compare with similar techniques. (2) It can serve as a practical guide for choosing the best WS sginals to train NR models given different availability levels of resources. (3) As the retrieve(-rerank)-read paradigm is generic and has been increasingly popular across NLP tasks like machine translation (He et al., 2021) and intent detection (Mehri and Eric, 2021), it can have broader impact in many other applications. Therefore, although this survey illustrates with the use case of ODQA, *the introduced techniques are intended to go beyond specific applications*.

In the following sections we first lay out the necessary background knowledge (§2), then explain popular WS signals when different resources are available in sections 3 to 5. In conclusion, we highlight promising directions for future work (§6).

2 Background

Neural Ranking (NR) for ODQA Let Q, D and A denote the question, document and answer set. Given a question $q \in Q$, the NR model assigns a relevance score $\mathcal{R}(q, d)$ to each $d \in D$ and selects top-k document $D_{topk} \in D$ with the highest relevance scores. Afterwards, a reader will estimate the score $\mathcal{G}(a|q, D_{topk})$ to predict the final answer $a \in A$ conditioned on both q and D_{topk}. The NR model can be implemented using various architecture with increasing model complexity. For computational efficiency, normally a bi-encoder architecture (Bromley et al., 1993) is first applied to pre-select top candidates from the whole document set, then a more complex cross-interaction model is applied to provide more accurate relevance scores only for the preselected candidates (Lee et al., 2021). The training objective for the NR model $\mathcal{R}$ can be formalized as:

$$\min_{\mathcal{R}} \mathbb{E}_{q,d^+,d^-_{1\sim n} \in Q \times D} \mathcal{L}(\mathcal{R}, q, d^+, d^-_{1\sim n}) \qquad (1)$$

where $Q \times D$ indicates the full set of question-document pairs, d^+ is a positive (relevant) document for q, $d^-_{1\sim n}$ is the sampled n negative (irrelevant) documents and $\mathcal{L}$ is the loss function. A common choice for $\mathcal{L}$ is the contrastive loss:

$$\mathcal{L} = -\log \frac{e^{\mathcal{R}(q,d^+)}}{e^{\mathcal{R}(q,d^+)} + \sum_{j=1}^{n} e^{\mathcal{R}(q,d^-_j)}} \qquad (2)$$

Neural Ranking with Weak Supervision In the standard supervised setting we need relevance annotations for $(q, d) \rightarrow \{+, -\}$ to train $\mathcal{R}$ with Eq 1. Obtaining high-quality relevance annotations requires tremendous human labor and is expensive to scale to multiple domains (Del Tredici et al., 2021; Ram et al., 2022). Weak supervision (WS) is a widely-used approach to reduce such cost by leveraging supervision signals from e,g., heuristic rules, knowledge bases or external models (Zhang et al., 2021). WS signals are cheap to obtain but might contain significant noise which will affect the NR performance. Therefore, understanding their working mechanisms and pros and cons are important to obtain a good NR model. We group WS signals into 3 classes by the resources that they need: (1) *Documents*: only document collection D is needed; (2) *Documents + Questions*: document collection D and question set Q are needed; (3) *Documents + QA Pairs*: document collection D and QA pairs (Q, A) are needed. In the next section, we will present the three classes of WS signals and discuss their pros and cons.

3 Resource: Documents

This section discusses two main techniques to produce WS signals requiring only the document set: (1) self-contrastive learning and (2) question generation. This makes the minimum assumption about

Method	Pseudo Question
Perturbation	d with added perturbation
Summary	(Pseudo) summary of d
Proximity	Nearby text of d
Cooccurrence	Text sharing cooccurred spans with d
Hyperlink	Text with a hyperlink to/from d

Table 2: Given a document d, different heuristics to construct pseudo questions. Contrastive samples made from these heuristics serve as WS signals to train the NR model.

resource availability since having the document set is a prerequisite for building an ODQA system.

3.1 Self Contrastive Learning

Self contrastive learning relies on heuristics to construct pseudo question-document pairs $(q', d'^{+/-})$ from D, then uses them to supervise training of a NR model. The objective is:

$$\min_{\mathcal{R}} \mathbb{E}_{q', d'^{+}, d'^{-}_{1\sim n} \in D} \mathcal{L}(\mathcal{R}, q', d'^{+}, d'^{-}_{1\sim n}) \quad (3)$$

where $\mathcal{L}$ is the ranking loss as in Eq 1. Since negative pairs can be easily constructed by random sampling, the main difficulty is to design good heuristics for constructing positive pseudo pairs (q', d'^{+}). There are 5 popular heuristics to construct such positive pairs: perturbation-based, summary-based, proximity-based, cooccurence-based and hyperlink-based. An overview is in Table 2.

Perturbation-based heuristics add perturbations to some text, then treat the perturbed text and the original text as a positive pair. The intuition is that *perturbed text should still be relevant to the original text*. Typical choices of perturbations include word deletion, substitution and permutation (Zhu et al., 2021b; Meng et al., 2021), adding drop out to representation layers (Gao et al., 2021), or passing sentences through different language models (Carlsson et al., 2021), among other.

Summary-based heuristics extract a summary from the document as the pseudo question based on the intuition that *questions should contain representative information about the central topic of the document*. The summary can be the document title (MacAvaney et al., 2017, 2019; Mass and Roitman, 2020), a random sentence from the first section of the document (Chang et al., 2020), randomly sampled ngrams (Gysel et al., 2018) or a set of keywords generated from a document language model (Ma et al., 2021a).

Proximity-based heuristics utilize the position information in the document to obtain positive pairs based on the intuition that *nearby text should be more relevant to each other*. The most famous one is the inverse-cloze task (Lee et al., 2019), where a sentence from a passage is treated as the question and the original passage, after removing the sentence, is treated as a positive document. They can be combined with typical noise injection methods like adding drop-out masks (Xu et al., 2022), random word chopping or deletion (Izacard et al., 2021) to further improve the model robustness. Other methods include using spans from the same document (Gao and Callan, 2022; Ma et al., 2022), sentences from the same paragraph, paragraphs from the same document as positive samples (Di Liello et al., 2022), etc.

Cooccurrence-based heuristics construct positive samples based on the intuition that *sentences containing cooccurred spans are more likely to be relevant* (Ram et al., 2021). For example, Glass et al. (2020) constructs a pseudo question with a sentence from the corpus. A term from it is treated as the answer and replaced with a special token. Passages retrieved with BM25 which also contains the answer term are treated as pseudo positive documents. Ram et al. (2022) treat a span and its surrounding context as the pseudo question and use another passage that contains the same span as a positive document.

Hyperlink-based heuristics leverage hyperlink information based on the intuition that *hyperlinked text are more likely to be relevant* (Zhang et al., 2020; Ma et al., 2021b). For example, Chang et al. (2020) takes a sentence from the first section of a page p as a pseudo question because it is often the description or summary of the topic. A passage from another page containing hyperlinks to p is treated as a positive document. Yue et al. (2022a) replace an entity word with a question phrase like "what/when" to form a pseudo question. A passage from its hyperlinked document that contains the same entity word is treated as a positive sample. Zhou et al. (2022) build positive samples with two typologies: "dual-link" where two passages have hyperlinks pointed to each other, and "co-mention" where two passages both have a hyperlink to the same third-party document.

Input	Document Document + Answer Document + Answer + Question Type Document + Answer + Question Type + Clue
Question Generator	Rule-based Generator Prompt-based Generator Fine-tuned Generator$^{\diamond}$
Filter	LM Score Round-trip Consistency Probability from pre-trained QA Influence Function Ensemble Consistency Entailment Score Learning to Reweight$^{\diamond}$ Target-domain Value Estimation$^{\diamond}$

Table 3: Different choices for a question generation model setup. $^{\diamond}$ means minimal relevance annotations are needed.

3.2 Question Generation

Self contrastive learning relies on sentences already present in D. Question generation leverage a question generator (QG) to generate new questions *not found* in D, which can then be used to provide WS signals for the NR model. It often employs a filter Fil to filter poorly generated questions. The training objective is:

$$\min_{\mathcal{R}} \mathbb{E}_{q=QG(d^+)\&Fil(q,d^+)=0}\mathcal{L}(\mathcal{R}, q, d^+, d^-_{1\sim n})$$

where the expectation is with respect to documents d^+ and $d^-_{1\sim n}$ drawn from D, the q are generated from d^+, $Fil(q, d^+) = 0$ requires that these questions not be discarded by Fil, and $\mathcal{L}$ is the standard ranking loss. There are various ways of designing the question generator and filter. We will cover the popular choices in the following section. An overview can be seen in Table 3.

Choices of Input A variety of information can be provided as input for the QG. The most straightforward approach is answer-agnostic which provides only the document (Du and Cardie, 2017; Kumar et al., 2019). In this way, the model can choose to attend to different spans of the document as potential answers and so generate different, corresponding questions. A more common method is answer-aware where an answer span is first extracted from a document, then the QG generates a question based on both the document and answer (Alberti et al., 2019; Shakeri et al., 2020). Finer-grained information can also be provided such as the question type ("what/how/...") (Cao and Wang, 2021; Gao et al., 2022) as well as additional clues (such as document context to disambiguate the question) (Liu et al., 2020). Adding more information reduces the entropy of the question and makes it easier for the model to learn, but also increases the possibility of error propagation (Zhang and Bansal, 2019). In practice, well-defined filters should be applied to remove low-quality questions.

Choices of Question Generator There are three popular choices for the question generator. (1) Rule-based methods (Pandey and Rajeswari, 2013; Rakangor and Ghodasara, 2015) rely on handcrafted templates and features. These are time-consuming to design, domain-specific, and can only cover certain forms of questions. (2) Prompt-based methods relying on pre-trained language models (PLMs). Documents can be presented to a PLM, with an appended prompt such as "*Please write a question based on this passage*" so that the PLM can continue the generation to produce a question (Bonifacio et al., 2022; Sachan et al., 2022; Dai et al., 2022). (3) Fine-tuned generators that are trained on annotated question-document pairs. When in-domain annotations are not enough, we can leverage out-of-domain (OOD) annotations, if any, to fine-tune the QG. The first two QGs require no training data, but their quality is often inadequate. In practice, we should only consider them when *there is a complete lack of high-quality supervised data for fine-tuning the QG*. When target-domain questions are available, we can also apply semi-supervised techniques such as back-training to adapt the QG to the target domain (Zhao et al., 2019; Kulshreshtha et al., 2021; Shen et al., 2022a).

Choices of Filter Filtering is a crucial part of QG since a significant portion of generated questions could be of low quality and would provide misleading signals when used to train the NR model (Alberti et al., 2019). A typical choice is filtering based on round-trip consistency (Alberti et al., 2019; Dong et al., 2019), where a pre-trained QA system is applied to produce an answer based on the generated question. A question is kept only when the produced answer is consistent with the answer from which the question is generated. We can also relax this strict consistency requirement and manually adjust an acceptance threshold based on the probability from the pre-trained QA system (Zhang and Bansal, 2019; Lewis et al., 2021), LM score from the generator itself (Shakeri et al., 2020; Liang et al., 2020), or an entailment score from a model trained on question-context-answer pairs (Liu et al., 2020). Influence functions (Cook and Weisberg, 1982) can be used to estimate the effect on the validation loss of including a synthetic example (Yang et al., 2020), but this does not achieve satisfying performances on QA tasks (Bartolo et al., 2021). Bartolo et al. (2021) propose filtering questions based on ensemble consistency, where an ensemble of QA models are trained with different random seeds and only questions agreed by most QA models are selected. When minimal target-domain annotation is available, we can also learn to reweight pseudo samples based on the validation loss (Sun et al., 2021), or use RL to select samples that lead to validation performance gains (value estimation) (Yue et al., 2022b).

3.3 Discussion

If the heuristics or QG are properly designed, NR models trained from their supervision can even match the fully-supervised performance (Wang et al., 2022; Ren et al., 2022). The biggest challenge is the difficulty to pick the most suitable heuristics or QG when we face a new domain. A general solution is to automatically select good pseudo pairs with reinforcement learning (RL) when minimal target-domain annotations are available (Zhang et al., 2020), so as avoiding the need to manually fixing the WS signals, but this would bring significant computational overhead. In practice hyperlink-based approaches often perform the best among the heuristics as they have additional reference information to leverage, which makes them most similar to the actual relevance annotations. However, hyperlink information is not available in most domains and thereby limits its use cases (Sun et al., 2021). QG-based WS signals are often preferred over heuristics-based ones as they can produce naturally-sound questions themselves without relying on the chance to find good pseudo questions in the documents. Nonetheless, obtaining a high-performing QG can also be non-trivial. One big challenge comes from the one-to-many mapping relations between questions and documents. Under this situation, standard supervised learning tends to produce safe questions with less diversity and high lexical overlap with the document. For example, Shinoda et al. (2021) found that QG reinforces the model bias towards high lexical overlap. We will need more sophisticated training techniques such as latent-variable models (Shen and Su, 2018; Xu et al., 2020; Li et al., 2022) and reinforcement learning (Yuan et al., 2017; Zhang and Bansal, 2019; Shen et al., 2019a) to alleviate the model bias towards safe questions.

4 Resource: Documents + Questions

This section includes WS signals that require additional access to a question set Q. In practice, annotating question-document relations usually requires domain experts to read long documents and careful sampling strategies to ensure enough positive samples, while unlabeled questions are much easier to obtain either through real user-generated content or simulated traffic. Therefore, it is common to have a predominance of unlabeled questions. The crucial point is to establish the missing relevance labels. Suppose a WS method can provide the missing label $\text{WS}(q,d)$ for a question-document pair (q,d), then we can use it to supervise the NR model by:

$$\min_{\mathcal{R}} \mathbb{E}_{q\in Q, d\in D}\mathcal{L}(\mathcal{R}(q,d), \text{WS}(q,d)) \qquad (4)$$

where $\mathcal{L}$ is the loss function that encourages similarity between $\mathcal{R}(q,d)$ and $\text{WS}(q,d)$.

There are three popular types models that can provide such WS signal here: (1) sparse retriever, (2) pre-trained language model and (3) supervised teacher model.

Sparse Retriever (SR) Recent research finds that NR and SR models are complementary. NR models are better at semantic matching while SRs are better at capturing exact match and handling long documents (Chen et al., 2021; Luan et al., 2021). SRs are also more robust across domains (Thakur et al., 2021; Chen et al., 2022). This motivates the use of unsupervised sparse retrievers like BM25 as WS signals. For example, Dehghani et al. (2017); Nie et al. (2018) train a NR model on samples annotated with BM25. Xu et al. (2019) apply four scoring functions to auto-label questions and documents with: (1) BM25 scores, (2) TF-IDF scores, (3) cosine similarity of universal embedding representation (Cer et al., 2018) and (4) cosine similarity of the last hidden layer activation of pre-trained BERT model (Devlin et al., 2019). Both papers observe that the resulting model outperforms BM25 on the test sets. Chen et al. (2021) further show that distilling knowledge from BM25 helps the retriever to better match rare entities and improves zero-shot out-of-domain performance.

Pre-trained Language Model (PLM) As PLMs already encode significant linguistic knowledge, there have also been attempts at using prompt-based PLMs to provide WS signals for question-document relations (Smith et al., 2022; Zeng et al., 2022). Similar as in question generation, we can use prompts like "*Please write a question based on this passage*", concatenate the document and question, then use the probability assigned by the PLM to auto-label question-document pairs. To maximize the chances of finding positive document, normally we first obtain a set of candidate documents by BM25, then apply PLM to auto-label the candidate set (Sachan et al., 2022). This can further exploit the latent knowledge inside PLMs that has been honed through pre-training, so it often shows better performance compared with weak supervision only using BM25 (Nogueira et al., 2020; Singh Sachan et al., 2022).

Supervised Teacher Model A very common choice is using a supervised teacher model to provide WS signals. The teacher model is "supervised" because it is explicitly fine-tuned on annotated question-document pairs. When in-domain annotations are not sufficient, we can leverage out-of-domain (OOD) annotations, if available, to train the teacher model. The teacher model usually employs a more powerful architecture such as with more complex interactions or larger sizes. It may not be directly applicable in downstream tasks due to the latency constraints, but can be useful in providing WS signals for training the NR model. For example, previous research has shown that models with larger sizes or late/cross-interaction structures generalize much better on OOD data (Pradeep et al., 2020; Lu et al., 2021; Ni et al., 2021; Rosa et al., 2022; Muennighoff, 2022; Zhan et al., 2022). After training a teacher model on OOD annotations, applying it to provide WS signals through target-domain question and document collections can significantly improve the in-domain performance of the NR model (Hofstätter et al., 2021; Lin et al., 2021b; Lu et al., 2022). Kim et al. (2022) further show that we can even use the same architecture and capacity to obtain a good teacher model. They expand the question with centroids of word embeddings from top retrieved passages (using BM25), and then use the expanded query for self knowledge distillation. Similar ideas of reusing the same architecture to provide WS signals have also been explored by Yu et al. (2021a); Kulshreshtha et al. (2021); Zhuang and Zuccon (2022).

Discussion The three WS signals listed above work directly on actual questions instead of pseudo pairs as in §3 so that the NR model can adapt better to the target-domain question distribution. The bottleneck is the quality of the WS signals. SRs and PLMs are unsupervised, which could be more robust when we face a completely different domain (Dai et al., 2022). Otherwise, if we already have certain amounts relevance annotations from the target or similar domains, usually using a supervised teacher model is preferred. Nevertheless, these WS signals inevitably contain noise, and can harm the downstream performance if the noise is significant. There are two main strategies to reduce the noise effects: (1) Apply less strict margin-based loss such as the hinge loss (Dehghani et al., 2017; Xu et al., 2019) and MarginMSE loss (Hofstätter et al., 2020; Wang et al., 2022), then models have fewer chances of overfitting to the exact labels, and (2) Apply noise-resistant training methods such as confidence-based filtering (Mukherjee and Awadallah, 2020; Yu et al., 2021b) and meta-learning-based refinement (Ren et al., 2018; Zhu et al., 2022). Another potential issue is that the amount of training data in this section relies on the amount of questions we have. Unlike the document set which we can obtain for free, the question set takes time to collect and are often orders of magnitudes smaller. If no sufficient questions are available, we can use synthetic questions from question generation, then apply same WS signals in this section, which has been shown to perform on par with using real questions in certain domains (Wang et al., 2020, 2022; Thakur et al., 2022).

5 Resource: Documents + QA Pairs

Many domains have large numbers of already answered questions from customer services, technical support or web forums (Huber et al., 2021). These QA pairs can provide richer information than only unlabeled questions. However, most answers are based on personal knowledge, derived from experience, and do not include a reference to any external document. This prevents their direct use as training data for the NR model. This section introduces three standard methods that exploit QA pairs to provide WS signals despite this difficulty: (1) Answer as document, (2) Answer-document mapping and (3) Latent-variable models.

5.1 Answer as a Document

As a straightforward way to leverage QA pairs, this method directly treats QA pairs as positive samples and does not distinguish between documents and answers (Lai et al., 2018). These QA pairs can provide direct WS signals to train the NR model:

$$\min_{\mathcal{R}} \mathbb{E}_{q,a^+,a^-_{1\sim n}\in Q\times A}\mathcal{L}(\mathcal{R},q,a^+,a^-_{1\sim n}) \quad (5)$$

where $(q, a^+) \in Q \times A$ are question-answer pairs in the target domain, $a^-_{1\sim n}$ are sampled n negative answers and $\mathcal{L}$ is the standard ranking loss.

Though simple, this has been a common practice to "warm up" the NR model when no sufficient relevance annotations are available. For large-sized models, this can be crucial to fully leverage the model capacity since we often have orders of magnitude more QA pairs than relevance annotations (Ni et al., 2021; Oğuz et al., 2021). However, the style, structure and format differ between the document and the answer. The answer is a direct response to the question, and so it is easier to predict due to its strong semantic correlation with the question. Whereas the document can be implicit and may contain fewer obvious clues that can imply an answer; deep text understanding is required to predict the relevance between questions and documents (Zhao et al., 2021; Shen et al., 2022b). Therefore, this approach may be insufficient to reach satisfying results as a standalone method.

5.2 Answer-document Mapping

This approach leverages an additional mapping function to automatically link answers to the corresponding documents. The NR model can get WS signals from the linked answers:

$$\min_{\mathcal{R}} \mathbb{E}_{q,a\in(Q,A),d^-_{1\sim n}\sim D}\mathcal{L}(\mathcal{R},q,M(a),d^-_{1\sim n})$$

where $(q, a) \in (Q, A)$ are question-answer pairs, M is a mapping function from an answer to its corresponding document, and $\mathcal{L}$ is the standard ranking loss. The mapping function is based on hand-crafted heuristics. For long-form descriptive answers, a popular way is to map them to documents with highest ROUGE scores (Lin, 2004) since the answers can be considered as summaries of the original documents (Fan et al., 2019). For short-span answers, a popular way is to map them to top-ranked documents retrieved using BM25 that contain the answer span (Karpukhin et al., 2020; Sachan et al., 2021; Christmann et al., 2022).

Answer-document mapping was widely adopted for constructing large-scale datasets in information retrieval (Joshi et al., 2017; Dunn et al., 2017; ElGohary et al., 2018). This can work well if the mapping has high accuracy, which is often difficult to achieve. Frequent answers or entities might lead to false positive mappings. It is also difficult to find positive documents for boolean and abstractive answers using only heuristics-based mapping functions (Izacard and Grave, 2021). Models can easily overfit to the biases introduced via such mapping function (Du et al., 2022).

Distribution of $\mathcal{R}(z\|q)$	Optimization Method
Categorical	Top-k approximation
Multinomial	EM algorithm Learning from attention

Table 4: Distribution assumptions made about the neural ranker and corresponding optimization methods, suppose we train the NR model following Equation 6.

5.3 Latent-Variable Model

We can still train the NR model on question-document pairs as in Answer-Document Mapping. However, instead of relying on a heuristic-based mapping function, we can treat this mapping as a "latent variable" within a probabilistic generative process (Lee et al., 2019; Shen, 2022). By this means, the NR model $\mathcal{R}$ gets WS signals from the QA reader $\mathcal{G}$ by maximizing the marginal likelihood:

$$\max_{\mathcal{R},\mathcal{G}} \mathbb{E}_{q,a\in(Q,A)} \log \sum_{z\sim Z} \mathcal{R}(z|q)\mathcal{G}(a|q,z) \quad (6)$$

where Z indicates all possible document combinations. Directly optimizing over Eq 6 is infeasible as it requires enumerating over all documents. A closed-form solution does not exist due to the deep neural network parameterization of $\mathcal{R}$ and $\mathcal{G}$. The following section explains popular optimization options. An overview can be seen in Table 4.

Top-k approximation A popular approach is to assume a categorical distribution for $\mathcal{R}(Z|q)$; that is, to assume for each question only a single document is selected and the answer is generated from that one document. Eq 6 can be approximated by enumerating over only the top-k documents, assuming the remaining documents having negligibly small contributions to the likelihood:

$$\max_{\mathcal{R},\mathcal{G}} \mathbb{D}_{q,a\in(Q,A)} \log \sum_{z\sim E_{topk}} \mathcal{R}(z|q)\mathcal{G}(a|q,z)$$

This has been a popular choice in end-to-end training of text generation models (Lee et al., 2019; Shen et al., 2019b; Guu et al., 2020; Lewis et al., 2020; Shuster et al., 2021; Ferguson et al., 2022). Despite its simplicity, the top-k approximation has two main drawbacks. (1) The approximation is performed on the top-k documents obtained from the NR model. If the NR model is very weak at the beginning of training, these top-k documents can be a bad approximation to the real joint likelihood and the model might struggle to converge. (2) The assumption that document follow a categorical distribution might be problematic especially if the answer requires evidence from multiple documents (Wang and Pan, 2022).

Expectation–Maximization (EM) algorithm To address the second drawback of the top-k approximation approach, we can assume a multinomial distribution for $R(Z|q)$ so that an answer can be generated from multiple documents. The cost of this relaxation is the increased difficulty of optimization. Approximating the joint likelihood from top-k samples becomes infeasible due to the combinatorial distribution of document. Singh et al. (2021) propose optimizating it with the EM algorithm under an independent assumption about the posterior distribution of $\mathcal{R}(z|q)$:

$$\max_{\mathcal{R},\mathcal{G}} \mathbb{E}_{q,a\in(Q,A)}[\log \sum_{z\in D_{topk}} \mathcal{R}(z|q) \times SG(\mathcal{G}(a|q,z)) + \log \mathcal{G}(a|q, D_{topk})] \tag{7}$$

where SG means stop-gradient (gradients are not backpropagated through $\mathcal{G}$). As can be seen, the training signal for the NR model is essentially the same as in the *Top-k Approximation* case, except that the reader is trained by conditioning on all top-k documents to generate the answer. Singh et al. (2021) also find that Eq 7 is quite robust with respect to parameter initialization. Similarly, Zhao et al. (2021) apply the hard-EM algorithm to train the NR model, which only treats documents with the highest likelihood estimated by the reader as positive. Izacard et al. (2022) further experiment with using the leave-one-out perplexity from the reader to supervise the ranker.

Learning from attention Another way to optimize the NR model in Eq 6 is to leverage attention scores from the reader $\mathcal{G}$. The assumption is that when training $\mathcal{G}$ to generate the answer, its attention score is a good approximation of question-document relevance. The training objective is:

$$\min_{\mathcal{R},\mathcal{G}} \mathbb{E}_{q,a\in(Q,A)} \sum_{z\sim E_{topk}} \mathcal{L}(A_z|\mathcal{R}(z|q)) - \log \mathcal{G}(a|q, Z = D_{topk}) \tag{8}$$

where $\mathcal{G}$ is trained to generate the right answer based on the question and the top-k document, same as in the EM algorithm. A_z is the attention score of $\mathcal{G}$ on the document z. $\mathcal{L}$ is the loss function to encourage the similarity between distributions of the attention scores and retrieving scores.

Izacard and Grave (2021) propose a training process that optimizes $\mathcal{R}$ and $\mathcal{G}$ iteratively. $\mathcal{R}$ is trained to minimize KL divergence between relevance and attention scores. (Lee et al., 2021) jointly optimize $\mathcal{R}$ and $\mathcal{G}$ and apply a stop-gradient operation on $\mathcal{G}$ when updating $\mathcal{R}$. Sachan et al. (2021) use retriever scores to bias attention scores on the contrary. These can be considered as first-order Taylor series approximations of Eq. 6 by replacing $\mathcal{R}(Z|q)$ with attention scores (Deng et al., 2018).

Discussion Training with latent-variable models can perform close to fully supervised models under certain scenarios (Zhao et al., 2021; Sachan et al., 2021). The main challenge is the training difficulty. In practice, we can often initialize the NR model using the *answer as document* or *answer-document mapping* to make the training more stable. If not enough QA pairs are available, we can use heuristics like masked salient entities (Guu et al., 2020) to form pseudo pairs, then apply the same WS techniques in this section. Combining supervision signals from various various optimization techniques such as learning from attention and EM algorithm can also be beneficial (Izacard et al., 2022). If the independence assumption made by Eq 7 does not hold, we need to resort to more complex optimization algorithms. A potential direction is to apply a Dirichlet prior over $R(z|q_t)$, which is a conjugate distribution to the multinomial distribution (Minka, 2000), with the result that the sampled document are not independent individuals but a combination set. Eq 6 can then be estimated by rejection sampling (Deng et al., 2018) or a Laplace approximation (Srivastava and Sutton, 2017) so as to avoid the independence assumption about the posterior distribution. Nonetheless, this will further increase the training complexity, which is already a key bottleneck for training the NR model.

6 Conclusions

We review standard WS signals used for training NR models in ODQA and provide a structured way of classifying them according to the required resource. For WS signal, we discuss different options and summarize the pros and cons. As a final wrap-up, we list promising directions that we believe worth exploring further: (1) *How to select the most suitable technique for a given scenario?* Despite the wide range of applicable techniques, it is non-trivial to decide how to select the best one except for an empirical experimentation. (2) *To which extent are these techniques complementary?* Existing work compares performance only between similar types of methods but not across the whole range of techniques and resources available. This makes it hard to decide whether different approaches could potentially complement each other and how they should be combined effectively. (3) *Do methods work across languages?* The vast majority of current research is conducted on English datasets. Even though all described methods in this survey have no explicit restrictions on languages they can be applied to, it is likely that their performance will vary across languages, especially for the methods relying on handcrafted heuristics.

Limitations

This survey covers introductions and related work of major WS algorithms used for neural ranking. Due to the space limit, most methods included in this paper are brief. Readers might not have a good understand on all the introduced methods. Interested readers can refer to existing surveys about general knowledge in QA (Zeng et al., 2020; Zhu et al., 2021a; Roy and Anand, 2021; Rogers et al., 2021; Pandya and Bhatt, 2021). Furthermore, we did not provide points to existing ODQA datasets and the performance of recent models. The conclusions in this survey also come from summaries of previous works. The lack of datasets including various resources needed for different WS algorithms prevents a comprehensive, fair comparison across algorithms. We hope future research can work on the creation of more datasets with various availabilitis of resources in different domains to enable this comparison. Lastly, we aim to create a big picture from the technology level, so we did not strictly limit our references only to the application of ODQA. The connection to specific ODQA applications might be loose, readers would need to extract useful information for the specific use cases.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

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. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174.

Wei-Cheng Chang, X Yu Felix, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

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

Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. 2022. Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. *arXiv preprint arXiv:2201.10582*.

Xilun Chen, Kushal Lakhotia, Barlas Oğuz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2021. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? *arXiv preprint arXiv:2110.06918*.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2022. Conversational question answering on heterogeneous sources. *arXiv preprint arXiv:2204.11677*.

R Dennis Cook and Sanford Weisberg. 1982. *Residuals and influence in regression*. New York: Chapman and Hall.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

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

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Marco Del Tredici, Gianni Barlacchi, Xiaoyu Shen, Weiwei Cheng, and Adriá de Gispert. 2021. Question rewriting for open-domain conversational qa: Best practices and limitations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2974–2978.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. *Advances in Neural Information Processing Systems*, 31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022. Pre-training transformer models with sentence-level objectives for answer sentence selection. *arXiv preprint arXiv:2205.10455*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32.

Pan Du, Jian-Yun Nie Nie, Yutao Zhu, Hao Jiang, Lixin Zou, and Xiaohui Yan. 2022. Pregan: Answer oriented passage ranking with weakly supervised gan. *arXiv preprint arXiv:2207.01762*.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1077–1083.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

James Ferguson, Hannaneh Hajishirzi, Pradeep Dasigi, and Tushar Khot. 2022. Retrieval data augmentation informed by downstream question answering performance. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 1–5, Dublin, Ireland. Association for Computational Linguistics.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2022. " what makes a question inquisitive?" a study on type-controlled inquisitive question generation. *arXiv preprint arXiv:2205.08056*.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, GP Shrivatsa Bhargav, Dinesh Garg, and Avirup Sil. 2020. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Christophe Van Gysel, Maarten De Rijke, and Evangelos Kanoulas. 2018. Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems (TOIS)*, 36(4):1–25.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Patrick Huber, Armen Aghajanyan, Barlas Oğuz, Dmytro Okhonko, Wen-tau Yih, Sonal Gupta, and Xilun Chen. 2021. Ccqa: A new web-scale question answering dataset for model pre-training. *arXiv preprint arXiv:2110.07731*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. *ICLR*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Jihyuk Kim, Minsoo Kim, and Seung-won Hwang. 2022. Collective relevance labeling for passage retrieval. *arXiv preprint arXiv:2205.03273*.

Devang Kulshreshtha, Robert Belfer, Iulian Vlad Serban, and Siva Reddy. 2021. Back-training excels self-training at unsupervised domain adaptation of question generation and passage retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7064–7078.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872.

Tuan Manh Lai, Trung Bui, Nedim Lipka, and Sheng Li. 2018. Supervised transfer learning for product information question answering. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1109–1114. IEEE.

Haejun Lee, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher D Manning, and Kyoung-Gu Woo. 2021. You only need one model for open-domain question answering. *arXiv preprint arXiv:2112.07381*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Jin Li, Peng Qi, and Hong Luo. 2022. Generating consistent and diverse qa pairs from contexts with bn conditional vae. In *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 944–949. IEEE.

Davis Liang, Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Embedding-based zero-shot retrieval through query generation. *arXiv preprint arXiv:2009.10270*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021a. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173.

Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043.

Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is more: Pretrain a strong siamese encoder for dense text retrieval using a weak decoder. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2791.

Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, et al. 2022. Erniesearch: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. *arXiv preprint arXiv:2205.09153*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction. *arXiv preprint arXiv:2204.10641*.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021a. Prop: Pre-training with representative words prediction for ad-hoc retrieval. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 283–291.

Zhengyi Ma, Zhicheng Dou, Wei Xu, Xinyu Zhang, Hao Jiang, Zhao Cao, and Ji-Rong Wen. 2021b. Pretraining for ad-hoc retrieval: Hyperlink is also you need. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1212–1221.

Sean MacAvaney, Kai Hui, and Andrew Yates. 2017. An approach for weakly-supervised deep information retrieval. *arXiv preprint arXiv:1707.00189*.

Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019. Content-based weak supervision for ad-hoc re-ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–996.

Yosi Mass and Haggai Roitman. 2020. Ad-hoc document retrieval using weak-supervision with bert and gpt2. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4191–4197.

Shikib Mehri and Mihail Eric. 2021. Example-driven intent prediction with observers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2979–2992.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34.

Thomas Minka. 2000. Estimating a dirichlet distribution.

Bhaskar Mitra, Nick Craswell, et al. 2018. *An introduction to neural information retrieval*. Now Foundations and Trends Boston, MA.

Iurii Mokrii, Leonid Boytsov, and Pavel Braslavski. 2021. A systematic evaluation of transfer learning and pseudo-labeling with bert-based ranking models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2081–2085.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems*, 33:21199–21212.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

Yifan Nie, Alessandro Sordoni, and Jian-Yun Nie. 2018. Multi-level abstraction convolutional model with weak supervision for information retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 985–988.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

Barlas Oğuz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen-tau Yih, Sonal Gupta, et al. 2021. Domain-matched pretraining tasks for dense retrieval. *arXiv preprint arXiv:2107.13602*.

Shivank Pandey and KC Rajeswari. 2013. Automatic question generation using software agents for technical institutions. *International Journal of Advanced Computer Research*, 3(4):307.

Hariom A Pandya and Brijesh S Bhatt. 2021. Question answering survey: Directions, challenges, datasets, evaluation matrices. *arXiv preprint arXiv:2112.03572*.

Ronak Pradeep, Xueguang Ma, Xinyu Zhang, Hang Cui, Ruizhou Xu, Rodrigo Nogueira, and Jimmy Lin. 2020. H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine. *Corpus*, 5(d3):d2.

Sheetal Rakangor and YR Ghodasara. 2015. Literature review of automatic question generation systems. *International journal of scientific and research publications*, 5(1):1–5.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3066–3079, Online. Association for Computational Linguistics.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2022. Learning to retrieve passages without supervision. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2687–2700, Seattle, United States.

Revanth Gangi Reddy, Vikas Yadav, Md Arafat Sultan, Martin Franz, Vittorio Castelli, Heng Ji, and Avirup Sil. 2021. Towards robust neural retrieval models with synthetic pre-training. *arXiv preprint arXiv:2104.07800*.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2022. A thorough examination on zero-shot dense retrieval. *arXiv preprint arXiv:2204.12755*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*, pages 232–241. Springer.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Guilherme Moraes Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. No parameter left behind: How distillation and model size affect zero-shot retrieval. *arXiv preprint*.

Rishiraj Saha Roy and Avishek Anand. 2021. Question answering for the curated web: Tasks and methods in qa over knowledge bases and text collections. *Synthesis Lectures onSynthesis Lectures on Information Concepts, Retrieval, and Services*, 13(4):1–194.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6648–6662.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. *arXiv preprint arXiv:2204.07496*.

Siamak Shakeri, Cicero dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460.

Xiaoyu Shen. 2022. Deep latent-variable models for text generation. *arXiv preprint arXiv:2203.02055*.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, Bill Byrne, and Adrià de Gispert. 2022a. Product answer generation from heterogeneous sources: A new benchmark and best practices. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 99–110.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, and Adrià de Gispert. 2022b. semipqa: A study on product question answering over semistructured data. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 111–120.

Xiaoyu Shen and Hui Su. 2018. Towards better variational encoder-decoders in seq2seq tasks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 8155–8156.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019a. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590.

Xiaoyu Shen, Svitlana Vakulenko, Marco Del Tredici, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. 2022c. Low-resource dense retrieval for open-domain question answering: A comprehensive survey. *arXiv preprint arXiv:2208.03197*.

Xiaoyu Shen, Yang Zhao, Hui Su, and Dietrich Klakow. 2019b. Improving latent alignment in text summarization by generalizing the pointer generator. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3762–3773.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2021. Can question generation debias question answering models? a case study on question–context lexical overlap. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 63–72.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. 2022. Questions are all you need to train a dense passage retriever. *arXiv e-prints*, pages arXiv–2206.

Ryan Smith, Jason A Fries, Braden Hancock, and Stephen H Bach. 2022. Language models in the loop: Incorporating prompting into weak supervision. *arXiv preprint arXiv:2205.02318*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *ICLR*.

Si Sun, Yingzhuo Qian, Zhenghao Liu, Chenyan Xiong, Kaitao Zhang, Jie Bao, Zhiyuan Liu, and Paul Bennett. 2021. Few-shot text ranking with meta adapted synthetic weak supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5030–5043.

Nandan Thakur, Nils Reimers, and Jimmy Lin. 2022. Domain adaptation for memory-efficient dense retrieval. *arXiv preprint arXiv:2205.11498*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William Cohen. 2021. Adaptable and interpretable neural memoryover symbolic knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3678–3691.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *NAACL*.

Liqiang Wang, Xiaoyu Shen, Gerard de Melo, and Gerhard Weikum. 2020. Cross-domain learning for classifying propaganda in online contents. *arXiv preprint arXiv:2011.06844*.

Wenya Wang and Sinno Pan. 2022. Deep inductive logic reasoning for multi-hop reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4999–5009.

Binxia Xu, Siyuan Qiu, Jie Zhang, Yafang Wang, Xiaoyu Shen, and Gerard de Melo. 2020. Data augmentation for multiclass utterance classification–a systematic study. In *Proceedings of the 28th international conference on computational linguistics*, pages 5494–5506.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2022. Laprador: Unsupervised pretrained dense retriever for zero-shot text retrieval. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3557–3569.

Peng Xu, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Passage ranking with weak supervision. *arXiv preprint arXiv:1905.05910*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021a. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 829–838.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2021b. Fine-tuning pretrained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1077.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25.

Xiang Yue, Xiaoman Pan, Wenlin Yao, Dian Yu, Dong Yu, and Jianshu Chen. 2022a. C-more: Pretraining to answer open-domain questions by consulting millions of references. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 371–377.

Xiang Yue, Ziyu Yao, and Huan Sun. 2022b. Synthetic question value estimation for domain adaptation of question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1340–1351.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences (2076-3417)*, 10(21).

Ziqian Zeng, Weimin Ni, Tianqing Fang, Xiang Li, Xinran Zhao, and Yangqiu Song. 2022. Weakly supervised text classification using supervision signals from a language model. *arXiv preprint arXiv:2205.06604*.

Jingtao Zhan, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. Evaluating extrapolation performance of dense retrieval. *arXiv preprint arXiv:2204.11447*.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. 2021. Wrench: A comprehensive benchmark for weak supervision. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kaitao Zhang, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2020. Selective weak supervision for neural information retrieval. In *Proceedings of The Web Conference 2020*, pages 474–485.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2495–2509.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. 2021. Distantly-supervised dense retrieval enables open-domain question answering without evidence annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9622, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Zhao, Xiaoyu Shen, Wei Bi, and Akiko Aizawa. 2019. Unsupervised rewriter for multi-sentence compression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2235–2240.

Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Lan Luo, Ke Zhan, Enrui Hu, Xinyu Zhang, Hao Jiang, Zhao Cao, Fan Yu, et al. 2022. Hyperlink-induced pretraining for passage retrieval in open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7135–7146.

Dawei Zhu, Xiaoyu Shen, Michael A Hedderich, and Dietrich Klakow. 2022. Meta self-refinement for robust learning with weak supervision. *arXiv preprint arXiv:2205.07290*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021a. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint*.

Yutao Zhu, Jian-Yun Nie, Zhicheng Dou, Zhengyi Ma, Xinyu Zhang, Pan Du, Xiaochen Zuo, and Hao Jiang. 2021b. Contrastive learning of user behavior sequence for context-aware document ranking. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2780–2791.

Shengyao Zhuang and Guido Zuccon. 2022. Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos. *arXiv preprint arXiv:2204.00716*.