Know What You Don't Know: Unanswerable Questions for SQuAD

Pranav Rajpurkar*, Robin Jia*, and Percy Liang
Stanford University

SQuAD2.0

The Stanford Question Answering Dataset

Pranav Rajpurkar*, Robin Jia*, and Percy Liang
Stanford University

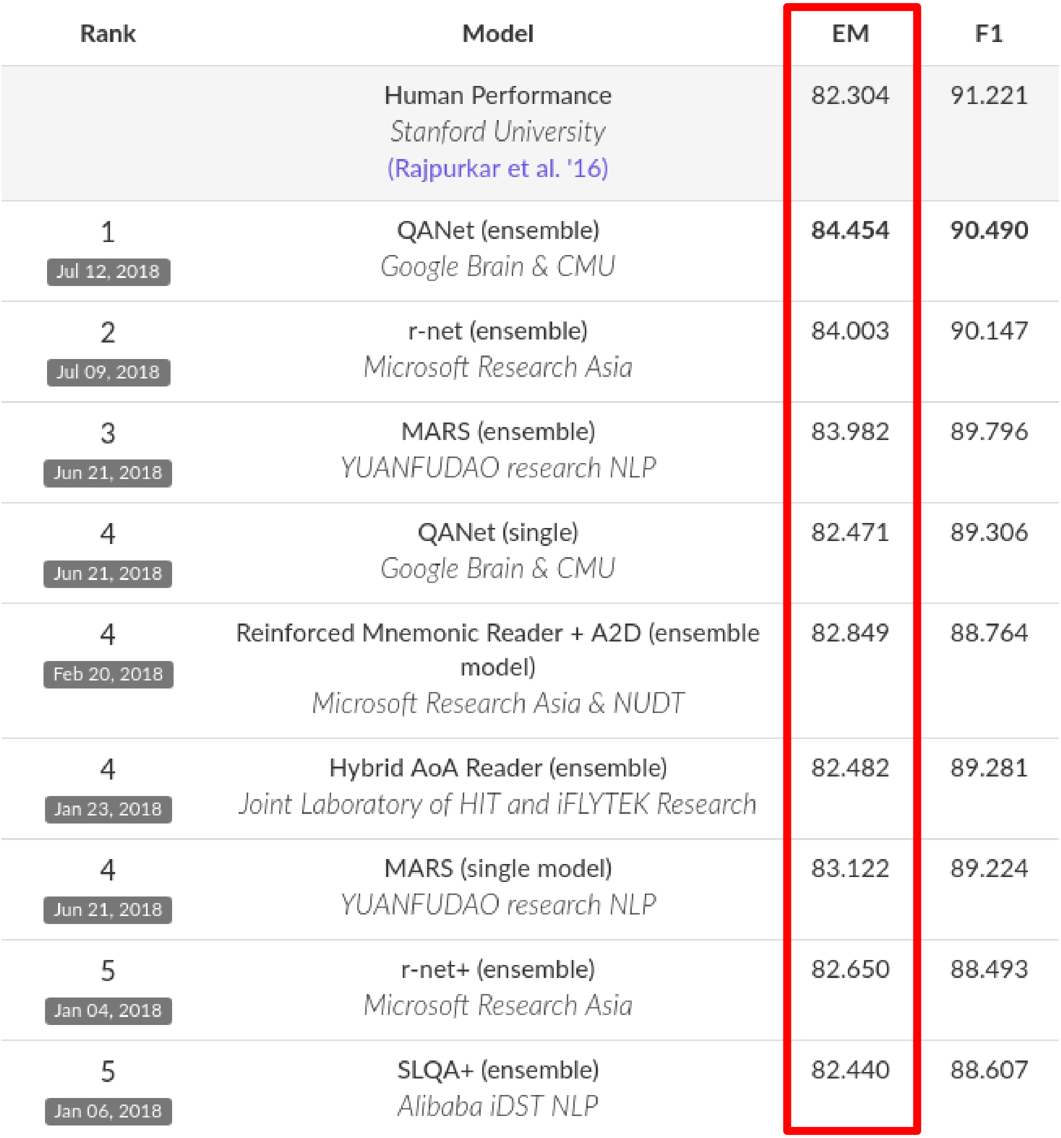
SQuAD (Rajpurkar et al., 2016)

Paragraph: *Victoria is a state in south-eastern Australia…Most of its population is concentrated in the area surrounding…its state capital and largest city, Melbourne…*

Question: *What city is the capital of Victoria?*

Answer: ***Melbourne***

Human-level abilities?



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Jul 12, 2018	QANet (ensemble) *Google Brain & CMU*	**84.454**	**90.490**
2 Jul 09, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
3 Jun 21, 2018	MARS (ensemble) *YUANFUDAO research NLP*	83.982	89.796
4 Jun 21, 2018	QANet (single) *Google Brain & CMU*	82.471	89.306
4 Feb 20, 2018	Reinforced Mnemonic Reader + A2D (ensemble model) *Microsoft Research Asia & NUDT*	82.849	88.764
4 Jan 23, 2018	Hybrid AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.482	89.281
4 Jun 21, 2018	MARS (single model) *YUANFUDAO research NLP*	83.122	89.224
5 Jan 04, 2018	r-net+ (ensemble) *Microsoft Research Asia*	82.650	88.493
5 Jan 06, 2018	SLQA+ (ensemble) *Alibaba iDST NLP*	82.440	88.607

A new challenge

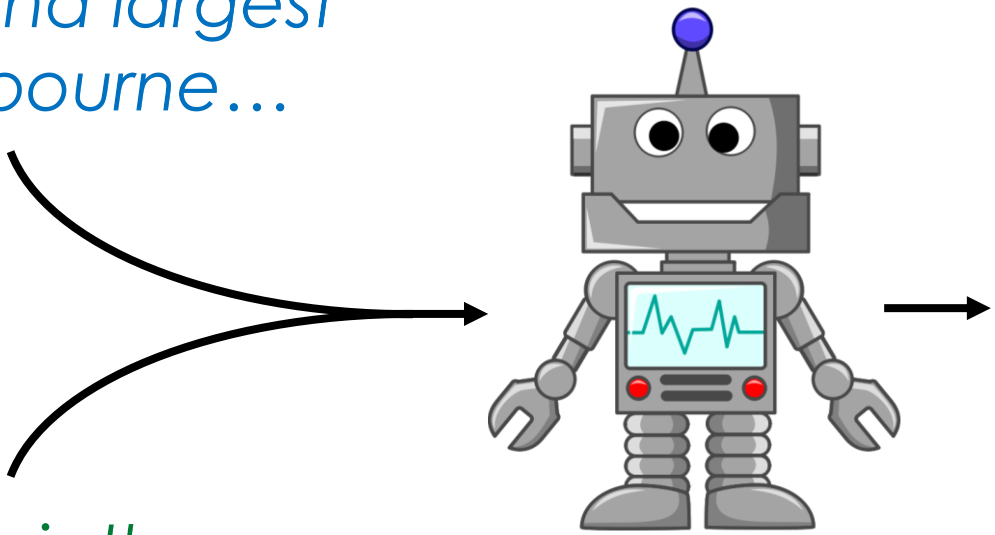
Paragraph: *Victoria is a state in south-eastern Australia…Most of its population is concentrated in the area surrounding…its state capital and largest city, Melbourne…*

Question: *What city is the capital of **Australia**?*

Answer: **`<No Answer>`**

SQuAD 2.0

Victoria's state capital and largest city, Melbourne…

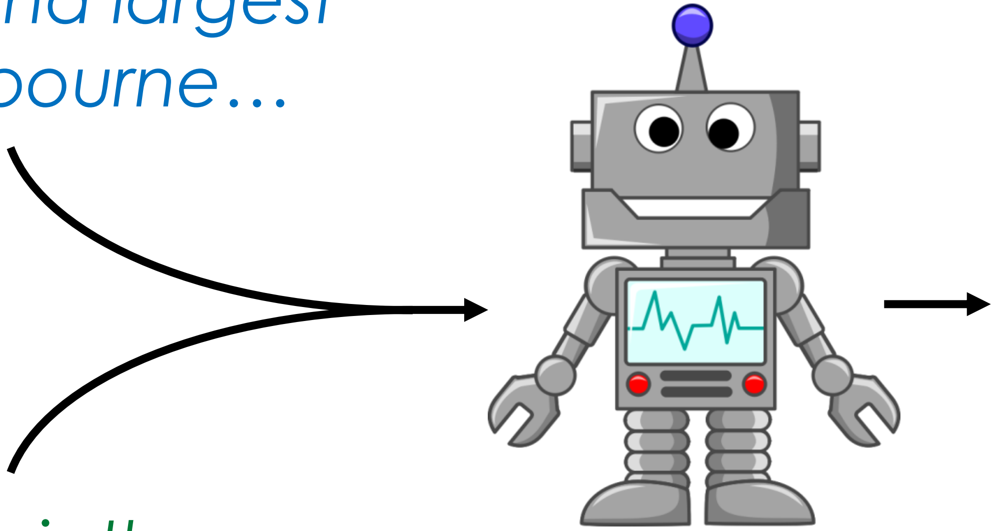


Melbourne!

What city is the capital of Victoria?

SQuAD 2.0

Victoria's state capital and largest city, Melbourne…



What city is the capital of ***Australia****?*

No answer!

Outline

- Why unanswerable questions?
- SQuAD 2.0
- Baseline systems, baseline datasets

Outline

- Why unanswerable questions?
- SQuAD 2.0
- Baseline systems, baseline datasets

Adversarial evaluation

Question: *The number of new Huguenot colonists declined after what year?*

Paragraph: *The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as 1700; thereafter, the numbers declined.*

Correct Answer: *1700*

Jia and Liang (2017)

Adversarial evaluation

Question: *The number of new Huguenot colonists declined after what year?*

Paragraph: *The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as 1700; thereafter, the numbers declined. The number of **old Acadian** colonists declined after the year of 1675.*

Correct Answer: *1700*

Predicted Answer: *1675*

Jia and Liang (2017)

A simpler adversary

Question: *The number of **old Acadian** colonists declined after what year?*

Paragraph: *The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689…but quite a few arrived as late as 1700; thereafter, the numbers declined.*

Correct Answer: **`<No Answer>`**

Predicted Answer: *1700*

Relation Extraction as QA

Relation query: `educated_at(AlbertEinstein, ?)`

Question: *Albert Einstein was a student at what school?*

Paragraph: *Albert Einstein was awarded a PhD by the* ***University of Zurich****, with his dissertation titled…*

Answer: ***University of Zurich***

Levy et al. (2017)

Relation Extraction as QA

Relation query: `educated_at(AlbertEinstein, ?)`

Question: *Albert Einstein was a student at what school?*

Paragraph: *Einstein became a full professor at the German* ***Charles-Ferdinand University*** *in Prague…*

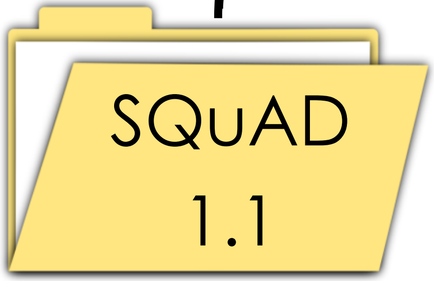
Answer: **`<No Answer>`**

Levy et al. (2017)

Outline

- Why unanswerable questions?
- SQuAD 2.0
- Baseline systems, baseline datasets

Data collection



Victoria's capital city, Melbourne, is Australia's second-largest city.

Inspiration questions:

- *Compared to other Australian cities, what is the size of Melbourne?*

New questions:

- *How populous is Melbourne compared to other Australian **states**?*
  - Plausible answer: *second-largest*

Crowdworker

Data summary

Property	SQuAD 1.1	SQuAD 2.0
Total size	108k	**151k**

Data summary

Property	SQuAD 1.1	SQuAD 2.0
Total size	108k	**151k**
Unanswerable questions at test time	0%	**48.9%**

Some unanswerable questions

Paragraph: *Typically, ministers or party leaders open debates, with* ***opening*** *speakers given between 5 and 20 minutes, and succeeding speakers allocated less time.*

Question: ***Closing*** *speakers are given between 5 and how many minutes?*

Category: Antonym (20%)

Some unanswerable questions

Paragraph: ***Newton****'s Law of Gravitation states that the force on a spherical object of mass due to the gravitational pull of mass is…*

Question: ***Cavendish****'s Law of Gravitation states what?*

Category: Entity Swap (21%)

Some unanswerable questions

Paragraph: *Dendritic cells…are named for their **resemblance to neuronal dendrites**, as both have many spine-like projections…*

Question: *What is named for its **resemblance to dendritic cells**?*

Category: Mutual Exclusion (15%)

Some unanswerable questions

Paragraph: *The Malkin Athletic Center…includes two cardio rooms, an* ***Olympic****-size swimming pool, …*

Question: *At what building do* ***Olympic*** *athletes train?*

Category: Neutral (24%)

Human validation

Victoria's state capital and largest city, Melbourne…

What city is the capital of Australia?

Votes from multiple crowdworkers

→ **No answer!**

Human validation

- Human test accuracy: **86.9% Exact**, **89.5% F1**
- People **can** do well on this dataset (if they're careful)

Outline

- Why unanswerable questions?
- SQuAD 2.0
- Baseline systems, baseline datasets

Baseline systems

- Three existing SQuAD systems that can be made to predict **`<No Answer>`**
  - BiDAF-No-Answer (Levy et al., 2017)
  - DocumentQA (Clark and Gardner, 2018)
  - DocumentQA + ELMo (Peters et al., 2018)

Baseline systems

System	SQuAD 1.1	SQuAD 2.0
“No answer” baseline	-	48.9

Test set F1 scores

Baseline systems

System	SQuAD 1.1	SQuAD 2.0
“No answer” baseline	-	48.9
BiDAF-No-Answer	77.3	62.1
DocumentQA	81.0	62.3
DocumentQA + ELMo	**85.8**	**66.3**

Test set F1 scores

Baseline systems

System	SQuAD 1.1	SQuAD 2.0
“No answer” baseline	-	48.9
BiDAF-No-Answer	77.3	62.1
DocumentQA	81.0	62.3
DocumentQA + ELMo	**85.8**	**66.3**
Human	91.2	89.5

Test set F1 scores

Baseline systems

System	SQuAD 1.1	SQuAD 2.0
“No answer” baseline	-	48.9
BiDAF-No-Answer	77.3	62.1
DocumentQA	81.0	62.3
DocumentQA + ELMo	**85.8**	**66.3**
Human	91.2	89.5
Human-Machine Gap	5.4	**23.2**

Test set F1 scores

Guessing answerability

- Can you guess that a question is unanswerable **without reading the paragraph**?

See e.g. Gururangan et al. (2018), Poliak et al. (2018)

Guessing answerability

System	Binary Classification Accuracy
Majority baseline	50.1
Question only	
Fasttext (Joulin et al., 2017)	60.2
Linear SVM with 1,2,3-grams	60.9

Development set

Guessing answerability

System	Binary Classification Accuracy
Majority baseline	50.1
Question only	
Fasttext (Joulin et al., 2017)	60.2
Linear SVM with 1,2,3-grams	60.9
Question + Context	
BiDAF-No-Answer	68.0
DocumentQA	70.1
DocumentQA + ELMo	72.0

Development set

Signs of unanswerability

- Negation words (“never”, “n’t”, “not”)
- Antonyms of common question words (“least”, “smallest”, “last”)
- In many cases, features are rare (<1% frequency) but do provide strong signal

Baseline datasets

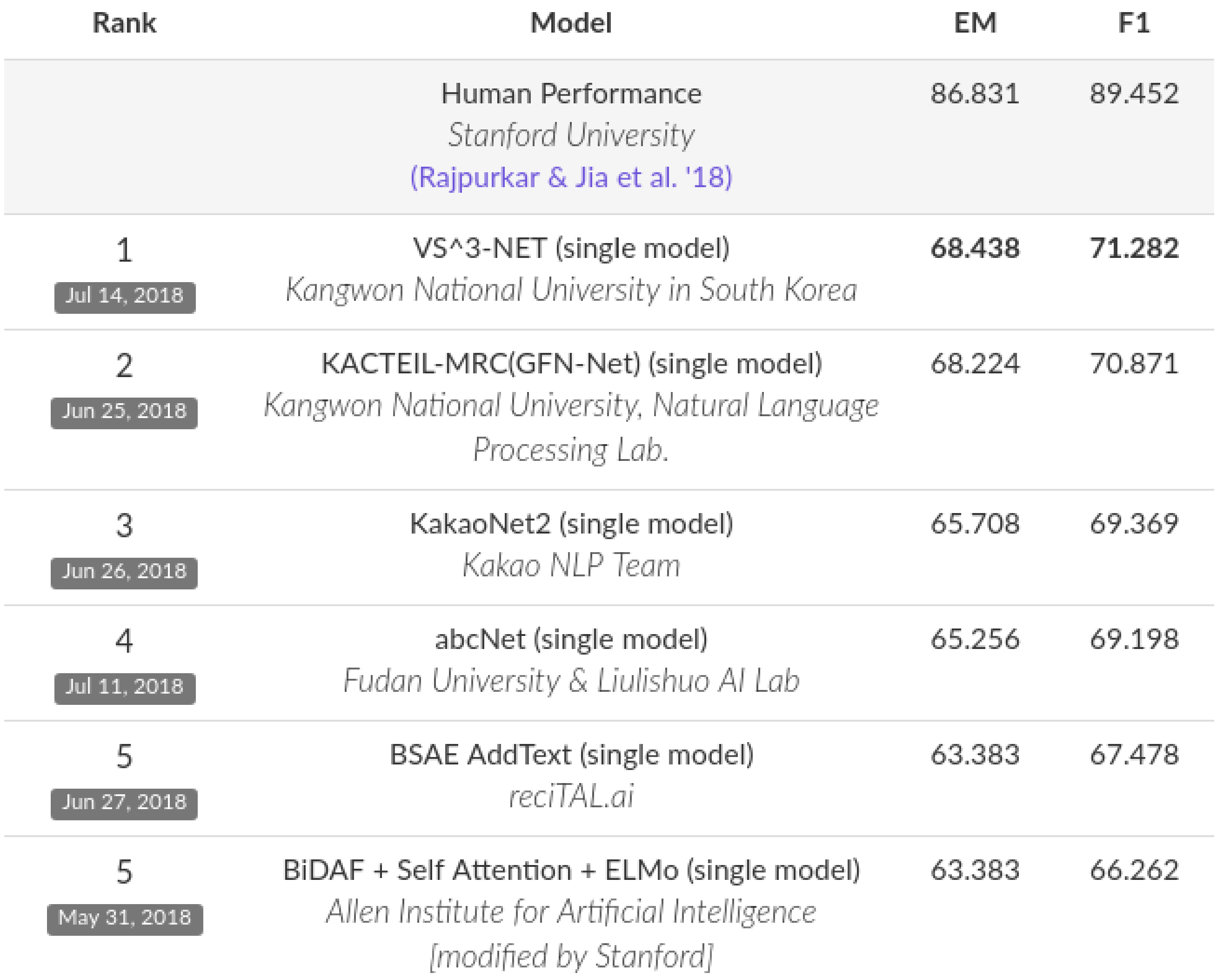
- Was all this effort necessary to make a challenging dataset?
- Automatically generated unanswerable questions
  - TF-IDF-based (Clark and Gardner, 2018)
  - Rule-based (Jia and Liang, 2017)

Baseline datasets

System	SQuAD 1.1 + TF-IDF	SQuAD 1.1 + Rule-based	SQuAD 2.0
BiDAF-No-Answer	76.6	84.8	62.6
DocumentQA	79.2	84.8	64.8
DocumentQA + ELMo	**83.0**	**89.6**	**67.6**

Development set F1 scores

Live leaderboard



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 14, 2018	VS^3-NET (single model) *Kangwon National University in South Korea*	**68.438**	**71.282**
2 Jun 25, 2018	KACTEIL-MRC(GFN-Net) (single model) *Kangwon National University, Natural Language Processing Lab.*	68.224	70.871
3 Jun 26, 2018	KakaoNet2 (single model) *Kakao NLP Team*	65.708	69.369
4 Jul 11, 2018	abcNet (single model) *Fudan University & Liulishuo AI Lab*	65.256	69.198
5 Jun 27, 2018	BSAE AddText (single model) *reciTAL.ai*	63.383	67.478
5 May 31, 2018	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	63.383	66.262

Thank you!

Visit stanford-qa.com

SQuAD2.0

The Stanford Question Answering Dataset

Submit models on

CodaLab