Distantly-Supervised Evidence Retrieval Enables Question Answering without Evidence Annotation

Chen Zhao
CS, UMIACS
University of Maryland
chenz@cs.umd.edu

Chenyan Xiong
Microsoft Research
chenyan.xiong@microsoft.com

Jordan Boyd-Graber
CS, UMIACS, iSchool, LSC
University of Maryland
jbg@umiacs.umd.edu

Hal Daumé III
CS, UMIACS, LSC
University of Maryland
Microsoft Research
me@hal3.name

Abstract

Open-domain question answering answers a question based on evidence retrieved from a large corpus. State-of-the-art neural approaches require intermediate evidence annotations for training. However, such intermediate annotations are expensive, and methods that rely on them cannot transfer to the more common setting, where only question–answer pairs are available. This paper investigates whether models can learn to find evidence from a large corpus, with only distant supervision from answer labels for model training, thereby generating no additional annotation cost. We introduce a novel approach (DISTDR) that iteratively improves over a weak retriever by alternately finding evidence from the up-to-date model and encouraging the model to learn the most likely evidence. Without using any evidence labels, DISTDR is on par with fully-supervised state-of-the-art methods on both multi-hop and single-hop QA benchmarks. Our analysis confirms that DISTDR finds more accurate evidence over iterations, which leads to model improvements. The code is available at https://github.com/henryzhao5852/DistDR.

1 Introduction

Open-domain question answering (ODQA) takes a question, retrieves evidence from a large corpus, and finds an answer based on that evidence (Voorhees et al., 1999). With the help of large scale datasets, state-of-the-art approaches to QA (Karpukhin et al., 2020; Zhao et al., 2021, *inter alia*) can answer both simple questions that require only a single evidence piece (i.e., one passage); and more challenging multi-hop questions: computers must jump or "hop" from passage to passage (we call these passages **evidence pieces**), building a reasoning chain to find the answer.

Question: What state does ***Sang-Wook Cheong*** work as a ***materials scientist***?
Answer: New Jersey
Annotated Evidence: Sang-Wook Cheong -> Rutgers University
Alternate Evidence: Sang-Wook Cheong -> History of Rutgers University

Evidence Piece 1: Sang-Wook Cheong
Sang-Wook Cheong is a Korean American ***materials scientist*** at Rutgers University.

Evidence Piece 2: Rutgers University
Rutgers, The State University of New Jersey, commonly referred to as Rutgers University, Rutgers, or RU, is an American public research university and the largest institution for higher education in ***New Jersey***.

Evidence Piece 3: History of Rutgers University
Rutgers University is an institution of higher learning with campuses across the State of ***New Jersey***. Its main flagship campus locates in New Brunswick and Piscataway, ***New Jersey***.

Figure 1: A multi-hop question example from HOTPOTQA that requires finding multiple evidence pieces to form a reasoning chain (Sang-Wook Cheong → Rutgers University). Red: Text that overlaps between question and evidence piece; Blue: Span that matches the answer. State-of-the-art systems use evidence labels for training, but acquiring labeled evidence pieces is expensive.

State-of-the-art (SOTA) methods, however, are trained with *all* of the intermediate evidence pieces (e.g., in Figure 1, the evidence pieces for Sang-Wook Cheong's workplace which point you to Rutgers University's location) needed for the answer. Creating such intricate training data is expensive. For example, Kwiatkowski et al. (2019) use additional experts to justify the correctness of annotated evidence. The annotation protocol is even more nuanced for multi-hop questions. For example, Yang et al. (2018) ask annotators to write multi-hop questions based on two linked Wikipedia passages as a pre-defined reasoning chain, which creates dataset artifacts (Min et al., 2019b). While plenty of question-answer pairs are available without evidence labels, we cannot directly train SOTA models on such data.

Our work focuses on training ODQA systems

without these expensive annotations (Section 2): we only start with a question-answer pair. With that starting point, we use distant supervision to infer which evidence helps us get to the answer. The technical challenge is how to find these evidence from millions of candidates. Previous methods (Joshi et al., 2017; Cheng et al., 2020) use term matching (e.g., TF-IDF) for evidence retrieval, but their goal is a single piece of evidence: linking a question to a passage. As shown in Figure 1, the key to finding some evidence pieces does not appear in the question: for example, you only know to figure out that Rutgers University is in New Jersey after learning where Professor Cheong works. Fortunately, navigating to an answer given a question from a search engine is not impossible: humans do it every day, building on their existing knowledge toward the answer (Russell, 2019). With each round of searching to find additional clues, human users accumulate information to find the right answer. This paper creates a computational approach (DISTDR) that can use similar techniques to find the evidence needed to answer a given question.

DISTDR starts with a weak retriever then iteratively improves it by finding more useful evidence (Section 3). Specifically, we model evidence as a latent variable and develop a hard-EM algorithm that alternates between using the up-to-date retriever to find evidence (hard E-step) and updating the model parameters to further encourage the most useful evidence in the next iteration (M-step).

To implement this idea, we need a trainable retrieval system. We use dense retrieval (Lee et al., 2019) as our retriever, which uses a neural network to encode the evidence pieces we collect at each round into query vectors. DISTDR provides iterative feedback within the context of a QA system to guide the encoder to better find evidence pieces.

We evaluate DISTDR on ODQA benchmarks, including both single-hop questions (Kwiatkowski et al., 2019, NATURALQUESTIONS), where the evidence is the target passage, and multi-hop questions (Yang et al., 2018, HOTPOTQA), where the evidence is a chain of passages. *Without* using any annotated evidence labels, DISTDR's accuracy, according to several measures, is *on par* with fully-supervised state-of-the-art approaches on both benchmarks (Section 4).

Our analyses confirm the intuition that over iterations, DISTDR selects more accurate evidence, which in turn improves the model. Although some retrieved evidence from DISTDR does not match the annotation, it gives useful training signal, as DISTDR finds *alternative* evidence—another advantage of an automated approach (Section 5). For example, you can connect Sang-Wook Cheong to New Jersey through another Wikipedia page (History of Rutgers University in Figure 1).

2 Why Weakly-Supervised ODQA

Our task is to answer questions over large textual corpora. Our approach is generally applicable to both single-hop and multi-hop questions. We use Wikipedia as the knowledge source, but we do not use the metadata such as hyperlinks, to ensure our method can apply to any corpus (e.g., ClueWeb).

State-of-the-art approaches on ODQA mainly focus on the *fully-supervised* setting, where the question, answer, and evidence are given in training. Figure 1 shows an example multi-hop question with answer and annotated evidence. The fully-supervised setting simplifies model training but has three major challenges: (1) Annotation is costly: the question–answer pair is widely available in the real world, but the evidence (Sang-Wook Cheong → Rutgers University in Figure 1) requires human annotation, which is expensive to get, especially for complex questions; (2) Domain generalization: the labeled evidence is only on a single corpus, generalization to other corpora (e.g., moving from Wikipedia to medical records) is non-trivial; (3) Alternative evidence: there are often multiple correct evidence candidates in the corpora (e.g., in Figure 1, both Sang-Wook Cheong → Rutgers University and Sang-Wook Cheong → History of Rutgers University are correct), but only one of them is annotated as "gold evidence". Therefore, we explore the more common but challenging *weakly-supervised* setting which only requires question–answer input pairs.

Formally, given a question q and a textual corpus with D passages, our task is to first find a small subset of relevant passages as evidence z to the answer, where each evidence combines evidence pieces $z = z_1, ..., z_n$ with length n ($n = 1$ for single-hop questions), then find a span a from evidence z as an answer. We focus on the weakly-supervised setting, where training examples consist only of question–answer pairs (q, a), we gather evidence $\hat{z}$ from corpus as the distant supervision signal, based on the assumption that the presence of the answer a (source of distant supervision) in an

evidence piece implies that the evidence is needed to answer the question. This is contrast to the fully-supervised setting, where the gold evidence z^* is also given during training.

3 Weakly-Supervised ODQA with DISTDR

We present DISTDR, a unified framework for weakly-supervised ODQA. DISTDR is trained by retrieving evidence from a large corpus with distant supervision.

DISTDR follows the retriever–reader framework for ODQA, using dense retrieval to find evidence (Section 3.1). However, in our approach, we only have questions q and answers a, so we need to induce evidence $\hat{z}$ for training our retriever. We expect our initial retriever will struggle to find evidence. However, if it can find *some* useful evidence, we can encourage it to follow the clues to more evidence that can answer questions. This intuition is the foundation for our iterative approach (Section 3.2) for evidence retrieval: an initial retriever attempts to find evidence (Dense Retrieval). If the evidence contains the answer (Answer Matching), we label it as positive evidence (otherwise it's negative evidence); then we use the retrieved evidence as labels to retrain the retriever and reader (Model Update). While this idea forms the basis of our algorithm, it can be led astray by false positives: evidence that contains the answer but is irrelevant to question. Returning to our running example, while "named after a lawn ornament store in Wayne, New Jersey" has the state where Professor Cheong works, it is irrelevant to condensed matter physics. Thus, a reader filters spurious evidence (Reader Filter) to keep DISTDR on target.

3.1 Preliminary: Fully-Supervised ODQA

This section reviews state-of-the-art systems for fully-supervised ODQA, where a dense retriever finds evidence from a large corpus, and a reader—multi-tasked with evidence reranking and span extraction—outputs a span as the answer.

Dense Retrieval Dense retrieval (Lee et al., 2019) is based on a dual-encoder architecture, which uses BERT to represent both the query q and the passage p with dense vectors. The model learns a scoring function (e.g., dot product) between question and passage vectors:

$$f(q, p) = \text{sim}(\text{Enc}_Q(q), \text{Enc}_P(p)). \quad (1)$$

These models are highly scalable, since passages can be encoded offline, and are efficiently retrieved over maximum inner product search (MIPS) with the query (Shrivastava and Li, 2014).

Multi-step Evidence Retrieval We apply multiple dense retrieval steps to find evidence with a sequence of evidence pieces (each piece is a passage) $z_1, \ldots, z_n$, After each retrieval step, we create the new query by concatenating retrieved evidence pieces to the original question. Specifically, at retrieval step t, we form a new query q_t by appending our already retrieved evidence pieces to the original question

$$q_t = [q; z_1; \ldots; z_{t-1}], \quad (2)$$

and retrieve a new evidence piece z_t. During inference, a beam search finds the top-k evidence, where the score is the product of individual evidence pieces' score. For training, given positive evidence $z^+ \equiv z_1^+, \ldots, z_n^+$ (In Figure 1, z_1^+: Sang-Wook Cheong, z_2^+: Rutgers University) and a set of negative evidence Z^-, we use negative log likelihood loss (NLL) over each step:

$$L(q, z^+, Z^-) = \quad (3)$$
$$\sum_t \frac{\exp^{f([q;z_{t-1}^+], z_t^+)}}{\exp^{f([q;z_{t-1}^+], z_t^+)} + \sum_{z^- \in Z^-} \exp^{f([q;z_{t-1}^-], z_t^-)}}.$$

Evidence Reranking and Span Extraction After the retriever, QA systems need a reader, which combines evidence selection (reranking) and span extraction. Unlike retrieval, the reader encodes pairwise information between question and evidence, thus giving a more accurate (but slower) prediction. Following Karpukhin et al. (2020), we use a BERT encoder with input formatted as `[CLS] question [SEP] title`$_1$ `[SEP] evi`$_1$ `[SEP] ... title`$_n$ `[SEP] evi`$_n$`[SEP]`, where `[CLS]` and `[SEP]` are special tokens, and each `evi`$_t$ is an evidence piece z_t. First, we use `[CLS]` token's representation $\mathbf{U}_{\texttt{[CLS]}}$ to estimate the probability that the collected evidence z contains the answer:

$$P(z \mid q) = \text{softmax}(\mathbf{U}_{\texttt{[CLS]}}^\intercal \mathbf{w}_{\text{rank}}), \quad (4)$$

where $\mathbf{w}_{\text{rank}}$ is a weight vector. Then answers are predicted with a span start and end classifier:

$$P(\text{start} \mid q, z) = \text{softmax}(\mathbf{U}\mathbf{w}_{\text{start}}), \quad (5)$$
$$P(\text{end} \mid q, z) = \text{softmax}(\mathbf{U}\mathbf{w}_{\text{end}}), \quad (6)$$

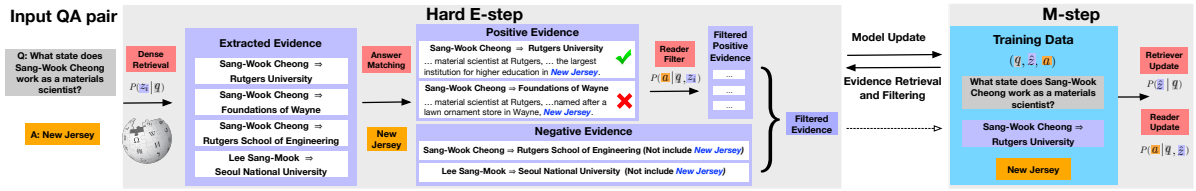


Figure 2: Our DISTDR model, with question: q, evidence: z, answer: a, model component: Dense Retrieval. ✔indicates reader output matches the correct answer, thus the positive evidence is kept (otherwise filtered out, presented by ✗). Left: at E-step, DISTDR finds the most relevant evidence using the current dense retriever on the training examples, uses both answer string matching and reader filter to form positive and negative evidence. Right: at M-step, DISTDR updates both retriever and reader components using the training data from E-step as distant supervision.

where $\mathbf{w}_{\text{start}}$ and $\mathbf{w}_{\text{end}}$ are weights. From the highest-scored evidence $\hat{z}$, we select the answer with the highest span probability $P(\text{start} \mid \hat{z}, q) \times P(\text{end} \mid \hat{z}, q)$. The training objective is the log-likelihood of the positive evidence for reranking and maximum marginal likelihood over all spans in the positive evidence for span extraction.

3.2 Learning DISTDR: Hard-EM

This section introduces training DISTDR with weak supervision. We treat the evidence z as a latent variable (Figure 2) with a multinomial distribution (parameterized by probability of selecting correct evidence from the retriever) over solution set $Z = D \times n$, where D is the corpus size and n is the number of retrieval steps. Formally, given the question q and answer a, our goal is to maximize the probability of finding correct evidence $P(z \mid q)$ (For single-hop questions, z is retrieved with single step; for multi-hop questions, it takes multiple steps to find z: after the first retrieval step, the query q is not just the original question but also appended evidence as in Equation 2.), and selecting an answer from the question $P(a \mid q, z)$. We use expectation maximization (EM) to infer the latent variable z. We compute the likelihood of each z given q, yielding a vector of estimates $\hat{z}$ (E-step); then update the model parameters based on $\hat{z}$ (M-step). Since it's intractable to enumerate all evidence candidates to compute the expectation, we adopt hard-EM (Samdani et al., 2012) and approximate the E-step by picking the most likely solution as $\hat{z}$. We pass over all questions in the training set and repeat this process for multiple iterations until the model converges.

Hard E-step At the hard E-step, for each question q in the training set we find the most likely estimated evidence $\hat{z}$. This is implemented by multiple retrieval steps for multi-hop questions, specifically at each step t:

$$\hat{z}_t = \arg\max_{\hat{z} \in Z} P(\hat{z}_t \mid q, \hat{z}_1, ..., \hat{z}_{t-1}); \qquad (7)$$

and single-step retrieval for single-hop questions. We use an up-to-date retriever to find the top-k evidence from corpus (with beam search for the k-highest scoring chains, Section 3.1).

Given the retrieval output, we will eventually need to retrain the retriever. This requires knowing which evidence is useful and which is not. As a proxy, we look for the *answer* to split the top-k candidate evidence into positive (has the answer) $\hat{Z}^+$ and negative $\hat{Z}^-$ evidence (lacks the answer) sets. And as a by-product, we generate the most challenging negative evidence at each iteration, which makes training more robust (Guu et al., 2020).

Evidence Filter Although using the answer to filter evidence ensures that the positive evidence contains the answer, it does not always mean this evidence is relevant. For example, the answer to "Who played in the most world series MLB games" is New York Yankees, but "New York Yankees is an American professional baseball team" is not the *correct* evidence. This issue is more pronounced at the beginning of the process when the retriever is weaker. To mitigate this issue, the reader filters spurious positives: if it does not believe New York Yankees is the answer to the question, the evidence is not usable. Specifically, for each evidence $\hat{z}^+$ in the positive evidence set, the current reader model outputs the most likely answer $\hat{a}$. We only keep positive evidence if $\hat{a}$ matches the correct answer.

M-step Now that we have our estimated evidence $\hat{z}$, including both the highest scored positive evidence $\hat{z}^+$ (after filtering), which we assume is a true solution, and $\hat{Z}^-$ as negative evidence set. We

have $(q, \hat{z}, a)$ for each training example to update both the retriever $P(\hat{z} \mid q)$ and reader $P(a \mid q, \hat{z})$ (Section 3.1).

4 Experiments

In this section, we evaluate DISTDR on both multi-hop and single-hop QA benchmarks. DISTDR is generally applicable to both questions by adopting different evidence retrieval steps.

4.1 Datasets

We evaluate on two datasets (statistics in Table 1), HOTPOTQA and NATURALQUESTIONS.

HOTPOTQA (Yang et al., 2018) is a multi-hop QA benchmark, where intermediate hops have been annotated by hand. We focus on the full wiki setting, where the corpus is first passage of all Wikipedia pages (5.23 million passages). We do not use its supporting facts (evidence) annotation in our setting. We only use its bridge questions subset, which are designed to be multi-hop. HOTPOTQA also includes comparison questions that compare properties of two question entities, but its yes/no answers are beyond the scope of this paper, as we cannot get distant supervision from them.

NATURALQUESTIONS (Kwiatkowski et al., 2019) is a QA benchmark, which mainly includes single-hop questions. Besides questions and answers, NATURALQUESTIONS also annotates passages as evidence, but we do not use it in the weakly-supervised setting. We follow Karpukhin et al. (2020) and use all of Wikipedia as a corpus, split into 100-token chunks (21 million passages).

4.2 Evaluation Metrics

On HOTPOTQA, we evaluate the retrieval component on ten evidence (chains), where each sequence has two passages. For retrieval, we follow Zhao et al. (2021) and report **answer recall** (the fraction of questions with the answer string in the retrieved passages), **passage recall** (if at least one gold passage is in the retrieved passages), and **chain recall** (if both gold passages are included in the retrieved passages) on the dev set. For the reader, we first use the same metrics as above on the top ten chains reranked from top-100 retrieval results for reranking. Then we report an exact match (EM) score on answer spans. On NATURALQUESTIONS, we report answer recall on top-k passages ($k = 1, 20$) from the retriever, and exact match (EM) on answer spans on test set, following Karpukhin et al. (2020).

Dataset	Train	Dev	Test
HOTPOTQA	72,424	5,918	-
NATURALQUESTIONS	79,168	8,757	3,610

Table 1: Number of questions on HOTPOTQA and NATURALQUESTIONS. We use the dev (sub)set to evaluate HOTPOTQA, since the test set is hidden and we only use its bridge questions.

4.3 Compared Methods

On HOTPOTQA retrieval, we compare DISTDR with unsupervised TF-IDF, and two recent state-of-the-art multi-step dense retrieval methods—BEAMDR (Zhao et al., 2021) and MDR (Xiong et al., 2021b)—under full supervision. For the reader, we first compare DISTDR with BEAMDR and two other top leaderboard entries, Transformer-XH (Zhao et al., 2020) and GRR (Asai et al., 2020), both of which use Wikipedia hyperlinks to find candidates. For fair comparison, all approaches use BERT-base as pre-trained model, and we use the released model checkpoints to do inference on bridge question subsets.[1]

On NATURALQUESTIONS, we compare DISTDR with two state-of-the-art dense retrieval methods, DPR (Karpukhin et al., 2020) and ANCE (Xiong et al., 2021a), which use same model architecture as DPR, with asynchronous negative evidence updates during training. We evaluate on fully-supervised and weakly-supervised settings. We directly use the released model checkpoint on full supervision and train models from published code and data[2] on distant supervision. All approaches use BERT-base as pre-trained model.

4.4 Implementation details

On HOTPOTQA, we initialize our retriever with a dense retrieval checkpoint from NATURALQUESTIONS for both hops.[3] To initialize the reader, we first run DISTDR's retrieval for one iteration (with-

[1] MDR uses RoBERTa-base for retrieval, and BERT/ELECTRA-large for reranking and span extraction. We include the retrieval results (though it gives slight gains) but do not compare the reader. We expect the reader results are close to BEAMDR, as both use similar models.

[2] The released data uses answers to match top BM-25 results for distant supervision.

[3] We make this choice because we need a cold starting point for our EM process (i.e., at the first iteration, find some training signals from the retrieved evidence). The initialization on HOTPOTQA does not use any evidence labels (which would be cheating in our setting). There are potentially alternative approaches to initializations that could be considered for future work.

Models	Ans	Passage	Chain
No Supervision			
TF-IDF	60.7	90.8	35.8
Full Supervision			
MDR	85.2	89.3	75.3
BEAMDR	84.9	91.1	73.6
Distant Supervision			
DISTDR w/ TF-IDF chains	55.6	52.3	22.3
DISTDR	86.2	92.2	75.1

Table 2: Comparison of DISTDR's retrieval (based on top-10 chains) with unsupervised and fully-supervised methods on HOTPOTQA dev set over answer, passage and chain recall. DISTDR matches fully-supervised dense retrieval approaches.

Models	Ans	Passage	Chain	Span
Full Supervision				
Transformer-XH	91.8	97.0	81.3	52.4
GRR	87.5	92.2	79.0	50.4
BEAMDR	90.5	94.7	83.0	52.6
Distant Supervision				
DISTDR	91.4	95.3	81.7	51.6

Table 3: Comparison of DISTDR with other fully-supervised methods with reranking over answer, passage and chain recall, and span extraction over span exact match. DISTDR is competitive to BEAMDR.

out the reader filter), and train the reader from scratch, using the top-50 retrieval outputs.[4] We use the same hyper-parameters as BEAMDR, and train DISTDR for eight iterations.[5] On NATURALQUESTIONS, we initialize DISTDR from a DPR checkpoint under distant supervision (so the evidence label is not used). We train DISTDR for ten iterations, using the same hyper-parameters as DPR (Karpukhin et al., 2020). We run DISTDR on eight 2080Ti GPUs, and training takes three days.

4.5 Main Results

Table 2 presents retrieval results on HOTPOTQA. Here, TF-IDF is only able to find one evidence piece (usually the first hop which overlaps with question), but fails to find all the evidence pieces. Hence, using evidence from TF-IDF as distant supervision cannot effectively train DISTDR. DISTDR is slightly better than BEAMDR and MDR—both are state-of-the-art dense retrieval methods with

[4]We follow the reader setup from Karpukhin et al. (2020): if multiple positives are in top-50 outputs, one of them is sampled as the positive instance in each training iteration. Our pilot experiment shows that reader model is robust: such initialization gives reasonable accuracy.

[5]DISTDR converges after five iterations (Figure 3).

Models	Top 1	Top 20	Span
Full Supervision			
DPR	46.3	78.4	41.5
ANCE	50.9	81.9	46.0
Distant Supervision			
DPR	45.2	78.2	37.9
ANCE	45.7	79.1	38.3
DISTDR	50.4	80.1	40.5

Table 4: Comparison of DISTDR with other dense retrieval systems on NATURALQUESTIONS dataset over answer recall @1, 20, and span exact match. DISTDR is competitive to fully-supervised approaches.

Models	Ans	Passage	Chain
DISTDR	86.2	92.2	75.1
DISTDR w/ Sampled Pos	83.9	90.9	72.1
Remove Reader Filter	85.4	91.9	72.1
Remove Non-gold Evidence	66.7	74.1	53.1

Table 5: Ablations evaluated on answer, passage, and chain recall. Removing all but the gold evidence hurts most, while the other components of DISTDR all contribute to finding the answer.

full supervision—even though DISTDR is only trained on (question, answer) pairs. When using the same model implementation but with distant supervision, DISTDR is competitive to BEAMDR on reader results (Table 3). On NATURALQUESTIONS (Table 4), unlike multi-hop questions, using IR to find evidence as distant supervision provides helpful training signals, which is confirmed by small gap between distant and full supervision.[6] Building on top of distantly-supervised DPR, DISTDR beats weakly-supervised systems, and is competitive with fully-supervised models.

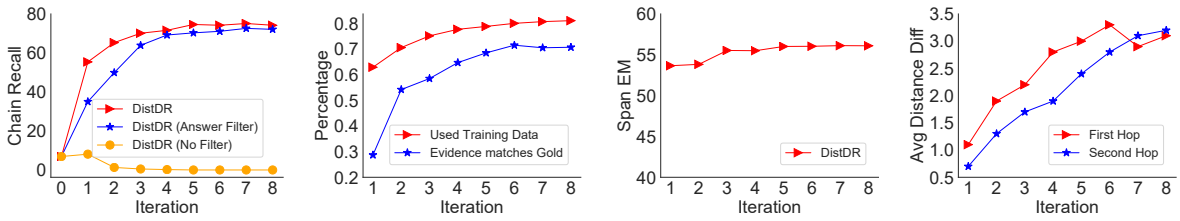
5 Analysis

This section explores HOTPOTQA results to understand why DISTDR is on par with fully-supervised approaches.

5.1 Analysis on Model Training

We first study how DISTDR finds better evidence from a large corpus with hard EM.

[6]Compared to retrieval, the gap is larger for the reader. This is due to training data processing for DPR (and ANCE). On retrieval, DPR replaces annotated gold passage with the corresponding 100-token passage in the candidate pool, and discard the questions if the matching is failed (25% of questions). On reader, DPR uses annotated passage as positive, and top retrieved passages that do not contain the answer are the negatives, thus entire training data is used.



(a) (b) (c) (d)

Figure 3: Quantitive analysis on DISTDR by iteration. (a): Compare different evidence filter strategies on dev set; (b): Statistics on training set extracted evidence; (c): Compare span extraction component over gold evidence; (d): Average distance difference on dev set from question to top-10 negative passages (Average) and positive passage. DISTDR finds better evidence over iterations, the improved evidence further helps model training.

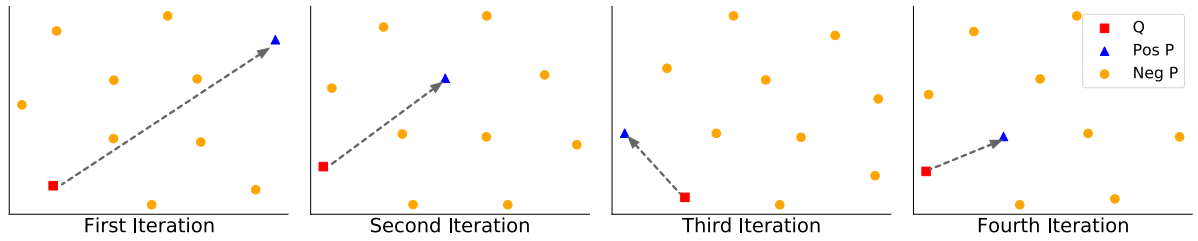


Figure 4: T-SNE visualization on an example with query (Q) and top-10 scored (first hop) passage (P) embeddings on first four iterations. DISTDR closes the distance between query and positive passage with more accurate evidence over iterations as distant supervision.

Effect of Hard EM We study the effect of hard EM by examining accuracy and extracted evidence statistics at every iteration. Chain recall for DISTDR on the dev set increases over the first five iterations (Figure 3a) and then converges. As the evidence finder improves, the E-step extracts better evidence for training examples (Figure 3b): both the percentage of used training examples and gold evidence in training examples increase with additional EM iterations. The improved evidence further helps DISTDR training. We compare multiple positive sampling strategies: using hard EM (top-1 as positive evidence) is slightly better than randomly sampling from top-k positives for the M-step update (Table 5, second row).

The Effect of Evidence Filtering Filtering positive evidence is crucial for high-quality evidence. We ablate different filtering methods in Table 5 and Figure 3a: No filter, only an answer filter, and an additional reader filter (DISTDR). No filter fails completely: evidence should not be considered correct if it does not contain the answer. With an additional reader filter, DISTDR beats using only the answer matching filter and converges faster (since the reader filter reduces false-positives, this makes training more robust).

Despite rapidly changing evidence, the reader's span extraction is robust over iterations (Figure 3c), even on the first iteration, suggesting it is robust against spurious false-positive evidence.

Visualizing DISTDR in Dense Space DISTDR improves evidence retrieval by bringing the question representation closer to positive evidence and pushing negative evidence away (Figure 3d). In Figure 4, T-SNE (Maaten and Hinton, 2008) visualizes the first hop query and passage representations (second hop representations are similar) in dense space at each iteration. As expected, at the beginning, the question is far from the positive passage, with negative passages between. After updating the representations, the distance gets closer until positive passage is the closest to the question.

5.2 Analysis on Retrieved Evidence

DISTDR is competitive to fully-supervised approaches on HOTPOTQA. However, even at the last iteration, only 65% of extracted evidence matches the gold. We ablate evidence and contrasts gold evidence with that retrieved by DISTDR.

Non-gold Evidence is Helpful If we only retain training examples that match labeled gold evidence, answer accuracy falters below DISTDR (Table 5, last row). Instead of providing noise, non-gold evidence improves model training.

Q: Jo Ann Terry won the 80m hurdles event at what Sao Paulo-based event from 1963?
A: Pan American Games
Human Annotation: Jo Ann Terry → 1963 pan american games **DISTDR**: Jo Ann Terry → Jovem Pan
Jo Ann Terry: Jo Ann Terry-Grissom is a retired female hurdler from the United States. Affiliated with the Tennessee State University she won the 80 m hurdles event at the 1963 Pan American Games. **1963 Pan American Games**: The 4th Pan American Games were held from April 20 to May 5, 1963, in São Paulo. **Jovem Pan**: Jovem Pan is the main Brazilian radio station based in São Paulo, Brazil.
Q: Who's the Hungarian-born US film director renowned for adapting Stephen King novellas to the screen, including The Mist and The Green Mile?
A: Frank Darabont
Human Annotation: The Mist (film) → Frank Darabont **DISTDR**: The Green Mile (film) → Frank Darabont
The Mist (film): The Mist is a 2007 American science-fiction horror film based on the 1980 novella "The Mist" by Stephen King. The film was written and directed by Frank Darabont. **The Green Mile (film)**: The Green Mile is a 1999 American fantasy crime drama film written and directed by Frank Darabont and adapted from the 1996 Stephen King novel of the same name. **Frank Darabont**: Frank Arpad Darabont is a French-Hungarian-American film director, screenwriter and producer. As a director he is known for his film adaptations of Stephen King novellas such as "The Green Mile" and "The Mist".
Q: Zimbabwe's Guwe Secondary School has a sister school in what New York county?
A: Nassau County
Human Annotation: Guwe Secondary School → Carle Place High School **DISTDR**: University of Zimbabwe → East Rockaway High School
Guwe Secondary School: Guwe Secondary School is located in Zimbabwe. It has a sister school in Carle Place, New York, United States. **Carle Place High School**: Carle Place Middle/High School is a six-year comprehensive public high school located in the hamlet of Carle Place in Nassau County, New York. **University of Zimbabwe**: The University of Zimbabwe (UZ) is the oldest and formerly largest university in Zimbabwe. **East Rockaway High School**: East Rockaway Junior-Senior High School is a co-educational six-year secondary school in East Rockaway, New York, and the sole high school in Nassau County, New York.

Table 6: Case study of DISTDR on "false positive" evidence from HOTPOTQA that does not match gold evidence. Despite the dataset goals, some examples are answerable with a single piece of evidence evidence (top). Other times, DISTDR finds alternate valid evidence (middle). However, it can also find incorrect evidence (bottom).

Human Analysis and Case Study We manually annotate fifty training examples where extracted DISTDR evidence differs from labeled evidence. We confirm that most extracted evidence is helpful. Specifically, 38% of the questions are answerable via a single piece of evidence, even though this dataset is supposed to require multiple hops (Table 6). In another 28% of cases, DISTDR finds alternate valid evidence. In the second example in Table 6, the question mentions both films The Mist and The Green Mile; therefore the reasoning chain from either film to the director Frank Darabont is correct (though only one is annotated). In the final of 34% of cases, DISTDR finds the wrong evidence, often because the extracted evidence only includes one span that matches the answer type, therefore the reader confidently outputs the span for the wrong reason. In the third example in Table 6, Nassau County is the only county it sees, and therefore the model has stumbled upon the right answer erroneously. Building a model with faithful predictions is an important ongoing research topic (Jacovi and Goldberg, 2020).

6 Related Work

Question Answering Datasets There is growing interest in NLP communities to build large-scale datasets (Rajpurkar et al., 2016; Jia and Liang, 2017; Dua et al., 2019, *inter alia*) for QA research. In addition to questions and answers, benchmark datasets often include annotated evidence, but it requires significant annotation protocol design and human annotations. In SQUAD (Rajpurkar et al., 2016), among the first large-scale reading comprehension datasets, annotators write questions conditioned on a passage, which creates dataset artifacts (Jia and Liang, 2017). To overcome dataset artifacts, NATURALQUESTIONS (Kwiatkowski et al., 2019) use real Google queries as questions and ask annotators to label both evidence passages and short answers. But such annotation is expensive, as they further ask additional experts to verify whether the evidence correctly leads to the answer. Instead, DISTDR focuses on the weakly-supervised setting with only question–answer pairs, which is significantly cheaper.

Annotation is more fraught for multi-hop QA

datasets (Yang et al., 2018). To construct HotpotQA, annotators are presented with a linked Wikipedia passages, as pilot studies indicate that it is difficult to ask a meaningful multi-hop question with arbitrary passages. However, some questions in HotpotQA include shortcuts that are answerable by a single passage (Min et al., 2019b), which is confirmed by our analysis (Section 5.2).

Distant supervision has been successfully adopted for many NLP tasks such as relation extraction (Mintz et al., 2009). Recent work builds QA datasets with distant supervision, such as TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), QBLink (Elgohary et al., 2018) by automatically gathering evidence documents from a corpus as distant supervision for available question-answer pairs. They use standard IR techniques to find relevant passages and match them with answer strings. These methods succeed for simple questions where terms overlap with evidence passages, This no longer holds for multi-hop questions that require a reasoning chain as evidence to the answer: evidence pieces do not overlap with the question but rather depend on the previous evidence pieces. Unsupervised IR methods cannot capture such implicit relations. DistDR removes the burden of little textual overlap through dense retrieval and its iterative process retrieves better evidence.

Open-domain QA systems Chen et al. (2017) first combine information retrieval and (neural) reading comprehension for open-domain QA. Several works aim to improve the neural reader (Clark and Gardner, 2018; Wang et al., 2018, *inter alia*), or use generative models to compose an answer (Lewis et al., 2020; Izacard and Grave, 2021, *inter alia*). Recent progress (Karpukhin et al., 2020; Xiong et al., 2021a,b, *inter alia*) uses dense retrieval to aid both single-hop and multi-hop questions. However, a crucial distinction is that these approaches assume the evidence is given for training, while DistDR iteratively finds evidence and uses it to improve the model. Min et al. (2019a) also use hard-EM for weakly-supervised QA, but—orthogonal to our approach—they assume the evidence is given and find the most likely answer mentions in the evidence, while we aim to find evidence from a large corpus.

7 Conclusion

We present DistDR, a distantly-supervised ODQA system that improves over a weak retriever by iteratively finding evidence from a corpus, and using the evidence as distant supervision for model training. Without using any evidence labels, DistDR matches the fully-supervised SOTA approaches on both multi-hop and single-hop QA benchmarks.

Annotating evidence for existing question-answer pairs is generally expensive, especially for complex questions. While DistDR can accurately find evidence for arbitrary complex machine reading-style questions, future work needs to validate whether this can work for other types of questions. This could improve the reader to answer numerical reasoning (Dua et al., 2019), temporal reasoning (Ning et al., 2020), multi-model reasoning (Lei et al., 2018), or combination of these skills (Bartolo et al., 2020).

Acknowledgments

We thank CLIP members, Tianze Shi, anonymous reviewers and meta-reviewer for their suggestions and comments. Zhao is supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the BETTER Program contract 2019-19051600005. Boyd-Graber is supported by NSF Grant IIS-1822494. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over Wikipedia graph for question answering. In *Proceedings of the International Conference on Learning Representations*.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the Association for Computational Linguistics*.

Hao Cheng, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2020. Probabilistic assumptions matter: Improved models for distantly-supervised document-level question answering. In *Proceedings of the Association for Computational Linguistics*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the Association for Computational Linguistics*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

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. Dataset and baselines for sequential open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the International Conference of Machine Learning*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the Association for Computational Linguistics*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of Empirical Methods in Natural Language Processing*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*.

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 Empirical Methods in Natural Language Processing*.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the Association for Computational Linguistics*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

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. In *Proceedings of Advances in Neural Information Processing Systems*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. A discrete hard EM approach for weakly supervised question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019b. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the Association for Computational Linguistics*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Association for Computational Linguistics*.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. Torque: A reading comprehension dataset of temporal ordering questions. In *Proceedings of Empirical Methods in Natural Language Processing*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Daniel M Russell. 2019. *The Joy of Search: A Google Insider's Guide to Going Beyond the Basics*. MIT Press.

Rajhans Samdani, Ming-Wei Chang, and Dan Roth. 2012. Unified expectation maximization. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Proceedings of Advances in Neural Information Processing Systems*.

Ellen M Voorhees et al. 1999. The TREC-8 question answering track report. In *Proceedings of Text Retrieval Conference*.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2018. R^3: Reinforced reader-ranker for open-domain question answering. In *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021a. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the International Conference on Learning Representations*.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2021b. Answering complex open-domain questions with multi-hop dense retrieval. In *Proceedings of the International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. 2021. Multi-step reasoning over unstructured text with beam dense retrieval. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *Proceedings of the International Conference on Learning Representations*.