A Discrete Hard EM Approach for Weakly Supervised Question Answering

Summer 2019 NLP Group Study

Proceedings on Empirical Methods in Natural Language Processing 2019 (Sewon Min et al., 2019)

Presented by Minbyul Jeong

Introduction

1. Why is it Discrete Hard EM Approach?

   A: Replacing Maximum Marginal Likelihood methodology to most likely solution.

2. Why is it Weakly Supervision Question Answering task?

   A: Assuming one correct answer exist in the pre-computed answer set.

Main Contribution

1. Obvious Task Setting & Task Definition
   - Only one mention is related to answer in most of existing Question Answering task
2. Convert an aspect of Question Answering task. (Supervised Learning → Weakly Supervised Learning)
3. Improve performance independent to model architecture.

Introduction - Task Setting

Supervised Setting → Weakly Supervised Setting

1. Multi-Mention Reading Comprehension (TRIVIAQA, NARRATIVEQA, TRIVIAQA-OPEN & NATURALQ

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
era. (...) Robert Schumann himself refers to it as "an affliction of the whole hand". (...) Robert Schumann is
Jacket". (...) Clara Schumann was a German musician and composer, considered one of the most distinguished
the composer Robert Schumann <Childhood> (...) At the age of eight, the young Clara Wieck performed at
she met another gifted young pianist who had been invited to the musical evening, named Robert Schumann , w
Clara' s playing so much that he asked permission from his mother to discontinue his law studies. (...) In the spr
Brahms met Joachim in Hanover, made a very favorable impression on him, and got from him a letter of introdu
Answer (y): Robert Schumann

2. Reading Comprehension with Discrete Reasoning (DROP_{num})

Question: How many yards longer was Rob Bironas' longest field goal compared to John Carney's only f
Document: (...) The Chiefs tied the game with QB Brodie Croyle completing a 10 yard td pass to WR S
Kicker Rob Bironas managing to get a 37 yard field goal. Kansas city would take the lead prior to halfti
Kris Wilson. In the third quarter Tennessee would draw close as Bironas kicked a 37 yard field goal. Th
a 36 yard field goal. Afterwards the Titans would retake the lead with Young and Williams hooking up v
(...) Tennessee clinched the victory with Bironas nailing a 40 yard and a 25 yard field goal. With the w
Answer (y): 4

Multi-mention reading comprehension (TriviaQA)
Question: Which composer did pianist Clara Wieck marry in 1840?
Answer: Robert Schumann
Document: Robert Schumann was a German composer and influential music critics. … 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.

Reading comprehension with discrete reasoning (DROP)
Question: How many yards longer was Rob Bironas' longest field goal compared to John Carney's only field goal?
Answer: 4
Document: … 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. The Chiefs answered with kicker John Carney getting a 36 yard field goal. Titans would retake the lead with Young and Williams hooking up with each other again on a 41 yard td pass. In the fourth quarter Tennessee clinched the victory with Bironas nailing a 40 yard and a 25 yard field goal.

41 - 37 ✗
41 - 37 ✗
40 - 36 ✓

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.

Introduction - Task Setting

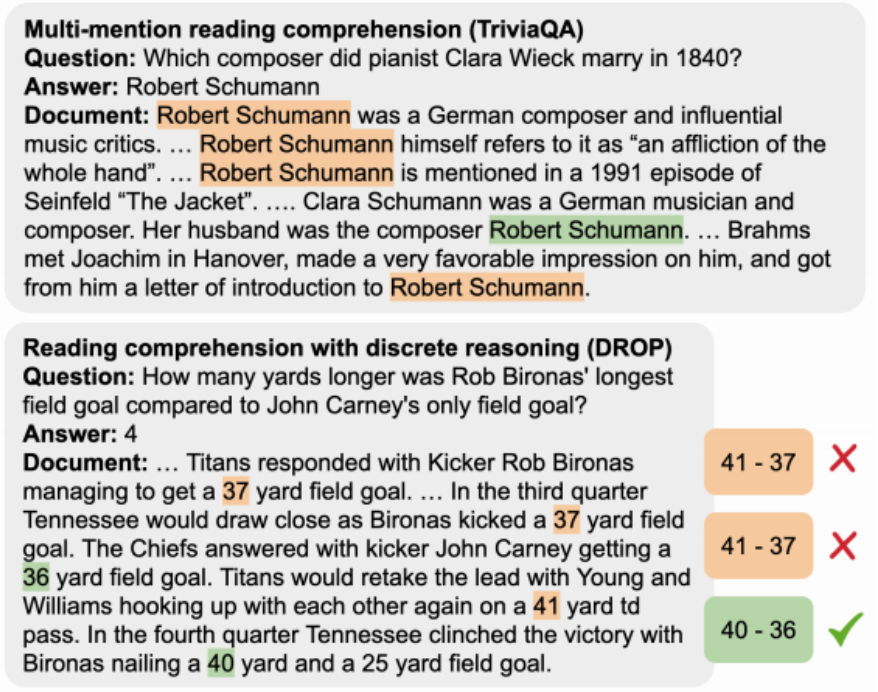


Figure 1: Examples from two different question answering tasks. **(Top) Multi-mention reading comprehension.** The answer text is mentioned five times in the given document, however, only the fourth span actually answers the question. **(Bottom) Reading comprehension with discrete reasoning.** There 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.

- Pre-computed possible solution set (one answer exists)

  : Let this solution set as a discrete latent variable.

- Hard-EM approach (procedure of giving model a inductive bias or prior knowledge that only one answer exists in this problem)

1. Predict most likely solution in pre-computed possible solution set.
2. Update model parameter with most likely solution

Related Work

1. Reading Comprehension – only considering model architecture

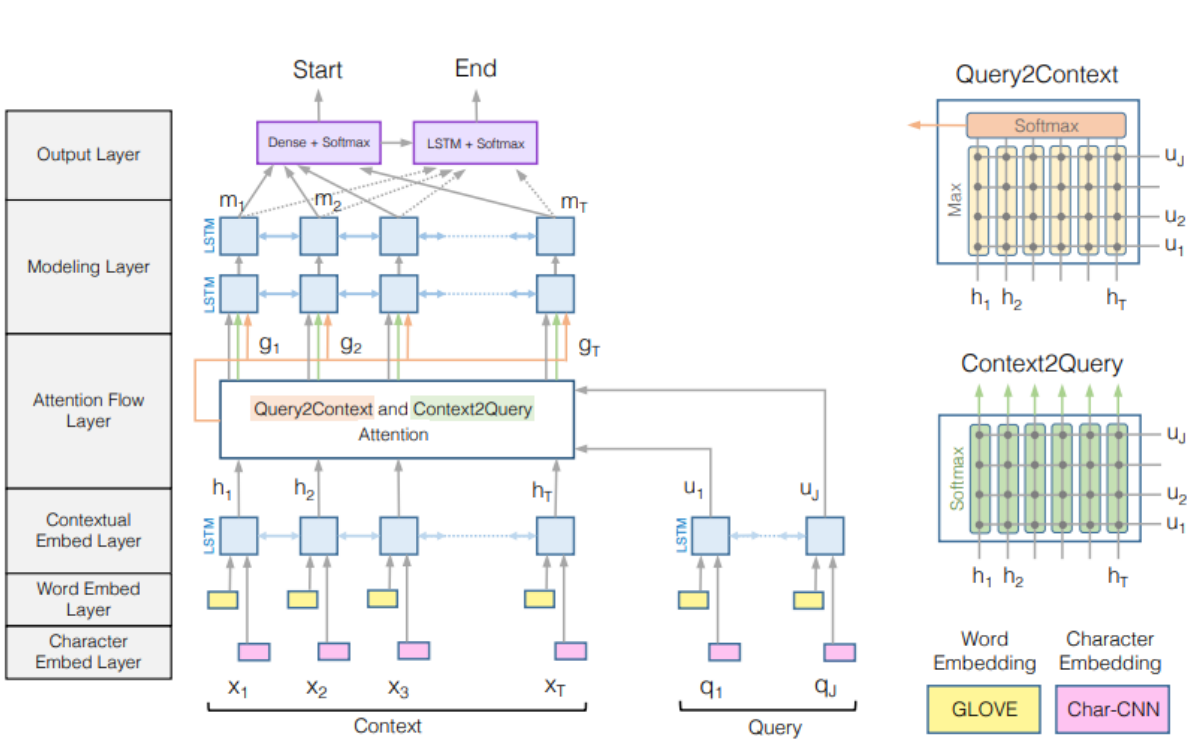
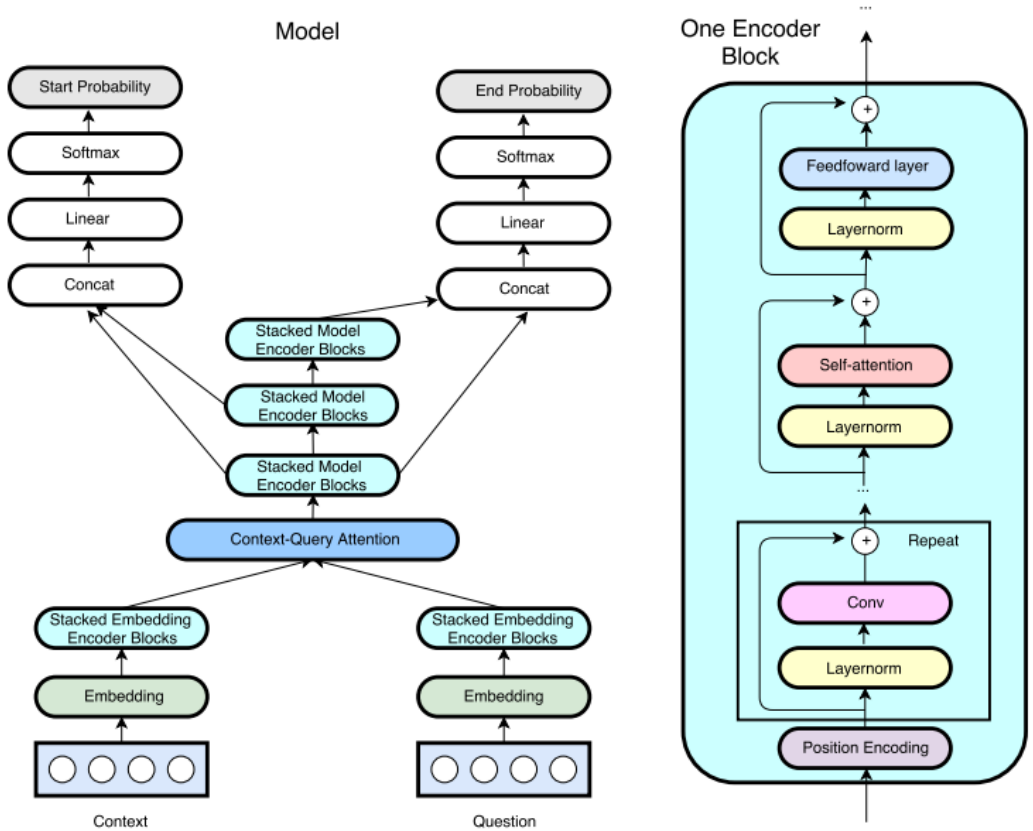


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

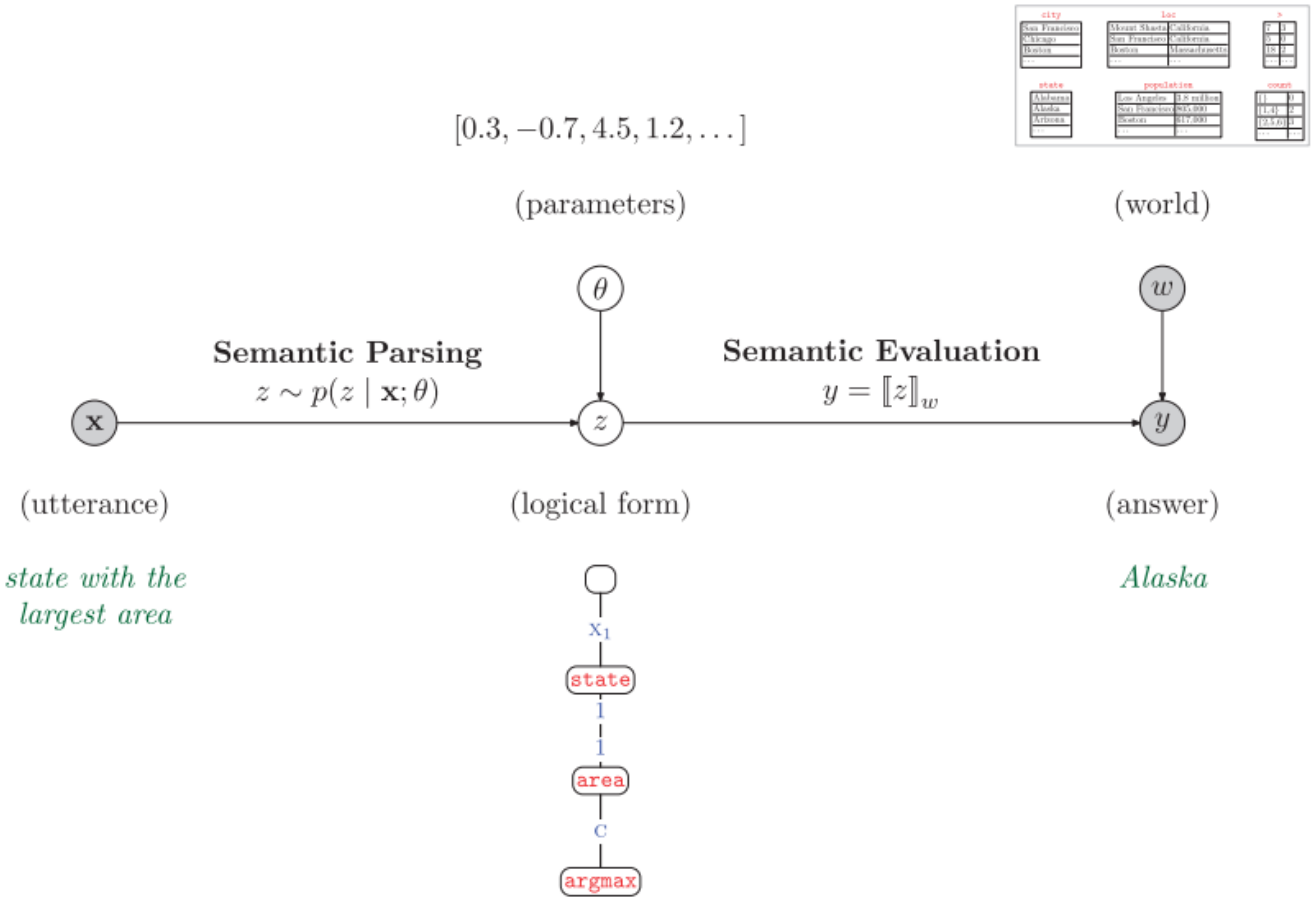
BIDAF (Seo et al., 2017)



QANet (Yu et al., 2018)

Related Work

2. Semantic Parsing – Weakly Supervised Setting (Discrete Latent Variable)



- Question and Answer pair (x, y) is given
- Logical form to compute answer is not given
- Maximum Marginal Likelihood

$$\sum_{z \in \hat{Z}} \mathbb{P}(z|x),$$

Z_hat : approximation of a set of logical forms

- Reward Based functions

Dataset

Task & Dataset	# Examples			$\lvert Z \rvert$	
	Train	Dev	Test	Avg	Median
1. Multi-mention reading comprehension					
TRIVIAQA (Joshi et al., 2017)	61,888	7,993	7,701	2.7	2
NARRATIVEQA (Kočiskỳ et al., 2018)	32,747	3,461	10,557	4.3	5
TRIVIAQA-OPEN (Joshi et al., 2017)	78,785	8,837	11,313	6.7	4
NATURALQUESTIONS-OPEN (Kwiatkowski et al., 2019)	79,168	8,757	3,610	1.8	1
2. Reading comprehension with discrete reasoning					
DROP_{num} (Dua et al., 2019)	46,973	5,850	-	8.2	3
3. Semantic Parsing					
WIKISQL (Zhong et al., 2017)	56,355	8,421	15,878	346.1	5

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.

Method - Setup

Input X : Question & Paragraph

Input Y : Answer Text (e.g., 'Robert Schumann' or '4')

Solution Z

Function F : task-specific and deterministic function, mapping a solution to textual form of the Y

(Ex : $F(Z) = Y$)

Z_{tot} : Finite set of all the possible solutions

$Z = \{Z \in Z_{tot} : F(Z) = Y\}$ → one of these solutions is True Solution

Task Setup

1. Multi-Mention Reading Comprehension (e.g., TriviaQA, NarrativeQA, open-domain QA)

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 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, considered one of the most distinguished pianists of the Romantic era. Her husband was the composer Robert Schumann . <Childhood> (...) 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 his 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 (y): Robert Schumann
f**:** Text match
Z_{tot}: All spans in the document
Z**:** Spans which match 'Robert schumann' (red text)

possible solutions $Z = \{z_1, \ldots, z_n\}$

$$g_{\max} = \max_{1 \le s_i \le e_i \le L} g([d_{s_i}, \ldots, d_{e_i}], y)$$

$$Z = \{z_i = (s_i, e_i) \text{ s.t. } g(s_i, e_i) = g_{\max}\},$$

s_i, e_i : start and end token indices

g: string match function (exact match)

Task Setup

2. Reading Comprehension with Discrete Reasoning (e.g., DROP_num)

2. Reading Comprehension with Discrete Reasoning (DROP_{num})

Question: How many yards longer was Rob 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 Samie 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 (y): 4
f: Equation executor
Z_{tot}: Equations with two numeric values and one arithmetic operation
Z: { 41-37, 40-36, 10-6, ... }

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

N_D, N_Q : Numeric value in Document and Question

S : set of pre-defined special numbers

Task Setup

3. SQL Query Generation : logical form to execute the answer is not given

3. SQL Query Generation (WIKISQL)

Question: What player played guard for Toronto in 1996-1997?
Table Header: `player, year, position, ...`
Answer (y): John Long
f: SQL executor
Z_{tot}: Non-nested SQL queries with up to 3 conditions
Z: `Select player where position=guard and year in toronto=1996-97`

```
Select max(player) where position=guard and year in toronto=1996-97
Select min(player) where position=guard
Select min(player) where year in toronto=1996-97
Select min(player) where position=guard and year in toronto=1996-97
```

$$
\begin{aligned}
Z_{\text{tot}} = \{z_i \quad &= \quad (z_i^{\text{sel}}, z_i^{\text{agg}}, \{z_{i,j}^{\text{cond}}\}_{j=1}^{3}\} \text{ s.t.} \\
z_i^{\text{sel}} \quad &\in \quad [1, n_L] \\
z_i^{\text{agg}} \quad &\in \quad \{\text{none}\} \cup A \\
z_{i,j}^{\text{cond}} \quad &\in \quad \{\text{none}\} \cup C \text{ for } j \in [1,3]\},
\end{aligned}
$$

Q = [q_1, ..., q_l]

H = [h_1, ..., h_n_L] : n_L is the number of headers

A : {sum, mean, max, min, count}

C : { (h, o, t) s.t. h ∈ [1, n_L], o ∈ {=, <, >}, t ∈ spans in Q }

Method – Learning

- Fully Supervised Setting

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

- Weakly Supervised Setting

$$Z = \{z_1, z_2, \ldots, z_n\}$$

$$\begin{aligned}\mathbb{P}(y|x;\theta) &= \sum_{z_i \in Z_{\text{tot}}} \mathbb{P}(y|z_i)\mathbb{P}(z_i|x;\theta) \\ &= \sum_{z_i \in Z} \mathbb{P}(z_i|x;\theta)\end{aligned}$$

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

Maximum Marginal Likelihood Limitation

1. MML method could assign high probability to any subset of solution.

This could be assigning low probability to real solution Z

2. Discrepancy between training and inference.

Training : optimize the sum over probabilities of Z

Inference : predict maximum probability solution

Method - Learning

Hard EM Approach

$$\tilde{z} = \text{argmax}_{z_i \in Z} \mathbb{P}(z_i|x;\theta)$$

$$\begin{aligned} J_{\text{Hard}}(\theta|x,Z) &= -\log \mathbb{P}(\tilde{z}|x;\theta) \\ &= -\log \max_{z_i \in Z} \mathbb{P}(z_i|x;\theta) \\ &= -\max_{z_i \in Z} \log \mathbb{P}(z_i|x;\theta) \\ &= \min_{z_i \in Z} J_{\text{Sup}}(\theta|x,z_i) \end{aligned}$$

→ tilde z : Assuming z as a true solution

→ encourage highest likelihood $\mathbb{P}(z_i|x;\theta)$

→ encourage lowest loss value $J_{\text{Sup}}(\theta|x,z_i)$

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

```
if self.loss_type=='first-only':
    total_loss = torch.sum(start_losses[0]+end_losses[0]+switch_losses[0])
```

```
def _take_mml(self, loss_tensor):
    return -torch.sum(torch.log(torch.sum(torch.exp(-loss_tensor - 1e10 * (loss_tensor==0).float()), 1)))
```

```
def _take_min(self, loss_tensor):
    return torch.sum(torch.min(
        loss_tensor + 2*torch.max(loss_tensor)*(loss_tensor==0).type(torch.FloatTensor).to(self.device), 1)[0])
```

Results

	TriviaQA (F1)		NarrativeQA (ROUGE-L)		TriviaQA -open (EM)		NaturalQ -open (EM)		$DROP_{num}$ w/ BERT (EM)	$DROP_{num}$ w/ QANet (EM)
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Dev
First Only	64.4	64.9	55.3	57.4	48.6	48.1	23.6	23.6	42.9	36.1
MML	64.8	65.5	55.8	56.1	47.0	47.4	26.6	25.8	39.7	43.8
Ours	66.9	67.1	**58.1**	**58.8**	**50.7**	**50.9**	**28.8**	**28.1**	**52.8**	**45.0**
SOTA	-	**71.4**	-	54.7	47.2	47.1	24.8	26.5	43.8	

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 $DROP_{num}$. 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.

ROUGE-L : longest co-occurring in sequence n-grams automatically (Longest Common Subsequence)

ROUGE-N : Overlap of N-grams between the system and reference summaries.

Results

Model	Accuracy Dev	Accuracy Test
Weakly-supervised setting		
REINFORCE (Williams, 1992)	< 10	
Iterative ML (Liang et al., 2017)	70.1	
Hard EM (Liang et al., 2018)	70.2	
Beam-based MML (Liang et al., 2018)	70.7	
MAPO (Liang et al., 2018)	71.8	72.4
MAPOX (Agarwal et al., 2019)	74.5	74.2
MAPOX+MeRL (Agarwal et al., 2019)	74.9	74.8
MML	70.6	70.5
Ours	**84.4**	**83.9**
Fully-supervised setting		
SQLNet (Xu et al., 2018)	69.8	68.0
TypeSQL (Yu et al., 2018b)	74.5	73.5
Coarse2Fine (Dong and Lapata, 2018)	79.0	78.5
SQLova (Hwang et al., 2019)	87.2	86.2
X-SQL (He et al., 2019)	**89.5**	**88.7**

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.

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$_{num}$ and WIKISQL, especially when $|Z|$ is large.

Group	Avg $\|Z\|$	Median $\|Z\|$	# train
3	3.0	3	10k
10	10.2	9	10k
30	30.0	22	10k
100	100.6	42	10k
300	300.0	66	10k

Figure 3: **Varying the size of solution set ($|Z|$) at training.** (Left) Subsets of the train set on WIKISQL varying in the size of solution set ($|Z|$). All subsets contain 10k training examples (total in the original train set is 55k). All subsets are evaluated on the same, original development set for a fair comparison. (Right) Performance across subsets of the training set with varying $|Z|$. Our method achieves substantial gains over MML.

Results

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 $\mathbb{P}(z\|x;\theta_t)$)			
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

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.

- Uniform probability distribution assigned to Z
- Gradually learns to favor the true solution