A Discrete Hard EM Approach for Weakly Supervised Question Answering

Sewon Min\textsuperscript{1}, Danqi Chen\textsuperscript{2,3}, Hannaneh Hajishirzi\textsuperscript{1,4}, Luke Zettlemoyer\textsuperscript{1,3}
\textsuperscript{1}University of Washington, Seattle, WA
\textsuperscript{2}Princeton University, Princeton, NJ
\textsuperscript{3}Facebook AI Research, Seattle, WA
\textsuperscript{4}Allen Institute for Artificial Intelligence, Seattle, WA
\{sewon,hannaneh,lsz\}@cs.washington.edu danqic@cs.princeton.edu

Abstract

Many question answering (QA) tasks only provide weak supervision for how the answer should be computed. For example, TRIVIAQA answers are entities that can be mentioned multiple times in supporting documents, while DROP answers can be computed by deriving many different equations from numbers in the reference text. In this paper, we show it is possible to convert such tasks into discrete latent variable learning problems with a precomputed, task-specific set of possible solutions (e.g. different mentions or equations) that contains one correct option. We then develop a hard EM learning scheme that computes gradients relative to the most likely solution at each update. Despite its simplicity, we show that this approach significantly outperforms previous methods on six QA tasks, including absolute gains of 2–10%, and achieves the state-of-the-art on five of them. Using hard updates instead of maximizing marginal likelihood is key to these results as it encourages the model to find the one correct answer, which we show through detailed qualitative analysis.\footnote{Our code is publicly available at https://github.com/shmsw25/qa-hard-em.}

1 Introduction

A natural setting in many question answering (QA) tasks is to provide weak supervision to determine how the question should be answered given the evidence text. For example, as seen in Figure 1, TRIVIAQA answers are entities that can be mentioned multiple times in supporting documents, while DROP answers can be computed by deriving many different equations from numbers in the reference text. Such weak supervision is attractive because it is relatively easy to gather, allowing for large datasets, but complicates learning because there are many different spurious ways to derive the correct answer. It is natural to model such ambiguities with a latent variable during learning, but most prior work on reading comprehension has rather focused on the model architecture and used heuristics to map the weak signal to full supervision (e.g. by selecting the first answer span in TRIVIAQA (Joshi et al., 2017; Tay et al., 2018; Talmor and Berant, 2019)). Some models are trained with maximum marginal likelihood (MML) (Kadlec et al., 2016; Swayamdipta et al., 2018; Clark and Gardner, 2018; Lee et al., 2019), but it is unclear if it gives a meaningful improvement over the heuristics.

In this paper, we show it is possible to formulate a wide range of weakly supervised QA tasks as discrete latent-variable learning problems. First, we define a solution to be a particular derivation of a model to predict the answer (e.g. a span in

Figure 1: Examples from two different question answering tasks. (Top) Multi-mention reading comprehension. The answer text is mentioned five times in the given document, however, only the fourth span actually answers the question. (Bottom) Reading comprehension with discrete reasoning. There are many potential equations which execute the answer (‘4’), but only one of them is the correct equation (‘40-36’) and the others are false positives.

1 Our code is publicly available at https://github.com/shmsw25/qa-hard-em.
the document or an equation to compute the answer). We demonstrate that for many recently introduced tasks, which we group into three categories as given in Table 1, it is relatively easy to precompute a discrete, task-specific set of possible solutions that contains the correct solution along with a modest number of spurious options. The learning challenge is then to determine which solution in the set is the correct one, while estimating a complete QA model.

We model the set of possible solutions as a discrete latent variable, and develop a learning strategy that uses hard-EM-style parameter updates. This algorithm repeatedly (i) predicts the most likely solution according to the current model from the precomputed set, and (ii) updates the model parameters to further encourage its own prediction. Intuitively, these hard updates more strongly enforce our prior beliefs that there is a single correct solution. This method can be applied to any problem that fits our weak supervision assumptions and can be used with any model architecture.

We experiment on six different datasets (Table 1) using strong task-specific model architectures (Devlin et al., 2019; Dua et al., 2019; Hwang et al., 2019). Our learning approach significantly outperforms previous methods which use heuristic supervision and MML updates, including absolute gains of 2–10%, and achieves the state-of-the-art on five datasets. It outperforms recent state-of-the-art reward-based semantic parsing algorithms (Liang et al., 2018; Agarwal et al., 2019) by 13% absolute percentage on WikiSQL, strongly suggesting that having a small precomputed set of possible solutions is a key ingredient. Finally, we present a detailed analysis showing that, in practice, the introduction of hard updates encourages models to assign much higher probability to the correct solution.

### 2 Related Work

**Reading Comprehension.** Large-scale reading comprehension (RC) tasks that provide full supervision for answer spans (Rajpurkar et al., 2016) have seen significant progress recently (Seo et al., 2017; Xiong et al., 2018; Yu et al., 2018a; Devlin et al., 2019). More recently, the community has moved towards more challenging tasks such as distantly supervised RC (Joshi et al., 2017), RC with free-form human generated answers (Kočiský et al., 2018) and RC requiring discrete or multi-hop reasoning (Dua et al., 2019; Yang et al., 2018). These tasks introduce new learning challenges since the gold solution that is required to answer the question (e.g. a span or an equation) is not given.

Nevertheless, not much work has been done for this particular learning challenge. Most work on RC focuses on the model architecture and simply chooses the first span or a random span from the document (Joshi et al., 2017; Tay et al., 2018; Talmor and Berant, 2019), rather than modeling this uncertainty as a latent choice. Others maximize the sum of the likelihood of multiple spans (Kadlec et al., 2016; Swayamdipta et al., 2018; Clark and Gardner, 2018; Lee et al., 2019), but it is unclear if it gives a meaningful improvement. In this paper, we highlight the learning challenge and show that our learning method, independent of the model architecture, can give a significant gain. Specifically, we assume that one of

---

**Table 1:** Six QA datasets in three different categories used in this paper (detailed in Section 5) along with the size of each dataset. An average and median of the size of precomputed solution sets (denoted by $Z$) are also reported. Details on how to obtain $Z$ are given in Section 4.

<table>
<thead>
<tr>
<th>Task &amp; Dataset</th>
<th># Examples</th>
<th>Avg</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Multi-mention reading comprehension</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRIVIAQA (Joshi et al., 2017)</td>
<td>61,888</td>
<td>7,993</td>
<td>7,701</td>
</tr>
<tr>
<td>NARRATIVEQA (Kočiský et al., 2018)</td>
<td>32,747</td>
<td>3,461</td>
<td>10,557</td>
</tr>
<tr>
<td>TRIVIAQA-OPEN (Joshi et al., 2017)</td>
<td>78,785</td>
<td>8,837</td>
<td>11,313</td>
</tr>
<tr>
<td>NATURALQUESTIONS-OPEN (Kwiatkowski et al., 2019)</td>
<td>79,168</td>
<td>8,757</td>
<td>3,610</td>
</tr>
<tr>
<td><strong>2. Reading comprehension with discrete reasoning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DROPnum (Dua et al., 2019)</td>
<td>46,973</td>
<td>5,850</td>
<td>-</td>
</tr>
<tr>
<td><strong>3. Semantic Parsing</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WikiSQL (Zhong et al., 2017)</td>
<td>56,355</td>
<td>8,421</td>
<td>15,878</td>
</tr>
</tbody>
</table>
mentions are related to the question and others are false positives because (i) this happens for most cases, as the first example in Table 2, and (ii) even in the case where multiple mentions contribute to the answer, there is often a single span which fits the question the best.

**Semantic Parsing.** Latent-variable learning has been extensively studied in the literature of semantic parsing (Zettlemoyer and Collins, 2005; Clarke et al., 2010; Liang et al., 2013; Berant et al., 2013; Artzi and Zettlemoyer, 2013). For example, a question and an answer pair \((x, y)\) is given but the logical form that is used to compute the answer is not. Two common learning paradigms are maximum marginal likelihood (MML) and reward-based methods. In MML, the objective maximizes

\[
\sum_{z \in \hat{Z}} P(z|x),
\]

where \(\hat{Z}\) is an approximation of a set of logical forms executing \(y\) (Liang et al., 2013; Berant et al., 2013; Krishnamurthy et al., 2017). In reward-based methods, a reward function is defined as a prior, and the model parameters are updated with respect to it (Iyyer et al., 2017; Liang et al., 2017, 2018). Since it is computationally expensive to obtain a precomputed set in semantic parsing, these methods typically recompute the set of logical forms with respect to the beam at every parameter update. In contrast, our learning method targets tasks that a set of solutions can be precomputed, which include many recent QA tasks such as reading comprehension, open-domain QA and a recent SQL-based semantic parsing task (Zhong et al., 2017).

### 3 Method

In this section, we first formally define our general setup, which we will instantiate for specific tasks in Section 4 and then we describe our learning approach.

#### 3.1 Setup

Let \(x\) be the input of a QA system (e.g. a question and a document) and \(y\) be the answer text (e.g. ‘Robert Schumann’ or ‘4’). We define a solution as a particular derivation that a model is supposed to produce for the answer prediction (e.g. a span in the document or an equation to compute the answer, see Table 2). Let \(f\) denote a task-specific, deterministic function which maps a solution to the textual form of the answer (e.g. by simply returning the string associated with a particular selected mention or solving an equation to get the final number, see Table 2). Our goal is to learn a model (with parameters \(\theta\)) which takes an input \(x\) and outputs a solution \(z\) such that \(f(z) = y\).

In a fully supervised scenario, a true solution \(\bar{z}\) is given, and \(\theta\) is estimated based on a collection of \((x, \bar{z})\) pairs. In this work, we focus on a weakly supervised setting in which \(\bar{z}\) is not given and we define \(Z_{tot}\) as a finite set of all the possible solutions. In the case that the search space is very large or infinite, we usually can approximate \(Z_{tot}\) with a high coverage in practice. Then, we obtain \(Z = \{z \in Z_{tot} : f(z) = y\}\) by enumerating all \(z \in Z_{tot}\). This results a set of all the possible solutions that lead to the correct answer. We assume it contains one solution that we want to learn to produce, and potentially many other spurious ones. In practice, \(Z\) is defined in a task-specific manner, as we will see in Section 4.

At inference time, the model produces a solution \(\hat{z} \in Z_{tot}\) from an input \(x\) with respect to \(\theta\) and predicts the final answer as \(f(\hat{z})\). Note that we cannot compute \(Z\) at inference time because the groundtruth \(y\) is not given.\(^2\)

#### 3.2 Learning Method

In a fully-supervised setting where \(\bar{z}\) is given, we can learn \(\theta\) by optimizing the negative log likelihood of \(\bar{z}\) given the input \(x\) with respect to \(\theta\).

\[
J_{\text{Sup}}(\theta|x, \bar{z}) = -\log P(\bar{z}|x; \theta)
\]

In our weak supervision scenario, the model has access to \(x\) and \(Z = \{z_1, z_2, \ldots, z_n\}\), and the selection of the best solution in \(Z\) can be modeled as a latent variable. We can compute the maximum marginal likelihood (MML) estimate, which marginalizes the likelihood of each \(z_i \in Z\) given the input \(x\) with respect to \(\theta\). Formally,

\[
P(y|x; \theta) = \sum_{z_i \in Z_{tot}} P(y|z_i)P(z_i|x; \theta)
\]

is used to compute the objective as follows:

\[
J_{\text{MML}}(\theta|x, Z) = -\log \sum_{z_i \in Z} P(z_i|x; \theta)
\]

However, there are two major problems in the MML objective in our settings. First, MML can be

\(^2\) This is a main difference from multi-instance learning (MIL) (Zhou et al., 2009), since a bag of input instances is given at inference time in MIL.
maximized by assigning high probabilities to any subset of \( z_i \in Z \); whereas in our problems, instances in \( Z \) other than one correct \( z \) are spurious solutions which the model should ideally assign very low probability. Second, in MML we optimize the sum over probabilities of \( Z \) during training but typically predict the maximum probability solution during inference, creating a discrepancy between training and testing.

We introduce a learning strategy with a hard-EM approach. First, the model computes the likelihood of each \( z_i \) given the input \( x \) with respect to \( \theta \), \( \mathbb{P}(z_i|x; \theta) \), and picks one of \( Z \) with the largest likelihood:

\[
\hat{z} = \arg\max_{z_i \in Z} \mathbb{P}(z_i|x; \theta)
\]

Then, the model optimizes on a standard negative log likelihood objective, assuming \( \hat{z} \) is a true solution. The objective can be re-written as follows:

\[
J_{\text{Hard}}(\theta|x, Z) = -\log \mathbb{P}(\hat{z}|x; \theta) = -\log \max_{z_i \in Z} \mathbb{P}(z_i|x; \theta) = -\max_{z_i \in Z} \log \mathbb{P}(z_i|x; \theta) = \min_{z_i \in Z} J_{\text{Sup}}(\theta|x, z_i)
\]

This objective can be seen as a variant of MML, where the sum is replaced with a max.

4 Task Setup

We apply our approach to three different types of QA tasks: multi-mention reading comprehension, RC with discrete reasoning and a semantic parsing task. In this section, we describe each task in detail: how we define a solution \( z \) and pre-compute a set \( Z \) based on input \( x \) and answer \( y \). The statistics of \( |Z| \) and examples on each task are shown in Table 1 and Table 2 respectively.

4.1 Multi-Mention Reading Comprehension

Multi-mention reading comprehension naturally occurs in several QA tasks such as (i) distantly-supervised reading comprehension where a question and answer are collected first before the evidence document is gathered (e.g. TRIVIAQA), (ii) abstractive reading comprehension which requires a free-form text to answer the question (e.g. NARRATIVEQA), and (iii) open-domain QA where only question-answer pairs are provided.

Given a question \( Q = [q_1, \ldots, q_l] \) and a document \( D = [d_1, \ldots, d_L] \), where \( q_i \) and \( d_j \) denote the tokens in the question and document, the output \( y \) is an answer text, which is usually mentioned multiple times in the document.

Previous work has dealt with this setting by detecting spans in the document through text matching (Joshi et al., 2017; Clark and Gardner, 2018). Following previous approaches, we define a solution \( z \) as a span in the document. We obtain a set of possible solutions \( Z = \{z_1, \ldots, z_n\} \) by finding exact match or similar mentions of \( y \), where \( z_i = (s_i, e_i) \) is a span of text with start and end token indices \( s_i \) and \( e_i \). Specifically,

\[
g_{\text{max}} = \max_{1 \leq s_i \leq e_i \leq L} g([d_{s_i}, \ldots, d_{e_i}], y) = \{z_i = (s_i, e_i) \text{ s.t. } g(s_i, e_i) = g_{\text{max}}\}
\]

where \( g \) is a string matching function. If the answer is guaranteed to be a span in the document \( D \), \( g \) is a binary function which returns 1 if two strings are the same, and 0 otherwise. If the answer is free-form text, we choose \( g \) as the ROUGE-L metric (Lin, 2004).

This complicates the learning because the given document contains many spans matching to the text, while most of them are not related to answering the question. As an example shown in Table 2, only the fourth span out of six is relevant to the question.

4.2 Reading Comprehension with Discrete Reasoning

Some reading comprehension tasks require reasoning in several discrete steps by finding clues from the document and aggregating them. One such example is mathematical reasoning, where the model must pick numbers from a document and compute an answer through arithmetic operations (Dua et al., 2019).

In this task, the input is also a question \( Q \) and a document \( D \), and the output \( y \) is given as a numeric value. We define a solution \( z \) to be an executable arithmetic equation. Since there is an infinite set of potential equations, we approximate \( Z_{\text{tot}} \) as a finite set of arithmetic equations with two numeric values and one operation, following Dua et al. (2019).

Specifically,

\[
Z_{\text{tot}} = \{z_i = (o_1, n_1, o_2, n_2) \text{ s.t. } o_1, o_2 \in \{+, -, \%\}, n_1, n_2 \in N_D \cup N_Q \cup S\}
\]

This approximation covers 93% of the examples in the development set.
1. Multi-Mention Reading Comprehension (TRIVIAQA, NARRATIVEQA, TRIVIAQA-OPEN & NATURALQUESTIONS-OPEN)

**Question:** Which composer did pianist Clara Wieck marry in 1840?

**Document:** Robert Schumann was a German composer and influential music critic. He is widely regarded as one of the greatest composers of the Romantic era. (...) Robert Schumann himself refers to it as “an afflication of the whole hand”. (...) Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket”. (...) Clara Schumann was a German musician and composer, considered one of the most distinguished pianists of the Romantic era. Her husband was the composer Robert Schumann. (...) At the age of eight, the young Clara Wieck performed at the Leipzig home of Dr. Ernst Carus. There she met another gifted young pianist who had been invited to the musical evening, named Robert Schumann, who was nine years older. Schumann admired Clara’s playing so much that he asked permission from her mother to discontinue his law studies. (...) In the spring of 1853, the then unknown 20-year-old Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.

**Answer:** Robert Schumann

2. Reading Comprehension with Discrete Reasoning (DROP$_{mult}$)

**Question:** How many yards longer was Bob Bironas’ longest field goal compared to John Carney’s only field goal?

**Document:** (...) The Chiefs tied the game with QB Brodie Croyle completing a 10 yard td pass to WR Samrie Parker. Afterwards the Titans responded with kicker Rob Bironas managing to get a 37 yard field goal. Kansas city would take the lead prior to halftime with croyle completing a 9 yard td pass to FB Kris Wilson. In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. The Chiefs answered with kicker John Carney getting a 36 yard field goal. Afterwards the Titans would retake the lead with Young and Williams hooking up with each other again on a 41 yard td pass. (...) Tennessee clinched the victory with Bironas nailing a 40 yard and a 25 yard field goal. With the win the Titans kept their playoff hopes alive at 8-6.

**Answer:** 4

3. SQL Query Generation (WIKI) SQL

**Question:** What player played guard for Toronto in 1996-1997?

**Table Header:** player, year, position, ...

**Answer:** John Long

Table 2: Examples of the input, answer text \(y\), \(f\), \(Z_{\text{tot}}\) and \(Z\). First, in multi-mention reading comprehension, the answer text ‘Robert Schumann’ is mentioned six times but only the fourth span is related to the question. Second, in reading comprehension with discrete reasoning, many equations yield to the answer 4, but only ‘40-37’ answers the question. Lastly, in SQL query generation, five SQL queries lead to the answer but only the first one is the correct query. See Section 4 for more details.

\[
Z = \{ z_i \in Z_{\text{tot}} \text{ s.t. } f(z_i) = y \}
\]

where \(f\) is an execution function of equations. Figure 1 shows an example \(Z\) given a question and a document. We can see that one equation is correct, while the others are false positives which coincidentally lead to the correct answer.

4.3 SQL Query Generation

To evaluate if our training strategy generalizes to other weak supervision problems, we also study a semantic parsing task where a question and an answer are given but the logical form to execute the answer is not. In particular, we consider a task of answering questions about a given table by generating SQL queries.

The input is a question \(Q = [q_1, \ldots, q_l]\) and a table header \(H = [h_1, \ldots, h_n]\), where \(q_i\) is a token, \(h_i\) is a multi-token title of each column, and \(n_L\) is the number of headers. The supervision is given as the SQL query result \(y\), which is always a text string.

We define a solution to be an SQL query. Since the set of potential queries is infinite, we approximate \(Z_{\text{tot}}\) as a set of non-nested SQL queries with at most three conditions. Specifically, given \(A\) as a set of aggregating operators \{sum, mean, max, min, count\} and \(C\) as a set of possible conditions \{(h, o, t) s.t. \(h \in [1, n_L]\), \(o \in \{=, \lt, >\}\), \(t \in \text{spans in } Q\}\}, we define \(Z_{\text{tot}}\):

\[
Z_{\text{tot}} = \{ z_i = (z_i^{\text{sel}}; z_i^{\text{agg}}; \{ z_{ij}^{\text{cond}} \}_{j=1}^3 ) \text{ s.t. } z_i^{\text{sel}} \in [1, n_L], z_i^{\text{agg}} \in \{\text{none}\} \cup A, z_{ij}^{\text{cond}} \in \{\text{none}\} \cup C \text{ for } j \in [1, 3] \}
\]

then,

\[
Z = \{ z_i \in Z_{\text{tot}} \text{ s.t. } f(z_i) = y \}
\]

This approximation covers 99% of the examples in the development set.
where \( f \) is an SQL executor. The third example in Table 2 shows \( Z \) may contain many spurious SQL queries, e.g. the third query in \( Z \) coincidentally executes the answer because ‘John Long’ is ranked first among all the guards in alphabetical order.

5 Experiments

We experiment on a range of question answering tasks with varied model architectures to demonstrate the effectiveness of our approach. Built on top of strong base models, our learning method is able to achieve state-of-the-art on NARRATIVEQA, TRIVIAQA-open, NnaturalQuestions-open, DROP_num and WikiSQL.

5.1 Multi-mention Reading Comprehension

We experiment on two reading comprehension datasets and two open-domain QA datasets. For reading comprehension, we evaluate on TRIVIAQA (Wikipedia) (Joshi et al., 2017) and NARRATIVEQA (summary) (Kočiský et al., 2018).

For open-domain QA, we follow the settings in Lee et al. (2019) and use the QA pairs from TRIVIAQA-unfiltered (Joshi et al., 2017) and NnaturalQuestions (Kwiatkowski et al., 2019) with short answers and discard the given evidence documents. We refer to these two datasets as TRIVIAQA-open and NnaturalQuestions-open.

We experiment with three learning methods as follows.

- First Only: \( J(\theta) = -\log P(z_1|x; \theta) \), where \( z_1 \) appears first in the given document among all \( z_i \in Z \).
- MML: \( J(\theta) = -\log \sum_{i=1}^{n} P(z_i|x; \theta) \).
- Ours: \( J(\theta) = -\log \max_{1 \leq i \leq n} P(z_i|x; \theta) \).

\( P(z_i|Q, D) \) can be obtained by any model which outputs the start and end positions of the input document. In this work, we use a modified version of BERT (Devlin et al., 2019) for multi-paragraph reading comprehension (Min et al., 2019).

Training details. We use uncased version of BERT_base. For all datasets, we split documents into a set of segments up to 300 tokens because BERT limits the size of the input. We use batch size of 20 for two reading comprehension tasks and 192 for two open-domain QA tasks. Following Clark and Gardner (2018), we filter a subset of segments in TRIVIAQA through TF-IDF similarity between a segment and a question to maintain a reasonable length. For open-domain QA tasks, we retrieve 50 Wikipedia articles through TF-IDF (Chen et al., 2017) and further run BM25 (Robertson et al., 2009) to retrieve 20 (for train) or 80 (for development and test) paragraphs. We try 10, 20, 40 and 80 paragraphs on the development set to choose the number of paragraphs to use on the test set.

To avoid local optima, we perform annealing: at training step \( t \), the model optimizes on MML objective with a probability of \( \min(t/\tau, 1) \) and otherwise use our objective, where \( \tau \) is a hyperparameter. We observe that the performance is improved by annealing while not being overly sensitive to the hyperparameter \( \tau \). We include full hyperparameters and detailed ablations in Appendix B.

Results. Table 3 compares the results of baselines, our method and the state-of-the-art on four datasets. First of all, we observe that First-Only is a strong baseline across all the datasets. We hypothesize that this is due to the bias in the dataset that answers are likely to appear earlier in the paragraph. Second, while MML achieves comparable result to the First-Only baseline, our learning method outperforms others by 2+ F1/ROUGE-L/EM consistently on all datasets. Lastly, our method achieves the new state-of-the-art on NARRATIVEQA, TRIVIAQA-open and NnaturalQuestions-open, and is comparable to the state-of-the-art on TRIVIAQA, despite our aggressive truncation of documents.

5.2 Reading Comprehension with Discrete Reasoning

We experiment on a subset of DROP (Dua et al., 2019) with numeric answers (67% of the entire dataset) focusing on mathematical reasoning. We refer to this subset as DROP_num. The current state-of-the-art model is an augmented version of QANET (Yu et al., 2018a) which selects two numeric values from the document or the question and performs addition or subtraction to get the an-

\(^5\) Following Lee et al. (2019), we treat the dev set as the test set and split the train set into 90/10 for training and development. Datasets and their split can be downloaded from https://bit.ly/2HK1Fqn.
Table 3: Results on multi-mention reading comprehension & discrete reasoning tasks. We report performance on five datasets with different base models. Note that we are not able to obtain the test result on the subset DROPnum. Previous state-of-the-art are from Wang et al. (2018), Nishida et al. (2019), Lee et al. (2019), Lee et al. (2019) and Dua et al. (2019), respectively. Our training method consistently outperforms the First-Only and MML by a large margin in all the scenarios.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Only</td>
<td>64.4</td>
<td>64.9</td>
<td>55.3</td>
</tr>
<tr>
<td>MML</td>
<td>64.8</td>
<td>65.5</td>
<td>55.8</td>
</tr>
<tr>
<td>Ours</td>
<td>66.9</td>
<td>67.1</td>
<td>58.1</td>
</tr>
<tr>
<td>SOTA</td>
<td>-</td>
<td>71.4</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4: Results on WikiSQL. We compare accuracy with weakly-supervised or fully-supervised settings. Our method outperforms previous weakly-supervised methods and most of published fully-supervised methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>REINFORCE</td>
<td>&lt; 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iterative ML</td>
<td>70.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard EM</td>
<td>70.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beam-based MML</td>
<td>70.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPO</td>
<td>71.8</td>
<td>72.4</td>
<td></td>
</tr>
<tr>
<td>MAPO+McrL</td>
<td>74.9</td>
<td>74.8</td>
<td></td>
</tr>
<tr>
<td>MML</td>
<td>70.6</td>
<td>70.5</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>84.4</td>
<td>83.9</td>
<td></td>
</tr>
</tbody>
</table>

5.3 SQL Query Generation

Finally, we experiment on the weakly-supervised setting of WikiSQL (Zhong et al., 2017), in which only the question & answer pair is used and the SQL query \(z_i\) is treated as a latent variable. \(\mathbb{P}(z_i|Q,H)\) can be computed by any query generation or semantic parsing models. We choose SQLova (Hwang et al., 2019), a competitive model on WikiSQL (designed for fully supervised setting), as our base model. We modify the model to incorporate either the MML objective or our hard-EM learning approach for the weakly-supervised setting.

We compare with both traditional and recently-developed reward-based algorithms for weak supervision, including beam-based MML (MML which keeps a beam during training), conventional hard EM\(^8\), REINFORCE (Williams, 1992), iterative ML (Liang et al., 2017; Abolafia et al., 2018) and a family of MAPO (Memory-augmented policy optimization) (Liang et al., 2018; Agarwal et al., 2019). For a fair comparison, we only consider single models without execution-guided decoding.

Training details. We truncate the document to be up to 400 words. We use the batch size of 14 and 10 for QANet and BERT, respectively.

Results. Table 3 shows the results on DROPnum. Our training strategy outperforms the First-Only baseline and MML by a large margin, consistently across two base models. In particular, with BERT, we achieve an absolute gain of 10%.

\(^7\)As we use a different set of operations for the two models, they are not directly comparable. Details of the model architecture are shown in Appendix A.

\(^8\)This method differs from ours in that it does not have a precomputed set, and uses a beam of candidate predictions to execute for each update.
Results. Table 4 shows that our training method significantly outperforms all the weakly-supervised learning algorithms, including 10% gain over the previous state of the art. These results indicate that precomputing a solution set and training a model through hard updates play a significant role to the performance. Given that our method does not require SQL executions at training time (unlike MAPO), it provides a simpler, more effective and time-efficient strategy. Comparing to previous models with full supervision, our results are still on par and outperform most of the published results.

6 Analysis

In this section, we will conduct thorough analyses and ablation studies, to better understand how our model learns to find a solution from a precomputed set of possible solutions. We also provide more examples and analyses in Appendix C.

Varying the size of solution set at inference time. Figure 2 shows a breakdown of the model accuracy with respect to the size of a solution set ($|Z|$) at test time. We observe that the model with our training method outperforms the model with MML objective consistently across different values of $|Z|$. The gap between MML and our method is marginal when $|Z| = 0$ or 1, and gradually increases as $|Z|$ grows.

Varying the size of solution set at training. To see how our learning method works with respect to the size of a solution set ($|Z|$) of the training data, particularly with large $|Z|$, we take 5 subsets of the training set on WikiSQL with $|Z| = 3, 10, 30, 100, 300$. We train a model with those subsets and evaluate it on the original development set, both with our training method and MML objective. Figure 3 shows statistics of each subset and results. We observe that (i) our learning method outperforms MML consistently over different values of $|Z|$, and (ii) the gain is particularly large when $|Z| > 3$.

Model predictions over training. We analyze the top 1 prediction and the likelihood of $z \in Z$ assigned by the model on DROP$_{num}$ with different number of training iterations (steps from 1k to 32k). Table 5 shows one example on DROP$_{num}$ with the answer text ‘4’, along with the model’s top 1 prediction and a subset of $Z$. We observe that
Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal? (Answer: 4)

P: ... The Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. The Chiefs answered with kicker John Carney getting a 36 yard field goal. The Titans would retake the lead with Young and Williams hooking up with each other again on a 41 yard td pass. ... Tennessee clinched the victory with Bironas nailing a 40 yard and a 25 yard field goal.

Table 5: An example from DROP$_{num}$ (same as Figure 1 and Table 2), with its answer text ‘4’ and a subset of the solution set (Z), containing two of ‘41-38’ (which ‘41’ come from different mentions; one denoted by ‡ for distinction), ‘40-36’ and ‘10-4’. For each training step t, the top 1 prediction and Z ordered by $P(z|x; \theta_t)$, a probability of $z \in Z$ with respect to the model at t through training procedure are shown. Note that at inference time Z is not given, so top 1 prediction is not necessarily an element of Z.

The model first begins by assigning a small, uniform probability distribution to Z, but gradually learns to favor the true solution. The model sometimes gives the wrong prediction—for example, at $t = 16k$, and changes its prediction from the true solution to the wrong solution, ‘37-36’—but again changes its prediction to be a true solution afterward. In addition, its intermediate wrong solution, ‘37-36’ indicates the model was confused with distinguishing the longest field goal of Rob Bironas (40 vs. 37), which is an understandable mistake.

We also compare the predictions from the model with our method to those from the model with MML, which is shown in Appendix C.

Quality of the predicted solution. We analyze if the model outputs the correct solution, since the solution executing the correct answer could be spurious. First, on NARRATIVEQA and DROP$_{num}$, we manually analyze 50 samples from the development set and find that 98% and 92% of correct cases produce the correct solution respectively. Next, on WIKI$SQL$, we compare the predictions from the model to the annotated SQL queries on the development set. This is possible because gold SQL queries are available in the dataset for the full supervision. Out of 8,421 examples, 7,110 predictions execute the correct answers. Among those, 88.5% of the predictions are exactly same as the annotated queries. Others are the cases where (i) both queries are correct, (ii) the model prediction is correct but the annotated query is incorrect, and (iii) the annotated query is correct and the model prediction is spurious. We show a full analysis in Appendix C.

Robustness to the noise in |Z|. Sometimes noise arises during the construction of |Z|, such as |Z| constructed based on ROUGE-L for NARRATIVEQA. To explore the effect of noise in Z, we experiment with more noisy solution set by picking all the spans with scores that is equal to or larger than the 5th highest. The new construction method increases |Z| from 4.3 to 7.1 on NARRATIVEQA. The result by MML objective drops significantly (56.07→51.14) while the result by ours drops marginally (58.77→57.97), suggesting that MML suffers more with a noisier Z while ours is more robust.

7 Conclusion

In this paper, we demonstrated that, for many QA tasks which only provide the answer text as supervision, it is possible to precompute a discrete set of possible solutions that contains one correct option. Then, we introduced a discrete latent variable learning algorithm which iterates a procedure of predicting the most likely solution in the precomputed set and further increasing the likelihood of that solution. We showed that this approach significantly outperforms previous approaches on six QA tasks including reading comprehension, open-domain QA, discrete reasoning task and semantic parsing, achieving absolute gains of 2–10% and setting the new state-of-the-art on five well-studied datasets.

Acknowledgements

The University of Washington research was supported by ONR (N00014-18-1-2826, N00014-17-S-B001), DARPA N66001-19-2-403, NSF (IIS-1616112, IIS-1252835, IIS-1562364), ARO (W911NF-16-1-0121), an Allen Distinguished Investigator Award, Samsung GRO and gifts from Allen Institute for AI, Google and Amazon.

The authors would like to thank the anonymous reviewers, Eunsol Choi, Christopher Clark, Victor Zhong and UW NLP members for their valuable feedback.
References


Rishabh Agarwal, Chen Liang, Dale Schuurmans, and Mohammad Norouzi. 2019. Learning to generalize from sparse and underspecified rewards. In ICML.


Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In ACL.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In ACL.

James Clarke, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. 2010. Driving semantic parsing from the world’s response. In CoNLL.


Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In ACL.


Rudolf Kadlec, Martin Schmid, Ondrej Baigár, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In ACL.


Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In EMNLP.


Chen Liang, Jonathan Berant, Quoc Le, Kenneth D Forbus, and Ni Lao. 2017. Neural symbolic machines: Learning semantic parsers on Freebase with weak supervision. In ACL.

Chen Liang, Mohammad Norouzi, Jonathan Berant, Quoc V Le, and Ni Lao. 2018. Memory augmented policy optimization for program synthesis and semantic parsing. In NIPS.


Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In EMNLP.


A Model details

We describe the detailed model architecture used as a base model. In other words, we describe how we obtain \( P(z|x; \theta) \).

The following paragraphs describe (i) BERT extractive model used for multi-mention RC (Section 5.1) and (ii) BERT sequence tagging model for discrete reasoning task (Section 5.2), respectively. For QANet for discrete reasoning and the model for SQL generation (Section 5.3), we use the open-sourced code of the original implementation\(^9\) of Dua et al. (2019) and Hwang et al. (2019) and do not make any modification except the objective function, so we refer to original papers.

All implementations are done in PyTorch (Paszke et al., 2017). For BERT, we modify the open-sourced implementation in PyTorch\(^10\) and use the uncased version of BERT\(_{base}\).

Extractive QA model for multi-mention RC

The model architecture is closed to that of Min et al. (2019) and Alberti et al. (2019), where the model operates independently on each paragraph, and selects the best matching paragraph and its associated answer span.

The input is a question \( Q \) and a set of paragraphs \( \{P_1, \ldots, P_N\} \), and the desired output is a span from one of paragraphs. Since our goal is to compute a probability of a specific span, \( z \), let’s say \( z \) is \( s \)-th through \( e \)-th word in \( k \)-th paragraph.

The model receives a question \( Q \) and a single paragraph \( P_i \) in parallel. Then, \( S_i = Q : [SEP] : P_i \), a list of \( m + n_i + 1 \) words, where \( : \) indicates a concatenation, [SEP] is a special token, \( m \) is the length of \( Q \), and \( n_i \) is the length of \( P_i \). This \( S_i \) is fed into BERT:

\[
\tilde{S}_i = BERT(S) \in \mathbb{R}^{h \times (m+n_i+1)}
\]

where \( h \) is the hidden dimension of BERT. Then,

\[
p_{i,\text{start}} = \text{Softmax}(\tilde{S}_i^T W_1) \in \mathbb{R}^{m+n_i+1}
\]

\[
p_{i,\text{end}} = \text{Softmax}(\tilde{S}_i^T W_2) \in \mathbb{R}^{m+n_i+1}
\]

where \( W_1, W_2 \in \mathbb{R}^h \) are learnable vectors.

Finally, the probability of \( z \), \( s \)-th through \( e \)-th word in \( i \)-th paragraph, is obtained by:

\[
P(z|Q, P_i) = p_{i,\text{start}}^d \times p_{i,\text{end}}^e
\]

where \( p^d \) denotes \( d \)-th element of the vector \( p \).

Separately, a paragraph selector is trained through \( p_{i,\text{exist}} = \text{Softmax}(W_3 \text{maxpool}(S_i)) \) \( \in \mathbb{R}^2 \) where \( W_3 \in \mathbb{R}^{h \times 2} \) is learnable vector. At inference time, \( k = \arg\max_{i} p_{i,\text{exist}} \) is computed and \( P(z|Q, P_k) \) is only considered to output a span.

Sequence Tagging model for discrete reasoning

The basic idea of the model is closed to that of Dua et al. (2019). The input is a question \( Q \) and a paragraph \( P \). Our goal is to compute a probability of an equation, \( z = (o_1, n_1, o_2, n_2) \), where \( o_1, o_2 \in \{+, -, +0.01\} \) and \( n_1, n_2 \in N_P \cup N_Q \cup S \), \( N_P \) and \( N_Q \) are all numeric values appearing in \( P \) and \( Q \), and \( S \) is a set of predefined special numbers.\(^11\)

First, BERT encodings of the question and the paragraph is obtained via

\[
\tilde{S} = BERT(Q : [SEP] : P) \in \mathbb{R}^{h \times (m+n+1)}
\]

where \( : \) indicates a concatenation, [SEP] is a special token, \( m \) is the length of \( Q \), \( n \) is the length of \( P \), and \( h \) is the hidden dimension of BERT. Then,

\[
p_{\text{input}} = \text{Softmax}(\tilde{S}^T W_1) \in \mathbb{R}^{(m+n+1) \times 4}
\]

\[
p_{\text{special}} = \text{Softmax}(\text{maxpool}(\tilde{S}) W_2) \in \mathbb{R}^{|S| \times 4}
\]

where \( W_1 \in \mathbb{R}^{h \times 4} \) and \( W_2 \in \mathbb{R}^{h \times |S| \times 4} \) are learnable matrices. Then,

\[
P(z|x) = \prod_{i=1}^{m+n+1} p_{\text{input}}^{g(i)} \prod_{j=1}^{|S|} p_{\text{special}}^{h(j)}
\]

where

\[
g(i) = \begin{cases} 
\alpha(o_1) & \text{if } [Q : [SEP] : P]^i = n_1 \\
\alpha(o_2) & \text{if } [Q : [SEP] : P]^i = n_2 \\
0 & \text{otherwise}
\end{cases}
\]

\[
h(j) = \begin{cases} 
\alpha(o_1) & \text{if } S^j = n_1 \\
\alpha(o_2) & \text{if } S^j = n_2 \\
0 & \text{otherwise}
\end{cases}
\]

\[
\alpha(o) = \begin{cases} 
1 & \text{if } o = '+' \\
2 & \text{if } o = '-' \\
3 & \text{if } o = '*0.01'$
\end{cases}
\]

Here, subscript \( i \) of the vector or the sequence indicate \( i \)-th dimension of the vector or \( i \)-th element of the sequence, respectively.


\(^{10}\)https://github.com/huggingface/pytorch-pretrained-BERT

\(^{11}\)\( S = \{1, 2, 3, 4, 5, 7, 10, 12, 100, 1000\} \)
Table 6: (Left) Ablations with varying values of \( \tau \) on TRIVIAQA. (Middle) Ablations with varying values of \( \tau \) on \( \text{DROP}_{\text{num}} \) with BERT. (Right) Final \( \tau \) chosen for the main results on each dataset. Note that for \( \text{DROP}_{\text{num}} \) with QANET and WIKISQL, we just report the number without annealing.

### B  Annealing

To prevent the model to be optimized on early decision of the model, we perform annealing: at training step \( t \), the model optimizes on MML objective with a probability of \( \min(t/\tau, 1) \) and otherwise use our objective, where \( \tau \) is a hyperparameter. We observe that the performance is improved by annealing while not being sensitive to the hyperparameter \( \tau \). Ablations and chosen \( \tau \) for each dataset are shown in Table 6. Note that for \( \text{DROP}_{\text{num}} \) with QANET and WIKISQL, we do not ablate with varying \( \tau \) and just report the number without annealing.

### C  Examples

To see if the prediction from the model is the correct solution to derive the answer, we analyze outputs from the model.

**TRIVIAQA.** Table 7 shows one example from TRIVIAQA where the answer text (Montgomery) is mentioned in the paragraph multiple times. Predictions from the model with our training method and that with MML objective are shown in the red text and the blue text, respectively. The span predicted by the model with our method actually answers to the question, while other spans with the answer text is not related to the question.

**DROP\text{num}.** Table 8 shows predictions from the model with our method and that with MML objective over training procedure. We observe that the model with our method learns to assign a high probability to the best solution (‘34-24’), while the model with MML objective fails to do so. Another notable observation is that the model with our method assign sparse distribution of likelihood over \( Z \), compared to the model with MML objective. We quantitatively define sparsity as

\[
\frac{|\{z \in Z \text{ s.t. } P(z|x; \theta) < \epsilon\}|}{|Z|}
\]

(Hurley and Rickard, 2009) and show that the model with our method gives higher sparsity than the model with MML (59 vs. 36 with \( \epsilon = 10^{-3} \), 54 vs. 17 with \( \epsilon = 10^{-4} \) on DROPR).

**WIKISQL.** WIKISQL provides the annotated SQL queries, makes it easy to compare the predictions from the model to the annotated queries. Out of 8421 examples from the development set, 7110 predictions execute the correct answers. Among those, 6296 predictions are exactly same as the annotated queries. For cases where the predictions execute the correct answers but are not exactly same as the groundtruth queries, we show four examples in Table 9. In the first example, both the annotated query and the prediction are correct, because the selected column does not matter for counting. Similarly in the second example, both queries are correct because \text{Capital (exonym)} and \text{Capital (endonym)} both indicate the capital city. In the third example, the prediction makes more sense than the annotated query because the question does not imply anything about \text{min}. In the last example, the annotated query makes more sense than the prediction because the prediction misses \text{Ship Type = battleship}. We conjecture that the model might learn to ignore some information in the question if the table header implies the table is specific about that information, hence does not need to condition on that information.
Table 9: Four examples from WIKIQA where the prediction from the model is different from annotated SQL query, although the executed answers are the same. Q, H, A and P indicate the given question, the given table header, annotated SQL query and predicted SQL query. First two example shows the case where both queries are correct. Next example shows the case that the model prediction makes more sense than the annotated query. The last example shows the cases that the annotated query makes more sense than the model prediction.