Weakly Supervised Problem / Transfer Learning in Question Answering

Department of Computer Science, Yonsei University

Seungone Kim

louisdebroglie@yonsei.ac.kr

Referenced Papers

- Prerequisites / Additional Papers
  - (Multi-mention RC) Text Understanding with the attention sum reader network [ACL 2016]
  - (Multi-mention RC) Simple and effective multi-paragraph reading comprehension [ACL 2018]
  - (Multi-mention RC) Latent retrieval for weakly supervised open domain question answering [ACL 2019]

  - (WikiSQL) Neural semantic parsing with type constraints for semi-structured tables [EMNLP 2017]
  - (WikiSQL) Memory augmented policy optimization for program synthesis and semantic parsing [NeurIPS 2018]
  - (WikiSQL) A Comprehensive Exploration on WikiSQL with Table-Aware Word Contextualization [NeurIPS 2019 Workshop]

  - (RC with discrete reasoning) DROP : A reading comprehension benchmark requiring discrete reasoning over paragraphs [NAACL 2019]
  - (RC with discrete reasoning) Fast and accurate reading comprehension by combining self-attention and convolution [ICLR 2018]

  - Multiqa : An empirical investigation of generalization and transfer in reading comprehension [ACL 2019]
  - Exploring the limits of transfer learning with a unified text-to-text transformer (JMLR 2020)
  - Comprehensive Multi-Dataset Evaluation of Reading Comprehension (EMNLP 2019 Workshop)
  - Massively multilingual neural machine translation in the wild : Findings and Challenges (NAACL 2019)

- Key Papers
  - A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]
  - UNIFIEDQA : Cross Format Boundaries with a Single QA System [EMNLP 2020]

A Discrete Hard EM Approach for Weakly Supervised Question Answering

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

EMNLP 2019

A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

- Many QA tasks only provide **weak supervision** for how the answer should be computed
  - TRIVIAQA : answers are entities that can be mentioned multiple times in supporting documents.
  - DROP : answers can be computed by deriving many different equations from numbers in the reference text.
  - It is natural to **model such ambiguities with a latent variable** during learning, but most prior work on RC focused on **model architecture** and used **heuristics** to map the weak supervision to full supervision.

- Formulate a wide range of weakly supervised QA tasks as **discrete latent-variable learning problems**
  - Define a **solution** to be a particular derivation of a model to predict the answer.
  - The learning challenge is to determine which solution in the set is the correct one, while estimating a complete QA model.

  - **Hard EM Learning Scheme** that computes gradients relative to the most likely solution at each update.
  - Predict the most likely solution according to the current model from the precomputed set -> Update model parameters to further encourage its own prediction.
  - Intuitively, these hard updates more strongly enforce **prior beliefs** that there is a single correct solution.

  - Using **hard updates** instead of maximizing marginal likelihood(MML) is key to SOTA results as it encourages the model to find the one correct answer.

A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

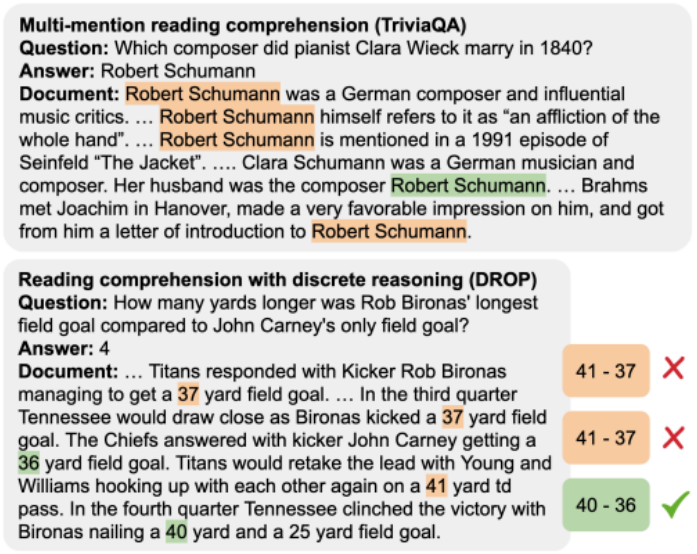


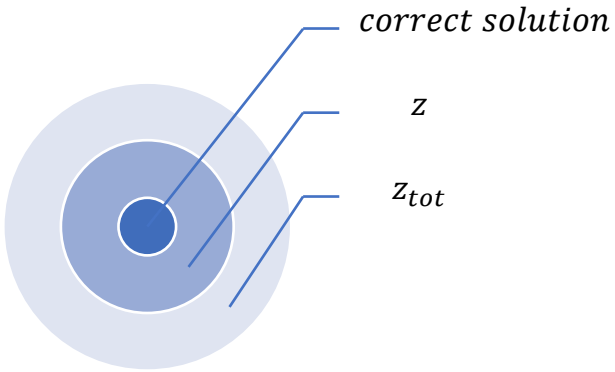
Figure 1: Examples from two different question answering tasks. **(Top) Multi-mention reading comprehension.** The answer text is mentioned five times in the given document, however, only the fourth span actually answers the question. **(Bottom) Reading comprehension with discrete reasoning.** There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

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.

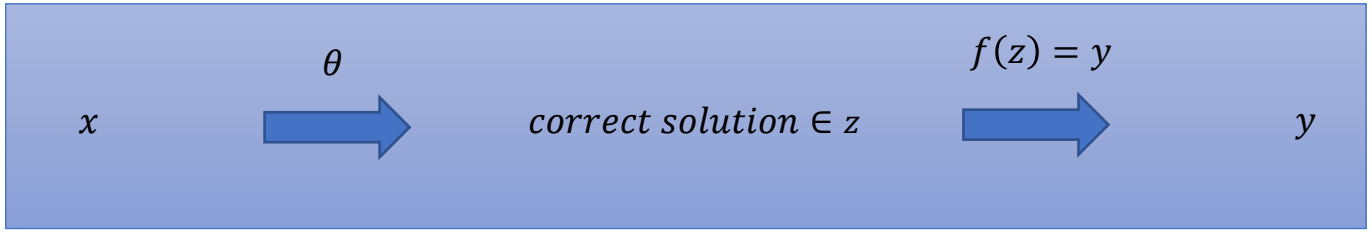
A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

- Setup
  - Let **input x** be the input of a QA system and y be the answer text.
  - Define a **solution z** as a particular derivation that a model is supposed to produce for the answer prediction.
  - Let **function f** denote a task-specific, deterministic function which maps a solution to the textual form of the answer.
  - Goal is to learn a model (with parameters θ) 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 θ is estimated based on a collection of $(x, \bar{z})$ pairs.
  - In a **weakly supervised setting**, $\bar{z}$ is not given, so z_{tot} (a finite set of all possible solutions) should be defined.
  - Then it is possible to obtain $Z = \{z \in z_{tot} : f(z) = y\}$ by enumerating all $z \in z_{tot}$, where Z is a set of all possible solution that lead to the correct answer.
  - Z contains one solution that we want to learn to produce, and potentially many other spurious ones.

  - At inference time, model produces a solution $\tilde{z} \in z_{tot}$ from input x with respect to θ and predicts the final answer as $f(\tilde{z})$.



A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

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)

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, ... }

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
```

Table 2: Examples of the input, answer text (y), f, Z_{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.

A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

- Learning Method
  - In a **fully supervised scenario**, a true solution $\bar{z}$ is given, so θ can be learned by optimizing NLL of $\bar{z}$ given input x with respect to θ.

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

  - In a **weakly supervised setting**, $\bar{z}$ is not given, and instead, input x and $Z = \{z_1, z_2, \ldots, z_n\}$ are given.
  - We can compute the **maximum marginal likelihood(MML)** estimate, which marginalizes the likelihood of each $z_i \in Z$ given input x with respect to θ.

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

  - However, the MML can maximize other z that are spurious solutions, and discrepancy can occur during training and testing when optimizing every z.

  - In **hard EM approach**, model first computes the likelihood of each z_i given input x with respect to θ, and picks one of Z with largest likelihood.
  - Then, optimize on a standard negative log likelihood objective, assuming $\bar{z}$ is a true solution.

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

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

A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

- Learning Method
  - For **Multi-Mention Reading Comprehension**, **Reading Comprehension with Discrete Reasoning** and **SQL Query Generation**, the authors first obtain a pre-computed solution set Z based on input x and output y.

$$
\begin{aligned}
g_{\max} &= \max_{1 \le s_i \le e_i \le L} g([d_{s_i}, \ldots, d_{e_i}], y) \\
Z &= \{z_i = (s_i, e_i) \text{ s.t. } g(s_i, e_i) = g_{\max}\},
\end{aligned}
$$

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

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

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

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

A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

- 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	

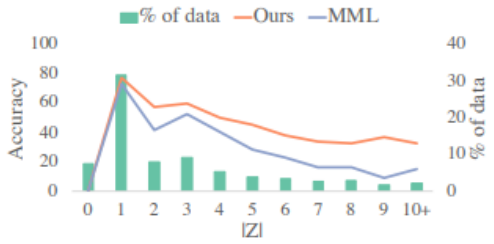
Table 3: **Results on multi-mention reading comprehension & discrete reasoning tasks.** We report performance on five datasets with different base models. Note that we are not able to obtain the test result on the subset DROP_{num}. Previous state-of-the-art are from Wang et al. (2018), Nishida et al. (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.

Model	Accuracy Dev	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**

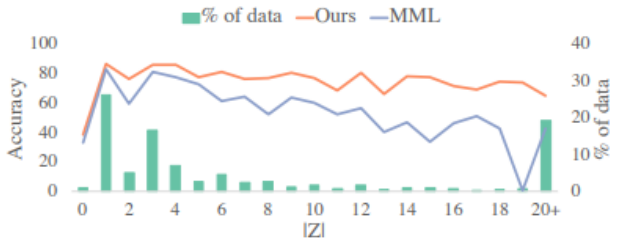
Table 4: **Results on WikiSQL.** We compare accuracy with weakly-supervised or fully-supervised settings. Our method outperforms previous weakly-supervised methods and most of published fully-supervised methods.

A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

- Analysis / Ablation Studies
  - **(Varying the size of solution set at inference time)**
  - The gap between MML and hard-EM method is marginal when $|Z| = 0$ *or* 1, and gradually increases as $|Z|$ grows.
  - **(Varying the size of solution set at training)**
  - Hard-EM approach outperforms MML consistently over different values of $|Z|$, and gain is particularly large when $|Z| > 3$.

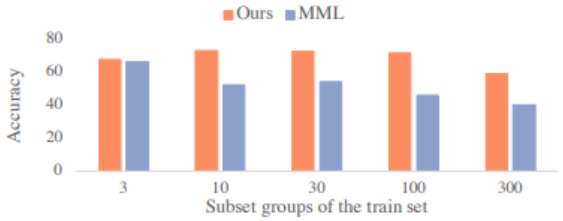


(a) DROP_{num}



(b) WikiSQL

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



A Discrete Hard EM Approach for Weakly Supervised Question Answering [EMNLP 2019]

- Analysis / Ablation Studies
  - **(Model predictions over training)**
  - Analyze the top-1 prediction and the likelihood of $z \in Z$ assigned by the model on dataset with different number of training iterations.
  - Model first begins by assigning a small, uniform probability distribution to Z, but **gradually learns to favor the true solution**.
  - (Quality of predicted solution)
  - Analyze if the model outputs the correct solution, since the solution executing the correct answer could be spurious.
  - NARRATIVEQA : 98%, DROP : 92%, WIKISQL : 88.5%
  - **(Robustness to the noise in $|Z|$)**
  - Sometimes noise arises during construction of $|Z|$, and explore the effect of noise in Z.
  - By picking all the spans with scores that is equal to or largest than the 5th highest.
  - MML drops significantly(56.07->51.14), while Hard-EM drops marginally(58.77->57.97), meaning that **Hard-EM is more robust**.

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

UNIFIEDQA: Crossing Format Boundaries with a Single QA System

Daniel Khashabi[1] **Sewon Min**[2] **Tushar Khot**[1] **Ashish Sabharwal**[1]
Oyvind Tafjord[1] **Peter Clark**[1] **Hannaneh Hajishirzi**[1,2]

[1]Allen Institute for AI, Seattle, U.S.A.
[2]University of Washington, Seattle, U.S.A.

EMNLP 2020

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- QA tasks have been posed using a **variety of formats**, and no single QA system targets all of these formats
  - The differences in format have motivated study in silos(kind of a bubble), often encoding QA format into the model architecture itself.
  - Can QA models learn linguistic reasoning abilities that generalize across format?
  - While question format and relevant knowledge may vary, the **underlying linguistic understanding and reasoning abilities** are largely common.

  - UNIFIEDQA exploits information across 4 different QA formats to achieve strong performance across 20 different factoid and commonsense QA datasets.
  - Crossing QA format boundaries is not only qualitatively desirable but also quantitatively beneficial.

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

Extractive [SQuAD]

Question: At what speed did the turbine operate?
Context: (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
Gold answer: 16,000 rpm

Abstractive [NarrativeQA]

Question: What does a drink from narcissus's spring cause the drinker to do?
Context: Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...
Gold answer: fall in love with themselves

Multiple-Choice [ARC-challenge]

Question: What does photosynthesis produce that helps plants grow?
Candidate Answers: (A) water (B) oxygen (C) protein (D) sugar
Gold answer: sugar

Yes/No [BoolQ]

Question: Was America the first country to have a president?
Context: (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
Gold answer: no

Figure 1: Four formats (color-coded throughout the paper) commonly used for posing questions and answering them: Extractive (EX), Abstractive (AB), Multiple-Choice (MC), and Yes/No (YN). Sample dataset names are shown in square brackets. We study generalization and transfer across these formats.

Datasets	SQuAD11	SQuAD2	NewsQA	Quoref	ROPES	NarQA	DROP	NatQA	RACE	MCTest	OBQA	ARC	QASC	CQA	WG	PIQA	SIQA	BoolQ	NP-BoolQ	MultiRC
Format	Extractive QA (EX)					Abstractive QA (AB)			Multiple-choice QA (MC)									Yes/NO QA (YN)		
Has paragraphs?	✓	✓	✓	✓	✓	✓	✓		✓	✓								✓	✓	✓
Has explicit candidate ans?									✓	✓	✓	✓	✓	✓	✓	✓	✓			
# of explicit candidates									4	4	4	4	8	5	2	2	3			
Para contains ans as substring?	✓	✓	✓	✓																
Has idk questions?		✓																		

Figure 2: Properties of various QA datasets included in this study: 5 extractive (EX), 3 abstractive (AB), 9 multiple-choice (MC), and 3 yes/no (YN). 'idk' denotes 'I don't know' or unanswerable questions. BoolQ represents both the original dataset and its *contrast-sets* extension BoolQ-CS; similarly for ROPES, Quoref, and DROP.

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- Multi-Format Training
  - A unified QA model can operate over k $formats$ $F_1, F_2, \ldots, F_k$ where each format F_i have l_i different datasets $D_1^i, D_2^i, \ldots, D_{l_i}^i$ where $D_j^i = (T_j^i, E_j^i)$.
  - If T_j^i is empty, it is an evaluation-only dataset and is used to see how well a model generalizes to such format or QA dataset overall.

  - To use the **text-to-text paradigm**, convert each training question q in format F_i into a plain-text input representation $enc_i(q)$.
  - We then obtain a mixed training pool consisting of all available training instances : $\tilde{T} = \cup_{i=1}^{k} \cup_{j=1}^{l_i} \{ enc_i(q) | \, q \in T_j^i \}$.

  - In **Multiple-Choice(MC) datasets**, the input is encoded as "$question$ \n (A) c_1 (B) c_2 (C) c_3 (D) c_4 \n $paragraph$".
  - In other datasets(**Extractive, Abstractive, Yes/No**), input is encoded as "$question$ \n $paragraph$".
  - Unlike T5, we do not specify any task-, dataset-, or format-specific prefixes in the input representation.

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- Training Procedure / Evaluation Metric
  - Empirically choose **8 seed datasets** for training UNIFIEDQA, based on their effectiveness in our pilot study.
  - **Evaluate on 20 existing datasets** that target different formats as well as various complex linguistic phenomena.

    - EX: SQuAD 1.1, SQuAD 2.0
    - AB: NarrativeQA
    - MC: RACE, ARC, OBQA, MCTest
    - YN: BoolQ

  - **EX format** : F1 score of the extracted span relative to the gold label.
  - **AB format** : ROUGE-L metric (For NatQA, use EM metric)
  - **MC format** : match the generated text with the closest answer candidate based token overlap and compute accuracy.
  - **YN format** : measure if the generated output matches the correct 'yes' or 'no' label. (If output is longer than one word, check if it contains the correct label but not the incorrect one)

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

EX	**Dataset**	SQuAD 1.1
	Input	`At what speed did the turbine operate? \n (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...`
	Output	`16,000 rpm`
AB	**Dataset**	NarrativeQA
	Input	`What does a drink from narcissus's spring cause the drinker to do? \n Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to ``Grow dotingly enamored of themselves.'' ...`
	Output	`fall in love with themselves`
MC	**Dataset**	ARC-challenge
	Input	`What does photosynthesis produce that helps plants grow? \n (A) water (B) oxygen (C) protein (D) sugar`
	Output	`sugar`
	Dataset	MCTest
	Input	`Who was Billy? \n (A) The skinny kid (B) A teacher (C) A little kid (D) The big kid \n Billy was like a king on the school yard. A king without a queen. He was the biggest kid in our grade, so he made all the rules during recess. ...`
	Output	`The big kid`
YN	**Dataset**	BoolQ
	Input	`Was America the first country to have a president? \n (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...`
	Output	`no`

Table 1: Example text-to-text encoding of instances.

Dataset	Train set size	Eval. set size	Best published	95% CI (%)	Input length	Output length
SQuAD 1.1	87k	10k	95.6	0.4	136.2	3.0
SQuAD 2.0	130k	11k	91.2	0.5	139.9	2.6
NewsQA	76k	4k	66.8	1.4	606.6	4.0
Quoref	22k	2k	86.1	1.5	352.7	1.7
Quoref-CS	-	700	55.4	3.6	324.1	2.2
ROPES	10k	1.4k	61.1	2.5	169.1	1.4
ROPES-CS	-	974	32.5	3.0	182.7	1.3
NarQA	65k	21k	58.9	0.7	563.6	6.2
NatQA	79k	3.6k	42.2	1.6	607.0	2.2
DROP	77k	9k	89.1	0.6	189.1	1.6
DROP-CS	-	947	54.2	3.2	206.0	2.1
RACE	87k	4k	89.5	0.9	317.9	6.9
OBQA	4k	501	80.0	3.3	28.7	3.6
MCTest	1.4k	320	86.5	3.4	245.4	4.0
ARC (easy)	2k	2k	80.0	1.7	39.4	3.7
ARC (chal.)	1k	1k	67.8	2.9	47.4	5.0
CQA	9.7k	1.2k	79.1	2.2	26.8	1.5
WG	40.3k	1.7k	67.5	2.2	25.2	3.0
PIQA	16.1k	3k	79.4	1.4	49.6	20.2
SIQA	33.4k	2.2k	78.0	1.7	37.3	4.7
BoolQ	9k	3k	91.0	1.0	105.1	1.0
BoolQ-CS	-	461	71.1	4.0	108.9	1.0
NP-BoolQ	10k	3k	78.4	1.4	106.2	1.0
MultiRC	-	312	91.7	2.6	293.3	1.0

Table 2: Dataset Statistics. CQA, OBQA, WG, and NarQA refer to CommonsenseQA, OpenBookQA, Winogrande, and NarrativeQA, respectively. The CI column shows the upper 95% confidence interval for the evaluation set as a percentage, based on the Wilson test around the mean score listed as a percentage in the best known performance column. Input and output representation lengths are measured in the number of tokens and averaged across the dataset.

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- Pilot Study : Can Out-of-Format Training Help?
  - Is the broad idea of benefiting from out-of-format training even viable?
  - For instance, is our intuition correct that an MC dataset can, in practice, benefit from training on an EX dataset?

  - Given a training set T_1^i (anchor dataset) of format F_i, is there an out-of-format training set T_1^j of format F_j such that training jointly on $T_1^i \cup T_1^j$ improves performance relative to training only on T_1^i?
  - This should be tested by evaluating both cases on evaluation set E_1^j as well as $E_2^j, E_3^j, \ldots$.

  - We generally observe that there is at least one out-of-format training set whose inclusion improves performance.
  - This pilot study thus provides a proof of concept that **out-of-format training can indeed help a QA model** in nearly every case.

Trained on ↓ - Evaluated on →	SQuAD11	SQuAD2	NewsQA	Quoref	Quoref-CS
SQuAD11	**85.9**	42.8	51.7	28.2	28.11
SQuAD11 + X	85.8	42.8	**52.1**	**29.4**	**29.84**
Best X	BoolQ	OBQA	OBQA	NarQA	OBQA

Trained on ↓ - Evaluated on →	RACE	OBQA	ARC-chal	MCTest
RACE	55.8	26.6	28.0	62.5
RACE + X	**59.1**	**32.2**	**28.4**	**69.4**
Best X	SQuAD11	NarQA	NewsQA	SQuAD11

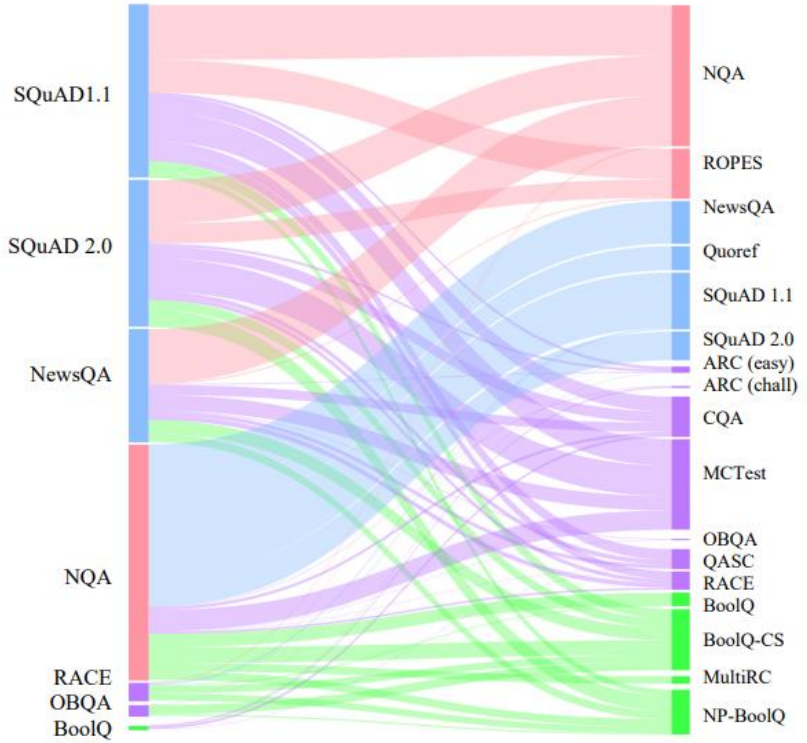
Trained on ↓ - Evaluated on →	BoolQ	MultiRC	NP-BoolQ	BoolQ-CS
BoolQ	76.4	64.1	51.3	53.4
BoolQ + X	**78.9**	**66.0**	**59.4**	**61.0**
Best X	SQuAD2	OBQA	SQuAD2	NarQA

Trained on ↓ - Evaluated on →	NarQA	DROP	DROP-CS
NarQA	51.5	10.2	11.1
NarQA + X	**53.0**	**14.4**	**14.6**
Best X	SQuAD2	SQuAD2	SQuAD2

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- Pilot Study : Can Out-of-Format Training Help?
  - There are strong connections between AB and EX datasets
  - NQA is most helpful dataset and even helps multiple formats.
  - Similarly, extractive datasets(SQuAD1.0, SQuAD 2.0, NewsQA) are also relatively more helpful.
  - RACE doesn't help that much.
  - Least helpful dataset in the mix is BoolQ.

  - From figure, (Left Side) Training dataset, (Right Side) Evaluation dataset.
  - The wider the edge from left dataset to a right dataset, the higher the contribution.



UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- Results (UnifiedQA vs 8 Dedicated Models)
  - Earlier works have observed that a system addressing multiple tasks often underperforms a focused system. (*Raffel et al., 2020*)
  - UNIFIEDQA performs almost as good as individual T5 models targeted to each dataset, and even performs better than the single-dataset experts.
  - UNIFIEDQA offers **flexibility** across multiple QA formats while compromising almost nothing compared to dataset-specific experts.

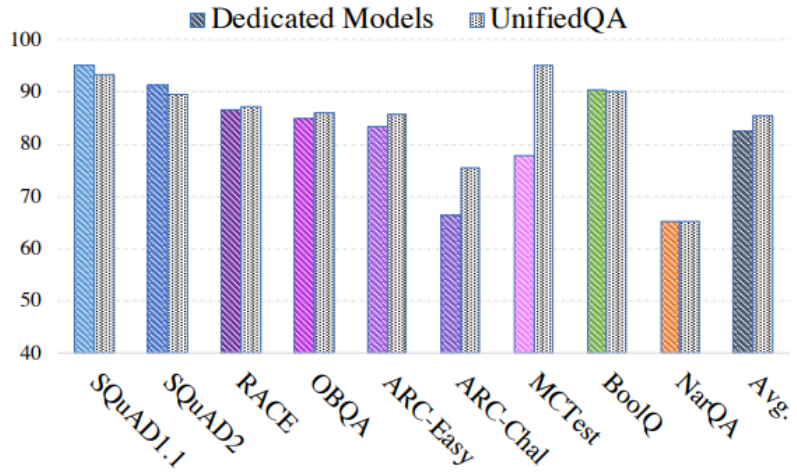


Figure 4: UNIFIEDQA is on-par with, and often outperforms, 9 different equally-sized T5-based systems tailored to individual datasets. The figure contains separate models for each of the two subsets of the ARC and Regents datasets.

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- Results (Generalization to Unseen Datasets)
  - Out on 9(out of 12) datasets, UNIFIEDQA shows a **better generalization** than any single-format expert.
  - One must be careful when reading these numbers as the best previous numbers follow the fully supervised protocol.

Seen dataset?	Model ↓ - Evaluated on →	NewsQA	Quoref	Quoref-CS	ROPES	ROPES-CS	DROP	DROP-CS	QASC	Common senseQA	NP-BoolQ	BoolQ-CS	MultiRC	Avg
No	UnifiedQA [EX]	58.7	64.7	53.3	43.4	29.4	24.6	24.2	55.3	62.8	20.6	12.8	7.2	38.1
	UnifiedQA [AB]	58.0	**68.2**	57.6	48.1	41.7	30.7	36.8	54.1	59.0	27.2	39.9	28.4	45.8
	UnifiedQA [MC]	48.5	67.9	**58.0**	61.0	44.4	28.9	37.2	67.9	75.9	2.6	5.7	9.7	42.3
	UnifiedQA [YN]	0.6	1.7	1.4	0.0	0.7	0.4	0.1	14.8	20.8	79.1	78.6	**91.7**	24.2
	UnifiedQA	**58.9**	63.5	55.3	**67.0**	**45.5**	**32.5**	**40.1**	**68.5**	**76.2**	**81.3**	**80.4**	59.9	**60.7**
Yes	Previous best	66.8	86.1	55.4	61.1	32.5	89.1	54.2	85.2	79.1	78.4	71.1	--	
		Retro Reader	TASE	XLNet	ROBERTa	RoBERTa	ALBERT	MTMSN	KF+SIR+2Step	reeLB-RoBERT	RoBERTa	RoBERTa	--	

Table 4: Generalization to unseen datasets: Multi-format training (UNIFIEDQA) often outperforms models trained the same way but solely on other in-format datasets (e.g., UNIFIEDQA [EX], which is trained on all extractive training sets of UNIFIEDQA. When averaged across all evaluation datasets (last column), UNIFIEDQA shows strong generalization performance across all formats. Notably, the "Previous best" models (last row) were trained on the target dataset's training data, but are even then outperformed by UnifiedQA (which has never seen these datasets during training) on the YN tasks.

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- Results (State-of-the-Art via Simple Fine-tuning)
  - When it comes to QA, is there a value in using UNIFIEDQA as a starting point for fine-tuning, as opposed to a vanilla LM that has not seen other QA datasets before?
  - Finetuning on UNIFIEDQA consistently dominates fine-tuning on T5 and BART, respectively.
  - It also dominates the best previous scores on the datasets.
  - The highlights the effectiveness of **cross-format training** is not limited only to T5, but is rather a general trend for text-to-text architectures.

Model ↓ - Eval. →	OBQA *	OBQA (w/ IR)	ARC-easy *	ARC-easy (w/ IR)	ARC-chal *	ARC-chal (w/ IR)	QASC	QASC (w/ IR)
Previous best published	RoBERTa (Clark et al.,2019c)	KF+SIR (Mitra et al., 2020)	RoBERTa (Clark et al.,2019c)	FreeLB-RoBERTa (Zhu et al., 2020)	RoBERTa (Clark et al.,2019c)	FreeLB- RoBERTa (Zhu et al., 2020)	--	KF+SIR +2Step (Mitra et al., 2020)
	75.7	80.0	69.9	80.0	55.9	67.8	--	85.2
$\text{BART}_{\text{large}}$ - FT	67.8	66.2	64.1	79.6	36.6	40.4	50.0	75.3
UnifiedQA$_{\text{BART}}$ - FT	63.8	70.0	68.0	82.7	52.1	55.0	53.2	78.2
T5 - FT	84.2	84.2	83.8	90.0	65.4	69.7	77.0	88.5
UnifiedQA - FT	**86.0**	**87.2**	**86.4**	**92.0**	**75.0**	**78.5**	**78.5**	**89.6**

Model ↓ - Eval. →	RACE *	ComQA	WG	PIQA	SIQA	ROPES	NatQ (w/ IR)
Previous best published	ALBERT (Lan et al.,2019)	FreeLB- RoBERTa (Zhu et al.,2020)	RoBERTa (Sakaguchi et al.,2019)	RoBERTa (Bisk et al., 2019)	RoBERTa (Mitra et al., 2020)	RoBERTa (Lin et al., 2019)	DPR+BART (Min et al.,2020)
	89.5	72.2	67.5	79.4	78.0	61.1	42.2
$\text{BART}_{\text{large}}$ - FT	78.8	62.5	62.4	77.4	74.0	60.5	42.1
UnifiedQA$_{\text{BART}}$ - FT	79.4	64.0	63.6	77.9	73.2	60.0	44.5
T5 - FT	87.1	78.1	84.9	88.9	**81.4**	74	**49.3**
UnifiedQA - FT	89.4	**79.1**	**85.7**	**89.5**	**81.4**	**75.2**	**49.3**

Table 5: Fine-tuning UNIFIEDQA (last row) results in new state-of-the-art performance on 11 datasets. Further, it consistently improves upon fine-tuned T5 (2nd last row) by a margin ranging from 1% for CommonsenseQA (CQA) to as much as 13% for ARC-challenge. '(w/ IR)' denotes relevant information is retrieved and appended as context sentences in the input encoding. Datasets marked with * are used in UNIFIEDQA's original training.

UNIFIEDQA : Crossing Format Boundaries with a Single QA System [EMNLP 2020]

- Ablation Studies (Training Set Contribution)
  - We take a system and assess how strong the model is when individual seed training datasets are dropped from the union.
  - BoolQ, SQuAD 2.0, OBQA, NarQA are the top-4 contributing datasets, each with a different format.
  - SQuAD1.0 has the least importance, presumably because it is mostly covered by SQuAD 2.0.
  - This study suggests that in order to build an effective unified QA system, it suffices to have a relatively small set of datasets as long as the set includes representatives from each format.

Model ↓ - Evaluated on →	SQuAD11	SQuAD2	NarQA	RACE	OBQA	ARC-easy	ARC-hard	MCTest	BoolQ	Avg	Δ
UnifiedQA	93.4	89.6	65.2	87.3	86.0	85.7	75.6	95.0	90.2	85.4	
excluding BoolQ	93.1	90.1	65.0	87.7	85.0	86.1	75.2	94.7	8.3	77.0	-8.4
excluding SQuAD 2	95.3	47.3	65.4	87.7	84.8	85.9	75.5	95.3	90.5	81.3	-4.2
excluding OBQA	93.6	89.3	65.2	87.4	77.8	85.7	74.0	94.7	90.1	84.2	-1.3
excluding NarQA	93.6	89.8	52.5	87.7	85.6	86.3	75.9	95.6	89.9	84.2	-1.2
excluding RACE	93.9	89.0	65.0	78.5	85.2	85.6	74.7	95.9	90.1	84.3	-1.2
excluding ARC-easy	93.4	89.8	65.0	87.0	83.8	84.0	75.9	94.7	89.9	84.9	-0.6
excluding ARC-hard	93.6	90.1	64.9	87.3	85.2	85.1	73.8	95.6	90.5	85.1	-0.4
excluding MCTest	92.8	90.6	65.0	87.1	84.6	85.6	75.4	95.6	90.2	85.2	-0.2
excluding SQuAD 1.1	92.6	90.3	65.3	87.4	85.8	86.5	75.9	95.3	90.7	85.6	0.1

Table 6: The results of a leave-one-out ablation. The first row indicates the performance of UNIFIEDQA on each dataset it was trained on. The rest of the rows exclude one dataset at a time. The rows are sorted based on the last column: the dataset with biggest contribution appear first. The red highlights indicate the top 3 performance drops for each column.

ANY QUESTIONS?