A Mutual Information Maximization Approach for the Spurious Solution Problem in Weakly Supervised Question Answering

ACL2021

- 论文探讨了弱监督场景下QA模型训练优化问题，通过最大化“互信息”的方式达到降噪目的。

符号表示

字符	意义	举例
d	文档	1990年我出生了，1998年我在吃火锅。
q	问题	什么时候出生?
a	答案	1990年
z	solution	1990年(位置0-4) 或者1998年(位置10-14)
Z	Solution集合	[1990年(位置0-4) ,1998年(位置10-14)]

- 其中**solution**是正确答案的来源处，且能够通过**solution**推导出正确的答案。

符号表示

Discrete Reasoning over Paragraphs

Question: How many years after the Battle of Powder River did Powerville Montana become the first establishment in the county?

Answer: 2

Paragraph: … From September 1-15, 1865, the Powder River Expedition (1865) battled Native Americans in the Powder River Battles (1865) near the future site of Broadus. On March 17, ①1876, the Battle of Powder River occurred in the south-central part of the county, about southwest of Broadus. In June ②1876 six companies of the 7th Cavalry Regiment (United States) led by Major Marcus Reno marched along the Powder River … On November 1, ③1878, Powderville, Montana became the first establishment in the county, … On April 5, 1879, the Mizpah Creek Incidents …

Possible Solution(s):

③1878 - ①1876 ✓

③1878 - ②1876 ✗

- 通过solution都能产生正确的答案，但是其过程可能是由于错误的信息造成的，这种solution称之为虚假的solution。

任务定义

- 弱监督QA

✓ 只提供最终答案作为监督信号，缺少推导答案的正确solution。

✓ 这种设置简化了数据收集，却使模型学习面临虚假solution的问题：可能存在许多虚假方法来推导正确的答案，用虚假的solution会损害模型的性能（例如误导模型产生不合理的solution或错误的答案）。

Multi-mention Reading Comprehension

Question: In the television series 'Thunderbirds', what is Lady Penelope's surname?
Answer: Creighton Ward
Document(s): Born on 24 December 2039, Lady Penelope is the 26-year old daughter of aristocrat Lord Hugh Creighton Ward and his wife, Amelia. The early years of her life were spent at Creighton Ward Mansion. … Lady Penelope Creighton Ward is a fictional character introduced in the British mid-1960s Supermarionation television series Thunderbirds, … Perce is the gardener for the 2000 acre Creighton Ward estate and a friend of Parker. …
Possible Solution(s):
``Creighton Ward'' across the document(s), only the third one is correct

Discrete Reasoning over Paragraphs

Question: How many years after the Battle of Powder River did Powerville Montana become the first establishment in the county?
Answer: 2
Paragraph: … From September 1-15, 1865, the Powder River Expedition (1865) battled Native Americans in the Powder River Battles (1865) near the future site of Broadus. On March 17, ①1876, the Battle of Powder River occurred in the south-central part of the county, about southwest of Broadus. In June ②1876 six companies of the 7th Cavalry Regiment (United States) led by Major Marcus Reno marched along the Powder River … On November 1, ③1878, Powderville, Montana became the first establishment in the county. … On April 5, 1879, the Mizpah Creek Incidents …
Possible Solution(s):
③1878 - ①1876 ✓
③1878 - ②1876 ✗

Semantic Parsing

Question: Give me the kickoff time of the game that was aired on CBS against the St. Louis Cardinals.
Answer: 1:00
Table Header: | week | date | opponent | result | kickoff[a] | game site | tv | attendance | …
Possible Solution(s):
SELECT (kickoff[a]) WHERE tv=CBS AND opponent=St. Louis Cardinals ✓
SELECT (kickoff[a]) WHERE opponent=St. Louis Cardinals ✗

- 之前方法的缺点:

在选择solution进行训练时没有明确考虑问题和solution之间的语义相关性。

本文动机

Multi-mention Reading Comprehension

Question: In the television series 'Thunderbirds', what is Lady Penelope's surname?
Answer: Creighton Ward
Document(s): Born on 24 December 2039, Lady Penelope is the 26-year old daughter of aristocrat Lord Hugh Creighton Ward and his wife, Amelia. The early years of her life were spent at Creighton Ward Mansion. … Lady Penelope Creighton Ward is a fictional character introduced in the British mid-1960s Supermarionation television series Thunderbirds, … Perce is the gardener for the 2000 acre Creighton Ward estate and a friend of Parker. …
Possible Solution(s):
``Creighton Ward'' across the document(s), only the third one is correct

Discrete Reasoning over Paragraphs

Question: How many years after the Battle of Powder River did Powerville Montana become the first establishment in the county?
Answer: 2
Paragraph: … From September 1-15, 1865, the Powder River Expedition (1865) battled Native Americans in the Powder River Battles (1865) near the future site of Broadus. On March 17, ①1876, the Battle of Powder River occurred in the south-central part of the county, about southwest of Broadus. In June ②1876 six companies of the 7th Cavalry Regiment (United States) led by Major Marcus Reno marched along the Powder River … On November 1, ③1878, Powderville, Montana became the first establishment in the county, … On April 5, 1879, the Mizpah Creek Incidents …
Possible Solution(s):
③1878 - ①1876 ✓
③1878 - ②1876 ✗

Semantic Parsing

Question: Give me the kickoff time of the game that was aired on CBS against the St. Louis Cardinals.
Answer: 1:00
Table Header: | week | date | opponent | result | kickoff[a] | game site | tv | attendance | …
Possible Solution(s):
SELECT (kickoff[a]) WHERE tv=CBS AND opponent=St. Louis Cardinals ✓
SELECT (kickoff[a]) WHERE opponent=St. Louis Cardinals ✗

- 一个问题通常包含关于如何得出答案的重要线索，而错误的解决方案及其上下文往往不能很好地与问题保持一致。

要回答这个问题，我们需要知道波德河战役的开始年份。

这一年需要发生一场战役。

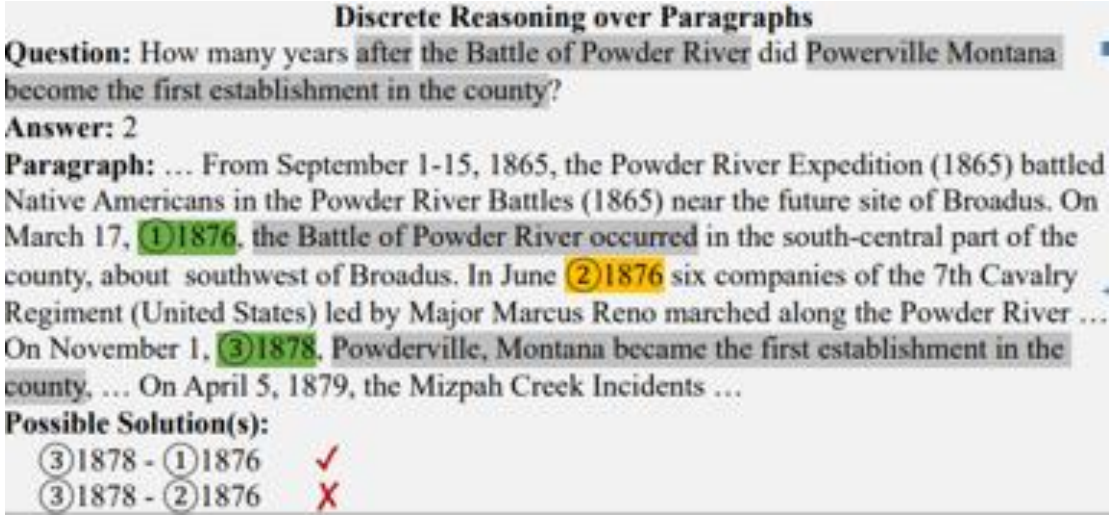
无关紧要

- 为了挖掘问题和解决方案之间的语义相关性，可以最大限度地提高问题-答案对和模型预测解决方案之间的相互信息。

- 可以预先计算一个包含正确答案的小范围、特定任务的可能解决方案集。然后着重研究在这种情况下如何处理伪解问题。

模型方法

- 给定一个实例[d，q，a]，解决方案集Z通常只包含一个最适合该实例的solution。
- 利用问题和其solution之间的语义关联，通过相互信息最大化来缓解虚假solution的问题。
- 目标是获得最佳的特定任务模型model，使以下条件的相互信息最大化。



- 在训练阶段: 每一个针对于特定任务的模型 $P_\theta(z|d,q)$ 都会和一个问题重构器 $P_\phi(q|d,z)$ 配对，从而最大化<q,a>和z 之间的互信息。

- 在测试阶段:输入<d,q> , 运用训练好的针对于特定任务的模型 ， 输出结果。

模型方法

$$\theta^* = \arg\max_\theta I_\theta(\langle q,a\rangle; z|d)$$

$$= \arg\max_\theta H(\langle q,a\rangle|d) - H_\theta(\langle q,a\rangle|d,z)$$

$$= \arg\max_\theta -H_\theta(\langle q,a\rangle|d,z)$$

$$= \arg\max_\theta E_{P(d,q,a)} E_{P_\theta(z|d,q,a)} \log P_\theta(q,a|d,z)$$

- 条件互信息:

$$\begin{aligned} I(X;Y \mid Z) &= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \\ &= H(X \mid Z) - H(X \mid Y,Z) \\ &= H(X \mid Z) + H(Y \mid Z) - H(X,Y \mid Z) \end{aligned}$$

$H(\langle q,a\rangle \mid d)$ 和模型参数无关

- 条件熵

$$\begin{aligned} \mathrm{H}(Y \mid X) &\equiv \sum_{x\in\mathcal{X}} p(x)\mathrm{H}(Y \mid X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y \mid x) \log p(y \mid x) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y \mid x) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log p(y \mid x) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)} \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log \frac{p(x)}{p(x,y)} \end{aligned}$$

不可计算的，引入了问题重构器。

因为得到的solution只依赖于参考文档d和提出的问题q。

$P_\theta(z \mid d,q,a)$ 和 $P_\theta(z \mid d,q)$ 等价

$$P_\theta(z \mid d,q,a) = \begin{cases} \frac{P_\theta(z|d,q)}{\sum_{z'\in Z} P_\theta(z'|d,q)} & z \in Z \\ 0 & z \notin Z \end{cases}$$

使用 $\mathbb{I}(f(z)=a)P_\phi(q \mid d,z)$ 来进行估计。

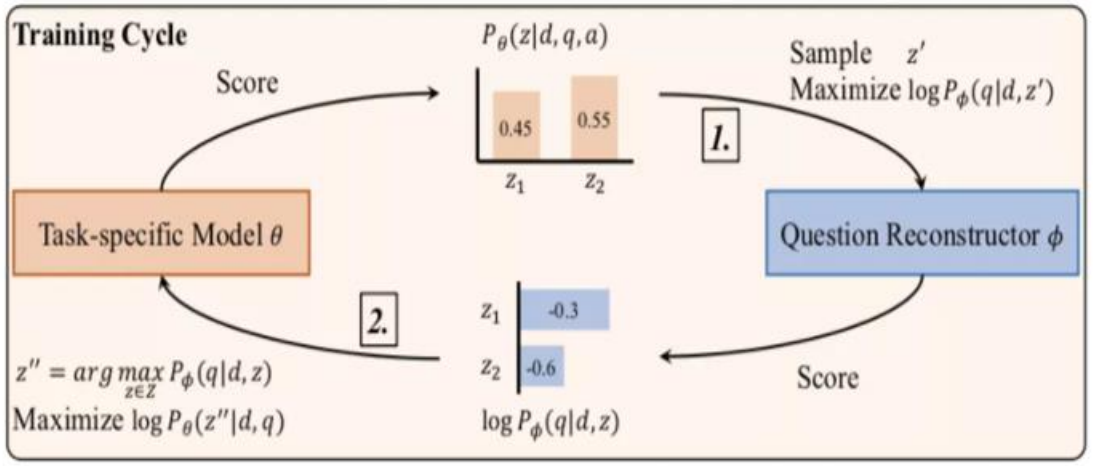
指示函数: $\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$

使得f(z)逼近真实答案的q才成立。

模型方法

$$\theta^* = \arg\max_\theta I_\theta(\langle q, a\rangle; z|d)$$
$$= \arg\max_\theta H(\langle q, a\rangle|d) - H_\theta(\langle q, a\rangle|d, z)$$
$$= \arg\max_\theta -H_\theta(\langle q, a\rangle|d, z)$$
$$= \arg\max_\theta E_{P(d,q,a)} E_{P_\theta(z|d,q,a)} \log P_\theta(q, a|d, z)$$

类似于EM算法



$$\theta^* = \arg\max_\theta \mathcal{L}_1 + \mathcal{L}_2$$
$$\mathcal{L}_1 = E_{P(d,q,a)} E_{P_\theta(z|d,q,a)} \log P_\phi(q \mid d, z)$$
$$\mathcal{L}_2 = E_{P(d,q,a)} E_{P_\theta(z|d,q,a)} \log \frac{P_\theta(q, a \mid d, z)}{P_\phi(q \mid d, z)}$$

To optimize Eq. 3 is to repeat the following training cycle, which is analogous to the EM algorithm:

1. Minimize $\mathcal{L}_2$ w.r.t. the question reconstructor ϕ to draw $P_\phi(q|d, z)$ close to $P_\theta(q, a|d, z)$, by sampling a solution $z' \in Z$ according to its posterior prediction probability $P_\theta(z|d, q, a)$ (see Eq. 2) and maximizing $\log P_\phi(q|d, z')$.
2. Maximize $\mathcal{L}_1$ w.r.t. the task-specific model θ. $\mathcal{L}_1$ can be seen as a reinforcement learning objective with $\log P_\phi(q|d, z)$ being the reward function. During training, the reward function is dynamically changing and may be of high variance. As we can compute the reward for all $z \in Z$, we therefore adopt a greedy but more stable update method, i.e., to maximize $\log P_\theta(z''|d, q)$ where $z'' = \arg\max_{z\in Z} \log P_\phi(q|d, z)$ is the best solution according to the question reconstructor.

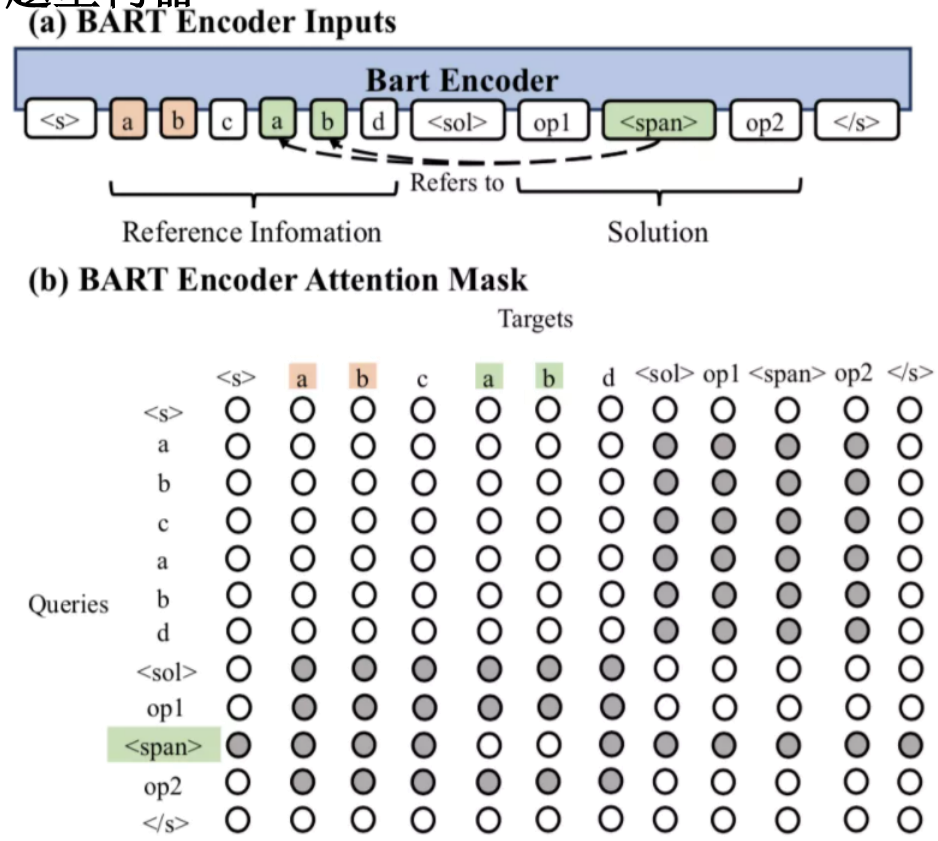
根据后验概率预测概率 $P_\theta(z \mid d, q, a)$ 在solution集合中进行采样 $z' \in Z$ 得到z'，同时最大化 $\log P_\phi(q \mid d, z')$，使得 $P_\phi(q \mid d, z)$ 逼近$P_\theta(q, a \mid d, z)$。

$$P_\theta(z \mid d, q, a) = \begin{cases} \frac{P_\theta(z|d,q)}{\sum_{z'\in Z} P_\theta(z'|d,q)} & z \in Z \\ 0 & z \notin Z \end{cases}$$

最大化L1.基于生成question的模型，再通过模型找到使得P(q|d,z)概率最大的q来间接获取Z。

模型方法

- 问题重构器



- ✓ 使用BART-base作为预训练的seq2seq模型的问题重构器。
- ✓ 如果仅仅是使用参考文本d和solution z进行拼接的话会出现问题，因为它们的上下文语义已经丢失，具有相同表面形式的不同solution就不能再被区分。

Discrete Reasoning over Paragraphs

Question: How many years after the Battle of Powder River did Powerville Montana become the first establishment in the county?

Answer: 2

Paragraph: … From September 1-15, 1865, the Powder River Expedition (1865) battled Native Americans in the Powder River Battles (1865) near the future site of Broadus. On March 17, ①1876, the Battle of Powder River occurred in the south-central part of the county, about southwest of Broadus. In June ②1876 six companies of the 7th Cavalry Regiment (United States) led by Major Marcus Reno marched along the Powder River … On November 1, ③1878, Powderville, Montana became the first establishment in the county. … On April 5, 1879, the Mizpah Creek Incidents …

Possible Solution(s):

③1878 - ①1876 ✓

③1878 - ②1876 ✗

- ✓ 论文在拼接的基础之上，通过限定encoder中，两种类型输入文本之间的attention的范围，获取更好的参考文档和solution语义表征。

- ✓ <s>和</s>分别表示输入序列的开始和结束。
- ✓ 灰色的圆圈表示相应位置的两个token之间不再互相“关注”。
- ✓ 只关注文本中的指定的solution。
- ✓ solution的选取来自于文本d不同位置中(不同位置的span，表示不同的solution)。

实验设置

数据集:

Datasets	# Examples			$\|Z\|$	
	Train	Dev	Test	Avg	Median
Multi-mention Reading Comprehension					
Quasar-T	37,012	3,000	3,000	8.1	4
WebQuestions	3,778	-	2,032	52.1	36
Discrete Reasoning over Paragraphs					
DROP	69,669	7,740	9,535	5.1	2
Semantic Parsing					
WikiSQL	56,355	8,421	15,878	315.4	4

Table 1: Statistics of the datasets we used. Statistics of the size of solution set $|Z|$ are computed on Train sets.

For convenience, we denote reference information as $d = [d_1, d_2, ..., d_{|d|}]$ and denote a question as $q = [q_1, q_2, ..., q_{|q|}]$ where d_i and q_j are a token of d and q respectively. A span from reference information and a question span is represented as $(s, e)^d$ and $(s, e)^q$ respectively, where s and e are start and end index of the span respectively.

Baseline:

First Only

MML

HardEM

HardEM-thres

VAE

实验结果

Multi-mention:

	Quasar-T				WebQuestions	
	Dev		Test		Test	
	EM	F1	EM	F1	EM	F1
First Only	36.0	43.9	35.6	42.8	16.7	22.6
MML	40.1	47.4	39.1	46.5	18.4	25.0
HardEM	41.5	49.1	40.7	47.7	18.0	24.2
HardEM-thres	42.8	50.2	41.9	49.4	19.0	25.3
Ours	**44.7**‡	**52.6**‡	**44.0**‡	**51.5**‡	**20.4**‡	**27.2**‡

Discrete Reasoning over Paragraph:

	Overall Test		Number (61.97%)		Span (31.47%)		Spans (4.99%)		Date (1.57%)	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
MML	58.99‡	62.30‡	55.38	55.58	69.96	75.51	39.29	66.01	42.57	49.05
HardEM	68.52‡	71.88‡	68.40	68.70	73.50	79.25	44.79	69.63	49.32	56.87
HardEM-thres	69.06	72.35‡	69.05	69.39	74.61	79.79	39.50	66.38	52.67	58.75
VAE	32.34‡	36.28‡	51.65	52.35	0.37	10.01	0.00	8.89	0.00	4.11
Ours	**69.35**	**72.92**	**69.96**	**70.27**	73.38	79.32	42.86	**70.42**	48.67	57.47

Text2SQL:

Model	Execution Accuracy	
	Dev	Test
Fully-supervised Setting		
SQLova (Hwang et al., 2019)	87.2	86.2
HydraNet (Lyu et al., 2020)	**89.1**	**89.2**
Weakly-supervised Setting		
MeRL (Agarwal et al., 2019)	74.9	74.8
GRAPPA (Yu et al., 2021)	85.9	84.7
MML(Min et al., 2019)	70.6	70.5
HardEM	84.5‡	84.1‡
HardEM-thres	85.2†	84.1‡
Ours	**85.9**	**85.6**

实验结果表明，在弱监督场景下论文提出的训练策略有着相当的效果提升，三种场景下的几个数据集提升基本在**2-3**个点上，提升幅度比较显著。

消融实验

- **不同大小的解决方案集对测试数据的性能**

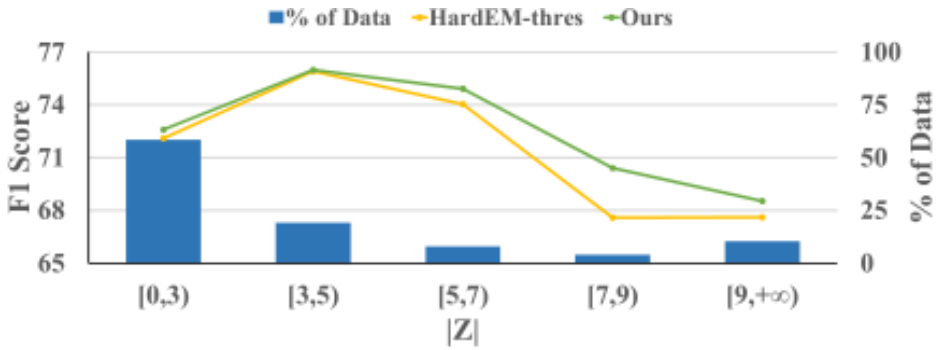


Figure 4: Performance on test examples with different size of Z on DROP.

Datasets	# Examples			$\lvert Z \rvert$	
	Train	Dev	Test	Avg	Median
Multi-mention Reading Comprehension					
Quasar-T	37,012	3,000	3,000	8.1	4
WebQuestions	3,778	-	2,032	52.1	36
Discrete Reasoning over Paragraphs					
DROP	69,669	7,740	9,535	5.1	2
Semantic Parsing					
WikiSQL	56,355	8,421	15,878	315.4	4

Table 1: Statistics of the datasets we used. Statistics of the size of solution set $|Z|$ are computed on Train sets.

始终优于**hardem -thres**

实验结果

- **|Z|的影响:**

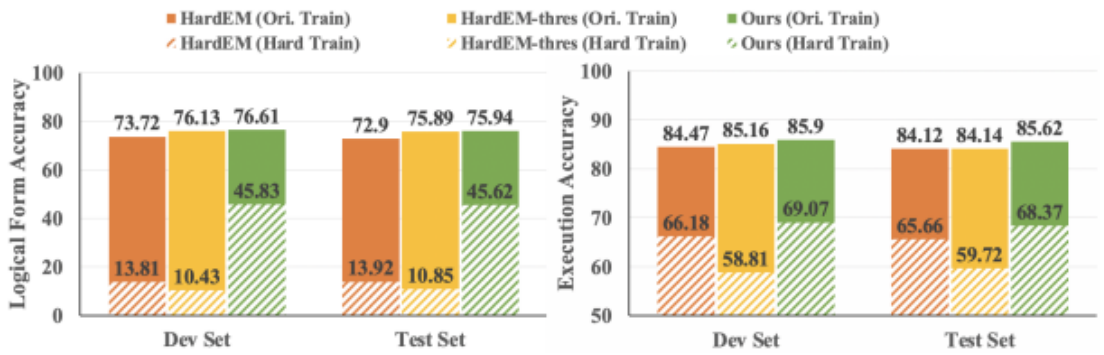


Figure 5: Logical form accuracy (left) and execution accuracy (right) on dev set and test set of WikiSQL. A method marked with *Ori. Train* or *Hard Train* means the evaluated model is trained on the original train set or a hard subset of training data, respectively. The hard train set consists of 10K training data with the largest solution set; the average size of solution set is 1,554.6.

- ✓ 验证文中提出的方法是否能够处理极度嘈杂的solution 集合。
- ✓ 作者从WikiSQL的原始训练集中提取了一个难训练集(hard train set)，该训练集合上的|Z|值(solution的数量)远大于原始数据的平均值。
- ✓ 目的是为了提高数据中噪声的大小。
- ✓ 然后，用不同的学习方法比较了在原始训练集和难训练集上训练的模型的性能。

从实验结果中可以看到，文中所提出的策略在噪声更多的数据上能够发挥更好的性能。

实验结果

■ 问题重构器的影响:

	DROP				WikiSQL (Hard Train Set)			
	Dev		Test		Dev		Test	
	EM	F1	EM	F1	LF. Acc	Exe. Acc	LF. Acc	Exe. Acc
T-scratch	**61.5**	66.3	69.0	72.4	24.7	67.9	24.9	67.5
T-DAE	**61.5**	66.3	**69.4**	72.7	**49.4**	68.9	**48.5**	**68.4**
$BART_{base}$	**61.5**	**66.4**	69.3	**72.9**	45.8	**69.1**	45.6	**68.4**

Table 7: Results with different question reconstructors. *LF. Acc* and *Exe. Acc* are logical form accuracy and execution accuracy, respectively. *T-scratch* is a Transformer without pre-training. *T-DAE* is a Transformer pre-trained as a denoising auto-encoder of questions.

✓ 为了验证不同种类问题重构器对模型训练的影响

✓ 文中选取了另外2类transformer-based生成器：
T-scratch是原生的transformer，没有进行预训练。
T-DAE是通过降噪自动编码机进行预训练了的transformer。

实验的结果表明了其实对于具体的任务，BART可能并不是最好的选择，问题的重构器对于不同的场景有着不同的要求，实操时可以试下不同的类型。

总结

- 论文利用问题重构，提出新颖的训练策略，最大化"问题-答案"pairs和solution(候选答案)之间的互信息，以达到更好的降噪效果。

- 其它任务中也可以考虑应用这种训练方式，降低标注成本，同时可能提高模型训练的性能。

THANK YOU