# On the "deep-learning getting near human reading-comprehension performance on SQUAD" claim

Yoav Goldberg

September 2017

(slight update early 2018)

# Near Human Performance on the SQUAD Dataset

- "Neural systems achieve near-human performance on Question Answering".

- Not really, let's see why.

# Restricted QA Setup

- Restricted to questions that can be answered by **span selection**.
- Need to find the answer in a **given paragraph**.
- The answer is **guaranteed to be** in the paragraph.

# Near Human Performance?

- Human performance: 91.2 F1.
  Current best system:  84.0 F1.

**(late addition: now approaching 89 F1.
doesn't change big picture)**

# Which humans? What are their motivations? How do they err?

- Humans on MTurk. Who are instructed to answer 5 answers in 2 minutes. That's 16 cent per question.

- Humans are wrong mostly in span boundaries.

- Also, when doing max-vote between several humans, human perf goes up substantially.

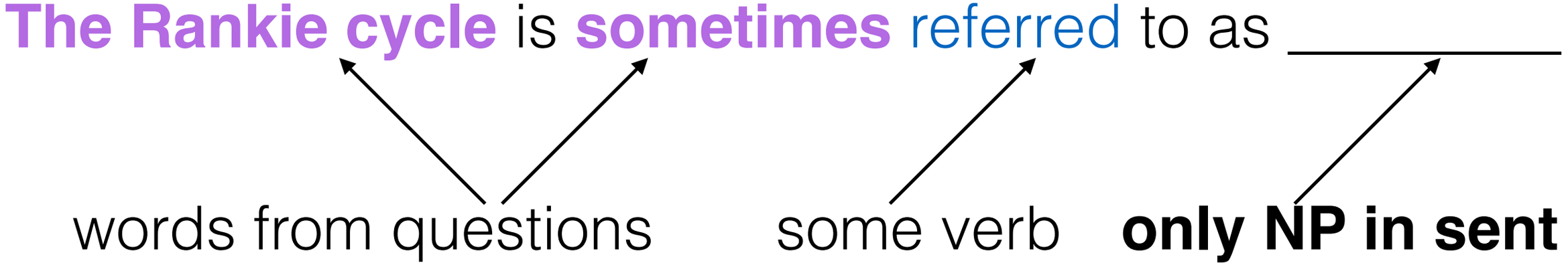
# How hard is the dataset?

- Do the questions require complex reasoning, or can they be "cheated" using superficial clues?

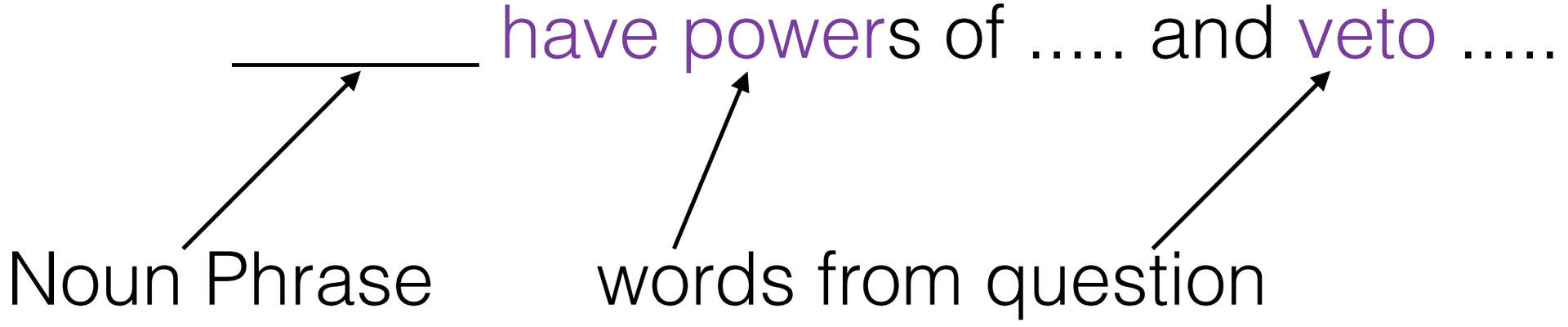
| Reasoning | Description | Example | Percentage |
|---|---|---|---|
| Lexical variation (synonymy) | Major correspondences between the question and the answer sentence are synonyms. | Q: What is the Rankine cycle sometimes **called**?<br>Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle. | 33.3% |
| Lexical variation (world knowledge) | Major correspondences between the question and the answer sentence require world knowledge to resolve. | Q: Which **governing bodies** have veto power?<br>Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process. | 9.1% |
| Syntactic variation | After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications. | Q: What Shakespeare scholar **is currently on the faculty**?<br>Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington. | 64.1% |
| Multiple sentence reasoning | There is anaphora, or higher-level fusion of multiple sentences is required. | Q: What collection does **the V&A Theatre & Performance galleries** hold?<br>Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance. | 13.6% |
| Ambiguous | We don't agree with the crowdworkers' answer, or the question does not have a unique answer. | Q: What is the main goal of criminal punishment?<br>Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment. | 6.1% |

**Table 3:** We manually labeled 192 examples into one or more of the above categories. Words relevant to the corresponding reasoning type are bolded, and the crowdsourced answer is underlined.

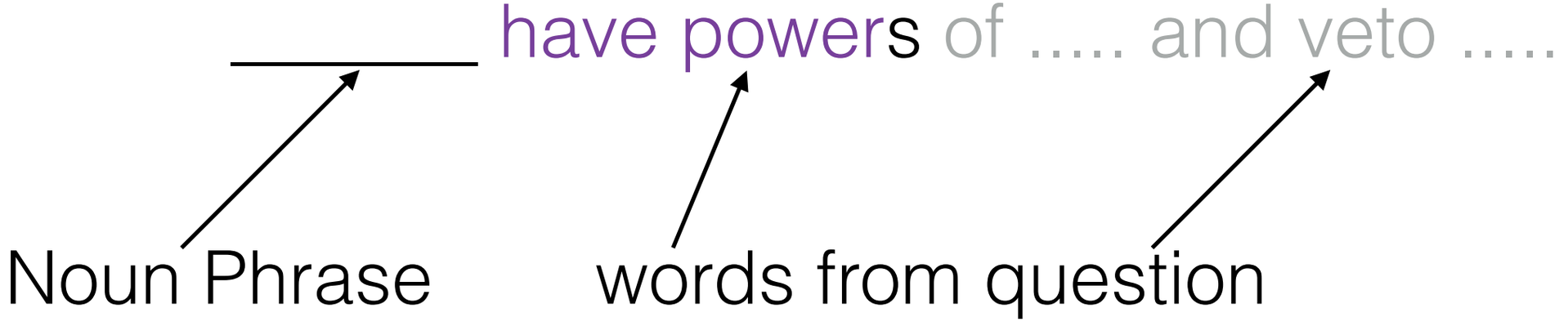
| Reasoning | Description | Example | Percentage |
|---|---|---|---|
| Lexical variation (synonymy) | Major correspondences between the question and the answer sentence are synonyms. | Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle. | 33.3% |



| | | | |
|---|---|---|---|
| Lexical variation (world knowledge) | Major correspondences between the question and the answer sentence require world knowledge to resolve. | Q: Which **governing bodies** have veto power?<br>Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process. | 9.1% |



| | | | |
|---|---|---|---|
| Lexical variation (world knowledge) | Major correspondences between the question and the answer sentence require world knowledge to resolve. | Q: Which **governing bodies** have veto power?<br>Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process. | 9.1% |



| | | | |
|---|---|---|---|
| Syntactic variation | After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications. | Q: What Shakespeare scholar **is currently on the faculty**?<br>Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington. | 64.1% |

What Shakespeare scholar is ....?

... Shakespeare scholar _____

hold

| | | | |
|---|---|---|---|
| Multiple sentence reasoning | There is anaphora, or higher-level fusion of multiple sentences is required. | Q: What collection does **the V&A Theatre & Performance galleries** hold?<br>Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance. | 13.6% |

hold

Noun Phrase

what is goal of criminal punishment

| | | | |
|---|---|---|---|
| Ambiguous | We don't agree with the crowd-workers' answer, or the question does not have a unique answer. | Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment. | 6.1% |

_____ is goal of criminal punishment

all can be "cheated away" using some smart template matching.

**Adversarial Examples for Evaluating Reading Comprehension Systems** (emnlp 2017)

**Robin Jia**
Computer Science Department
Stanford University
robinjia@cs.stanford.edu

**Percy Liang**
Computer Science Department
Stanford University
pliang@cs.stanford.edu

- It is really easy to trick the machine-learned systems by artificially creating texts that will match the patterns they expect, and appending them to the paragraph.

- Humans are not fooled by this.

**Article:** Super Bowl 50
**Paragraph:** "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
**Question:** "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
**Original Prediction:** John Elway
**Prediction under adversary:** Jeff Dean

from Jia and Liang.

The blue text was added to the paragraph
and confused the state-of-the-art learning-based system.

# To Summarize

- DL methods gets near human performance on SQUAD but:
  - Still ~~84~~ 89F1 vs. 91.2 F1.
  - Very restricted QA Setting (span selection, within paragraph, answer always present, high lexical overlap).
- Compared to under-incentivized humans.
  - (91.2 is a low estimate of human performance)
- Questions can be answered with "cheating".
  - (~~84.0~~ 89F1 is a high estimate of DL performance)