A Discrete Hard EM Approach for Weakly Supervised Question Answering

EMNLP 2019

Sewon Min¹, Danqi Chen²³, Hannaneh Hajishirzi¹⁴, Luke Zettlemoyer¹³

¹University of Washington, ²Princeton University, ³Facebook AI Research, ⁴Allen Institute of AI
A unified weak supervision scenario with a small set of possible solutions
Overview

- Reading comprehension
- Discrete reasoning task
- Semantic parsing

Task formulation

A unified weak supervision scenario with a small set of possible solutions

Learning method

A hard EM learning scheme
Q: Which composer did pianist Clara Wieck marry in 1840?  
A: Robert Schumann

Robert Schumann was a German composer and influential music critics of the Romantic era. (...) Robert Schumann himself refers to it as “an affliction of the whole hand” (...) Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket” (...) Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. (...) Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.
Q: Which composer did pianist Clara Wieck marry in 1840?
A: Robert Schumann

Robert Schumann was a German composer and influential music critics of the Romantic era. (…) Robert Schumann himself refers to it as “an affliction of the whole hand” (…) Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket” (…) Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. (…) Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.
Robert Schumann was a German composer and influential music critics of the Romantic era. (...) Robert Schumann himself refers to it as “an affliction of the whole hand” (...) Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket” (...) Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. (Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.)
Robert Schumann was a German composer and influential music critics of the Romantic era. He himself refers to it as “an affliction of the whole hand.” Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket” (...) Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. (...) Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.

Q: Which composer did pianist Clara Wieck marry in 1840?
A: Robert Schumann
Problem

Multi-mention Reading Comprehension

Q: Which composer did pianist Clara Wieck marry in 1840?
   A: Robert Schumann

Input: Q, Document

Solution z (span in this case)

Output: A (text)

From TriviaQA (Joshi et al. 2017)
Which composer did pianist Clara Wieck marry in 1840?

A: Robert Schumann

Robert Schumann was a German composer and influential music critic of the Romantic era. Robert Schumann himself refers to it as “an affliction of the whole hand.” Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket.” Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.
Problem

Multi-mention Reading Comprehension

Q: Which composer did pianist Clara Wieck marry in 1840?
A: Robert Schumann

Robert Schumann was a German composer and influential music critics of the Romantic era. (…) Robert Schumann himself refers to it as “an affliction of the whole hand.” (…) Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket” (…) Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. (…) Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.

Z = \{z1, z2, z3, z4, z5\}

From TriviaQA (Joshi et al. 2017)
Problem

Discrete Reasoning Task

Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?

A: 4

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

\[ 40 - 36 = 4 \]

From DROP (Dua et al 2019)
Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?
A: 4

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

\[40 - 36 = 4\]
Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?

A: 4

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

40 - 36 = 4

From DROP (Dua et al 2019)
Problem

Discrete Reasoning Task

Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?

A: 4

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

40 - 36 = 4

From DROP (Dua et al 2019)
Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?
A: 4

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

40 - 36 = 4
Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?

A: 4

Titan responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.
Problem

Discrete Reasoning Task

Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?

A: 4

Given

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

We can find a solution set \( Z \)

\[ Z = \{ \text{"41-37", "41-37", "40-36"} \} \]

From DROP (Dua et al 2019)
Related Work 1
– Reading Comprehension

1. Heuristics --- first span, random span
   • Very competitive baseline (partially because of dataset bias)

2. Maximum Marginal Likelihood (MML)
   • A latent variable learning method which maximizes \( \sum_{z \in Z} P(z|Q, D) \)

Example papers with first only: Joshi et al 2017, Tay et al 2018, Talmor & Berant 2019
Related Work 1
– Reading Comprehension

1. Heuristics --- first span, random span
   • Very competitive baseline (partially because of dataset bias)

2. Maximum Marginal Likelihood (MML)
   • A latent variable learning method which maximizes $\sum_{z \in Z} P(z | Q, D)$

We will show: First only & MML are similar; Our hard EM method outperforms them significantly

Example papers with first only: Joshi et al 2017, Tay et al 2018, Talmor & Berant 2019
Related Work 2
– Semantic Parsing


$x: \text{There is a small yellow item not touching any wall}

y: \text{True}

z: \text{Exist(Filter(ALL_ITEMS, } \lambda x. \text{And(And(IsYellow}(x),
\text{IsSmall}(x)), \text{Not(IsTouchingWall}(x, \text{Side.Any})))))$
Related Work 2
– Semantic Parsing


\[ x : \text{There is a small yellow item not touching any wall} \]
\[ y : \text{True} \]
\[ z : \text{Exist(\text{Filter(ALL_ITEMS, } \lambda x.\text{And(And(IsYellow}(x), \text{IsSmall}(x)), \text{Not(IsTouchingWall}(x, \text{Side.Any})))))} \]

CNLVR dataset (Suhr et al 2017); Figure from Goldman et al 2018
Related Work 2
– Semantic Parsing


Very large or infinite search space -> reward-based methods are used with no precomputed set of logical forms

\(x: \text{There is a small yellow item not touching any wall}
\)
\(y: \text{True}
\)
\(z: \text{Exist(\text{Filter(ALL_ITEMS, } \lambda x. \text{And(And(IsYellow}(x), \text{IsSmall}(x)), \text{Not(IsTouchingWall}(x, \text{Side.Any}))))))}
\)

CNLVR dataset (Suhr et al 2017); Figure from Goldman et al 2018
Related Work 2
– Semantic Parsing


Very large or infinite search space -> reward-based methods are used with no precomputed set of logical forms

This paper: only focus on problems where a solution set can be precomputed

CNLVR dataset (Suhr et al 2017); Figure from Goldman et al 2018
Related Work 2
– Semantic Parsing


Very large or infinite search space -> reward-based methods are used with no precomputed set of logical forms

This paper: only focus on problems where a solution set can be precomputed → Precomputing a solution set and using hard EM update is better than reward-based methods.

CNLVR dataset (Suhr et al 2017); Figure from Goldman et al 2018
A unified weak supervision scenario with a small set of possible solutions
Q: Which composer did pianist Clara Wieck marry in 1840?
A: Robert Schumann

Robert Schumann was a German composer and influential music critics of the Romantic era. (...) Robert Schumann himself refers to it as “an affliction of the whole hand” (...) Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket” (...) Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. (...) Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.
Robert Schumann was a German composer and influential music critics of the Romantic era. (…) Robert Schumann himself refers to it as “an affliction of the whole hand” (…) Robert Schumann is mentioned in a 1991 episode of Seinfeld “The Jacket” (…) Clara Schumann was a German musician and composer. Her husband was the composer Robert Schumann. (…) Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introduction to Robert Schumann.
Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?  
A: 4

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.
Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

\[
37 + 37 \quad 37 - 37 \quad 37 + 36 \quad 37 - 36 \quad 37 + 41 \quad 37 - 41 \quad 36 + 41 \quad 36 - 41 \quad ...
\]
Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?  
A: 4

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

37+37  37-37  37+36  37-36  37+41  37-41  36+41  36-41  ...

Execute

41-37  41-37  40-36  10-6  ...
Q: How many yards longer was Rob Bironas’ longest field goal compared to John Carney’s only field goal?  

A: 4

Titans responded with Kicker Rob Bironas managing to get a 37 yard field goal. ... In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. ... John Carney getting a 36 yard field goal. ... Young and Williams hooking up with each other on a 41 yard td pass. ... Bironas nailing a 40 yard and a 25 yard field goal.

37+37  37-37  37+36  37-36  37+41  37-41  36+41  36-41  ...

Execute  

41-37  41-37  40-36  10-6  ...

Average: 8.2  
Median: 3

<table>
<thead>
<tr>
<th>Player</th>
<th>No.</th>
<th>Position</th>
<th>Year in Toronto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kyle Lowry</td>
<td>3</td>
<td>Guard</td>
<td>2012-Present</td>
</tr>
<tr>
<td>Popeye Jones</td>
<td>54</td>
<td>Forward</td>
<td>1996-1998</td>
</tr>
</tbody>
</table>

All non-compositional SQL queries with up to 3 conditions
**Q:** Which player played guard for Toronto in **1996-1997**?  
**A:** John Long

<table>
<thead>
<tr>
<th>Player</th>
<th>No.</th>
<th>Position</th>
<th>Year in Toronto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kyle Lowry</td>
<td>3</td>
<td>Guard</td>
<td>2012-Present</td>
</tr>
<tr>
<td>Popeye Jones</td>
<td>54</td>
<td>Forward</td>
<td>1996-1998</td>
</tr>
</tbody>
</table>

Select player where No.="1996"  
Select max(player) where No.="1996"  
Select min(player) where No.="1996"  
...
**Task Formulation**

<table>
<thead>
<tr>
<th>Player</th>
<th>No.</th>
<th>Position</th>
<th>Year in Toronto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kyle Lowry</td>
<td>3</td>
<td>Guard</td>
<td>2012-Present</td>
</tr>
<tr>
<td>Popeye Jones</td>
<td>54</td>
<td>Forward</td>
<td>1996-1998</td>
</tr>
</tbody>
</table>

**Q:** Which player played guard for Toronto in 1996-1997?  
**A:** John Long

- Select player where No.="1996"
- Select max(player) where No.="1996"
- Select min(player) where No.="1996"
- Select player where No.="1997"
- Select max(player) where No.="1997"
- Select min(player) where No.="1997"
**Task Formulation**

**Reading comprehension**

**Discrete reasoning task**

**Semantic parsing**

---

**Q:** Which player played guard for Toronto in 1996-1997?  
**A:** John Long

<table>
<thead>
<tr>
<th>Player</th>
<th>No.</th>
<th>Position</th>
<th>Year in Toronto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kyle Lowry</td>
<td>3</td>
<td>Guard</td>
<td>2012-Present</td>
</tr>
<tr>
<td>Popeye Jones</td>
<td>54</td>
<td>Forward</td>
<td>1996-1998</td>
</tr>
</tbody>
</table>

Select player where No.="1996"  
Select max(player) where No.="1996"  
Select min(player) where No.="1996"  
Select player where position="guard" and Year in Toronto="1996-1997"  
Select max(player) where position="guard" and Year in Toronto="1996-1997"  
Select min(player) where position="guard"  
Select min(player) where Year in Toronto="1996-1997"
Q: Which player played guard for Toronto in 1996-1997?

A: John Long

<table>
<thead>
<tr>
<th>Player</th>
<th>No.</th>
<th>Position</th>
<th>Year in Toronto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kyle Lowry</td>
<td>3</td>
<td>Guard</td>
<td>2012-Present</td>
</tr>
<tr>
<td>Popeye Jones</td>
<td>54</td>
<td>Forward</td>
<td>1996-1998</td>
</tr>
</tbody>
</table>

Select player where No.="1996"  
Select max(player) where No.="1996"  
Select min(player) where No.="1996"

Select player where position="guard" and Year in Toronto="1996-1997"  
Select max(player) where position="guard" and Year in Toronto="1996-1997"  
Select min(player) where position="guard"  
Select min(player) where Year in Toronto="1996-1997"
A unified weak supervision scenario with a small set of possible solutions
Task Formulation

- Reading comprehension
- Discrete reasoning task
- Semantic parsing

Task formulation

A unified weak supervision scenario with a small set of possible solutions

Learning method

?
Goal: train $P(z|Q, D)$
Learning

Goal: train $P(z|Q, D)$
Groundtruth solution: $\tilde{z}$

Supervised model (given $\tilde{z}$) $P(\tilde{z}|Q, D)$
Learning

Goal: train $P(z|Q,D)$

Groundtruth solution: $\bar{z}$  \{z_1, z_2, ..., z_n\} is a solution set executing the correct answer

Model $P(z|Q,D)$

Supervised model (given $\bar{z}$)  \(P(\bar{z}|Q,D)\)
Learning

Goal: train $P(z|Q, D)$

Groundtruth solution: $\bar{z}$ \{ $z_1, z_2, \ldots, z_n$ \} is a solution set executing the correct answer

At each parameter update

Model

\[
P(z|Q, D)
\]

Supervised model (given $\bar{z}$)

$P(\bar{z}|Q, D)$
Learning - MML

Goal: train $P(z|Q,D)$

Groundtruth solution: $\bar{z}$ \quad \{z_1, z_2, \ldots, z_n\} is a solution set executing the correct answer

**MML:** Marginalize over $z_1, \ldots, z_n$

At each parameter update

**Model**

$P(z|Q,D)$

\[
\begin{align*}
Z_1 & \quad 0.32 \\
Z_2 & \quad 0.58 \\
Z_n & \quad 0.01 \\
\end{align*}
\]

Supervised model (given $\bar{z}$)

$P(\bar{z}|Q,D)$

$\sum_{z \in Z} P(z|Q,D)$

Encourage

Intro
Related Work
Method
Result
Learning - Ours

Goal: train $P(z|Q, D)$

Groundtruth solution: $\tilde{z} \{z_1, z_2, ..., z_n\}$ is a solution set executing the correct answer

Ours: Encourage the most likely solution

At each parameter update

Supervised model (given $\tilde{z}$)

MML

Ours

$P(\tilde{z}|Q, D)$

$\sum_{z \in Z} P(z|Q, D)$

$max_{z \in Z} P(z|Q, D)$
Learning - Ours

Goal: train $P(z|Q,D)$

Groundtruth solution: $\tilde{z} \{z_1, z_2, \ldots, z_n\}$ is a solution set executing the correct answer

In practice, we perform annealing:

- Start with MML
- Ratio of Replacing MML to hard EM
- Hyperparameter
- Mostly Hard EM later
- Training steps
## Datasets

1. **Multi-mention Reading Comprehension**
   - TriviaQA: Distantly-supervised RC
   - NarrativeQA: Generative RC
   - TriviaQA-open: Open-domain QA
   - Natural Questions-open: Open-domain QA

2. **Discrete Reasoning Task**
   - DROP-num: Numeric reasoning

3. **Semantic Parsing**
   - WikiSQL: Non-compositional SQL query generation

Datasets are from: Joshi et al 2017; Kocisky et al 2018; Joshi et al 2017; Kwiatkowski et al 2019; Dua et al 2019; Zhong et al 2017. Note that TriviaQA-open & Natural Questions-open are open-domain versions of TriviaQA & Natural Questions, respectively.
Datasets

1. Multi-mention Reading Comprehension
   - TriviaQA: Distantly-supervised RC
   - NarrativeQA: Generative RC
   - TriviaQA-open: Open-domain QA
   - Natural Questions-open: Open-domain QA

2. Discrete Reasoning Task
   - DROP-num: Numeric reasoning

3. Semantic Parsing
   - WikiSQL: Non-compositional SQL query generation

Base Model

Multi-paragraph
BERT-QA
(Devlin et al 2019 & others)

Augmented BERT
(Dua et al 2019)

SQLova
(Hwang et al 2019)

Datasets are from: Joshi et al 2017; Kocisky et al 2018; Joshi et al 2017; Kwiatkowski et al 2019; Dua et al 2019; Zhong et al 2017. Note that TriviaQA-open & Natural Questions-open are open-domain versions of TriviaQA & Natural Questions, respectively.
Result Summary

1) First-only and MML are similar.

SOTAs are from: Wang et al 2018; Nishida et al 2019; Lee et al 2019; Lee et al 2019 ; Dua et al 2019; Agarwal et al 2019
1) First-only and MML are similar.
2) Our Hard-EM method outperforms First-only & MML consistently.

SOTAs are from: Wang et al 2018; Nishida et al 2019; Lee et al 2019; Lee et al 2019; Dua et al 2019; Agarwal et al 2019
1) First-only and MML are similar.
2) Our Hard-EM method outperforms First-only & MML consistently.
3) SOTA on five datasets.

SOTAs are from: Wang et al 2018; Nishida et al 2019; Lee et al 2019; Lee et al 2019; Dua et al 2019; Agarwal et al 2019
Results - WikiSQL

We outperform a wide range of previous semantic parsing methods.
We outperform a wide range of previous semantic parsing methods.
We outperform a wide range of previous semantic parsing methods.
How does likelihood change over training?

Discrete Reasoning Task as example

**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. ...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. 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.

**Desired equation:**

“40-36”
How does likelihood change over training?

Discrete Reasoning Task as example

**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. ...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. 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.

**Desired equation:**

“40-36”

**Solution set:**

{“41-37”, “41-37”, “40-36”, “10-6”, …}
How does likelihood change over training?

### Discrete Reasoning Task as example

**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. ...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. 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: Solution set (ordered by likelihood)

| $t$ | Pred | $Z$ (ordered by $P(z|x; \theta_t)$) |
|-----|------|-----------------------------------|
| 1k  | 10-9 | 10-6 41-37 40-36 41-37†         |
| 2k  | 37-36| 40-36 41-37 41-37‡ 10-6         |
| 4k  | 40-36| 40-36 41-37‡ 41-37 10-6         |
| 8k  | 40-36| 40-36 41-37‡ 41-37 10-6         |
| 16k | 37-36| 40-36 41-37‡ 41-37 10-6         |
| 32k | 40-36| 40-36 41-37‡ 41-37 10-6         |

### Training step

Top 1 prediction

- $t$ denotes the time step.
- $Pred$ is the predicted field goal.
- $Z$ is the set of possible field goals, ordered by likelihood.
- The last column indicates the actual field goal.
**How does likelihood change over training?**

**Discrete Reasoning Task as example**

**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. ...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. 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.

| *t* | Pred | *Z* (ordered by $P(z|x; \theta_t)$) |
|-----|------|-----------------------------------|
| 1k  | 10-9 | 10-6 41-37 40-36 41-37‡           |
| 2k  | 37-36 40-36 41-37 41-37‡ 10-6     |
| 4k  | 40-36 40-36 41-37‡ 41-37 10-6     |
| 8k  | 40-36 40-36 41-37‡ 41-37 10-6     |
| 16k | 37-36 40-36 41-37‡ 41-37 10-6     |
| 32k | 40-36 40-36 41-37‡ 41-37 10-6     |

**Solution set (ordered by likelihood)**

Correct equation is ranked first since the early stage of training.
How does likelihood change over training?

Discrete Reasoning Task as example

**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. ...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. 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.

| t | Pred | $Z$ (ordered by $P(z|x; \theta_t)$) |
|---|------|----------------------------------|
| 1k | 10-9 | 10-6 41-37 40-36 41-37‡ |
| 2k | 37-36 | 40-36 41-37 41-37‡ 10-6 |
| 4k | 40-36 | 40-36 41-37‡ 41-37 10-6 |
| 8k | 40-36 | 40-36 41-37‡ 41-37 10-6 |
| 16k | 37-36 | 40-36 41-37‡ 41-37‡ 10-6 |
| 32k | 40-36 | 40-36 41-37‡ 41-37‡ 10-6 |

Solution set (ordered by likelihood)

Correct equation is ranked first since the early stage of training. “Pushing hard towards the most likely solution is helpful”
**Effect of solution set size**

![Graphs showing accuracy and solution set size](image)

Figure 2: **Varying the size of solution set** ($|Z|$) **at test time.** We compare the model trained on MML objective (blue) and our training strategy (orange). Our approach consistently outperforms MML on DROP\textsubscript{num} and WikiSQL, especially when $|Z|$ is large.

**More performance gains when the size of solution set is large!**
Figure 2: Varying the size of solution set ($|Z|$) at test time. We compare the model trained on MML objective (blue) and our training strategy (orange). Our approach consistently outperforms MML on DROP\textsubscript{num} and WIKISQL, especially when $|Z|$ is large.

More performance gains when the size of solution set is large!
We formulate various QA problems into a weak supervision problem where a solution is not given, but a small set of potential solutions can be precomputed.

We develop a hard EM learning scheme that computes gradients relative to the most likely solution at each parameter update.

Our method outperforms baselines significantly across 6 datasets, and set new SOTA on 5 datasets by only modifying the objective.
Use cases already!

A larger solution set with more extensive search, and further improved hard EM which encourages one or zero solution using thresholding.

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard EM with thresholding</td>
<td>80.58</td>
<td>83.42</td>
</tr>
<tr>
<td>Hard EM</td>
<td>73.72</td>
<td>77.46</td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td>63.96</td>
<td>67.98</td>
</tr>
</tbody>
</table>

Table 7: Results of different training algorithms on DROP development set.


(Disclaimer: we do not know who the authors are 😅)
Thank you for listening

**Code**  https://github.com/shmsw25/qa-hard-em


**Contact**  sewon@cs.washington.edu