

A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task

Danqi Chen, Jason Bolton and Christopher D. Manning

Presented by Aidana Karipbayeva

Summary

- Description of the CNN and Daily Mail news
- Two models of Chen et al.(2015)
 - Entity-center classifier
 - End-to-end Neural Network
- Results
- In-depth data analysis
- Conclusion

CNN and Daily Mail news

Passage:

@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question:

characters in " @placeholder " movies have gradually become more diverse

Answer:

@entity6

(p, q , a) = (passage, question, answer)

Passage is the news article.

Question is formed in Cloze style, where a single entity in the bullet summaries is replaced with a placeholder (@placeholder).

Answer is the replaced entity.

Goal: To predict the answer entity from all appearing entities in the passage, given the passage and question.

Data Statistics

- The text has been run through a Google NLP pipeline.
- It is tokenized, lowercased, and entities are replaced with abstract entity markers (@entityn)



Hermann et al. (2015):

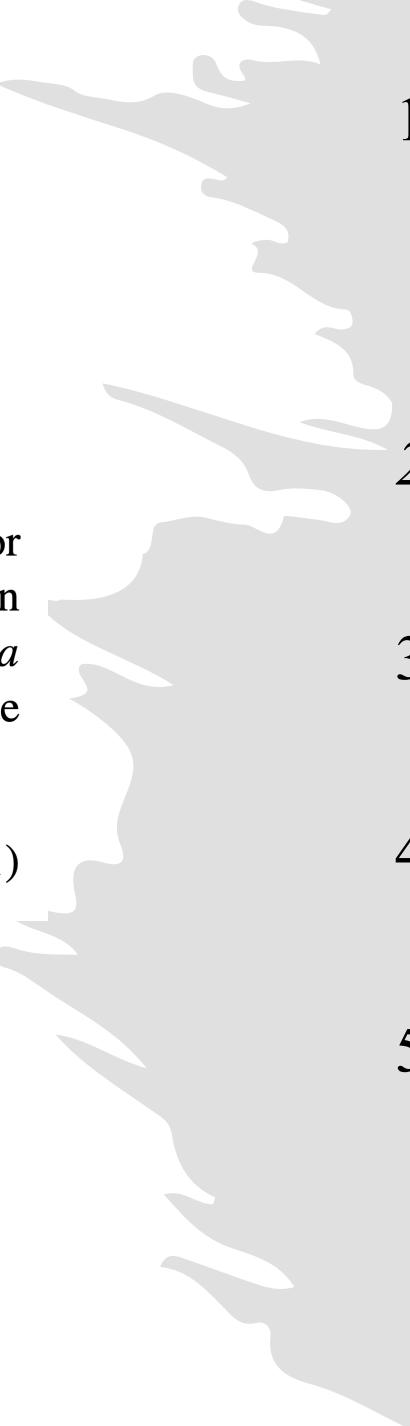
- Such a process ensures that their models are understanding the given passage, as opposed to applying world knowledge or co-occurrence.

	CNN	Daily Mail
# Train	380,298	879,450
# Dev	3,924	64,835
# Test	3,198	53,182
Passage: avg. tokens	761.8	813.1
Passage: avg. sentences	32.3	28.9
Question: avg. tokens	12.5	14.3
Avg. # entities	26.2	26.2

Entity-Centric Classifier

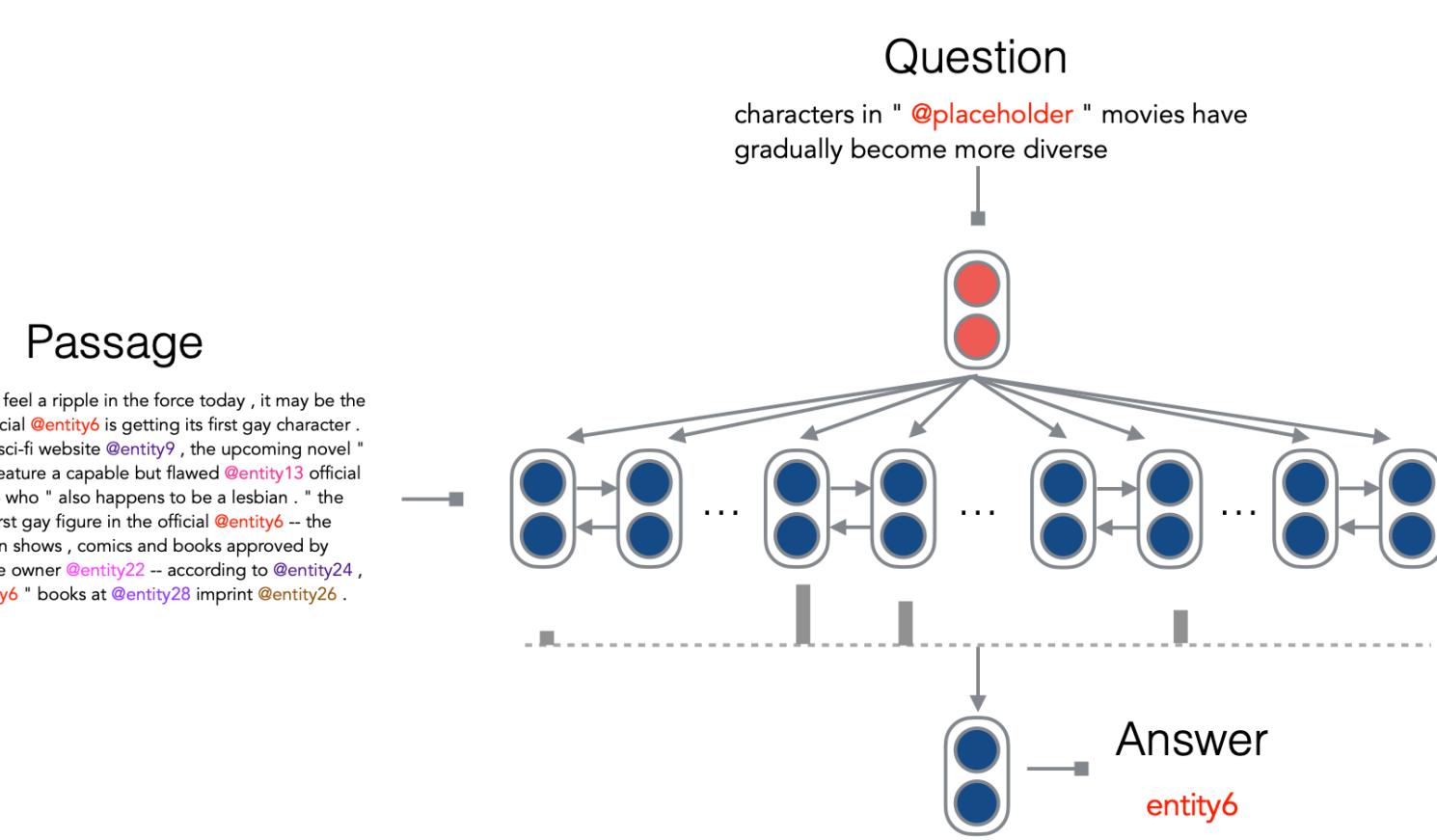
The setup of this system is to design a feature vector $f_{p,q}(e)$ for each candidate entity e , and to learn a weight vector θ such that the correct answer a is expected to rank higher than all other candidate entities:

$$\theta^\top f_{p,q}(a) > \theta^\top f_{p,q}(e), \forall e \in E \cap p \setminus \{a\} \quad (1)$$

- 
1. Whether entity e occurs in the passage, question, its frequency, first position of occurrence in the passage
 2. n -gram exact match
 3. Sentence co-occurrence
 4. Word distance
 5. Dependency parse match

End-to-end Neural Network

Passage $p: p_1, \dots, p_m \in \mathbb{R}^d$
 Question $q: q_1, \dots, q_l \in \mathbb{R}^d$
 Contextual emb.: \tilde{p}_i



Encoding:

$$\begin{aligned}\overrightarrow{\mathbf{h}}_i &= \text{LSTM}(\overrightarrow{\mathbf{h}}_{i-1}, \mathbf{p}_i), i = 1, \dots, m \\ \overleftarrow{\mathbf{h}}_i &= \text{LSTM}(\overleftarrow{\mathbf{h}}_{i+1}, \mathbf{p}_i), i = m, \dots, 1\end{aligned}$$

and $\tilde{\mathbf{p}}_i = \text{concat}(\overrightarrow{\mathbf{h}}_i, \overleftarrow{\mathbf{h}}_i) \in \mathbb{R}^h$, where $h = 2\tilde{h}$. Meanwhile, we use another bi-directional LSTM to map the question $\mathbf{q}_1, \dots, \mathbf{q}_l$ to an embedding $\mathbf{q} \in \mathbb{R}^h$.

Attention:

$$\begin{aligned}\alpha_i &= \text{softmax}_i \mathbf{q}^\top \mathbf{W}_s \tilde{\mathbf{p}}_i \\ \mathbf{o} &= \sum_i \alpha_i \tilde{\mathbf{p}}_i\end{aligned}$$

Prediction:

$$a = \arg \max_{a \in p \cap E} W_a^\top \mathbf{o}$$

Results

Model	CNN		Daily Mail	
	Dev	Test	Dev	Test
Frame-semantic model [†]	36.3	40.2	35.5	35.5
Word distance model [†]	50.5	50.9	56.4	55.5
Deep LSTM Reader [†]	55.0	57.0	63.3	62.2
Attentive Reader [†]	61.6	63.0	70.5	69.0
Impatient Reader [†]	61.8	63.8	69.0	68.0
Ours: Classifier	67.1	67.9	69.1	68.3
Ours: Neural net	72.4	72.4	76.9	75.8

- The **conventional feature-based classifier** obtains a 67.9% accuracy on the CNN test set, which actually outperforms the best neural network model from DeepMind.
- **Single-model neural network** surpasses the previous results of Attentive reader by a large margin (over 5%).

Questions to analyze

- i) Since the dataset was created synthetically, what proportion of questions are trivial to answer, and how many are noisy and not answerable?
- ii) What have these models learned?
- iii) What are the prospects of improving them?

To answer these, authors randomly sample 100 examples from the CNN dev dataset, to perform a breakdown of the examples.

Breakdown of the Examples

1. **Exact Match** - The nearest words around the placeholder in the question also appear identically in the passage, in which case, the answer is self-evident.
2. **Sentence-level paraphrase** - The question is a paraphrasing of exactly one sentence in the passage, and the answer can definitely be identified in the sentence.
3. **Partial Clue** - No semantic match between the question and document sentences exist but the answer can be easily inferred through partial clues such as word and concept overlaps.
4. **Multiple sentences** - Multiple sentences in the passage must be examined to determine the answer.
5. **Coreference errors** - This category refers to examples with critical coreference errors for the answer entity or other key entities in the question. **Not answerable.**
6. **Ambiguous / Very Hard** - This category includes examples for which even humans cannot answer correctly (confidently). **Not answerable.**

Data analysis

Distribution of these examples based on their respective categories:

No.	Category	(%)
1	Exact match	13
2	Paraphrasing	41
3	Partial clue	19
4	Multiple sentences	2
5	Coreference errors	8
6	Ambiguous / hard	17

“Coreference errors” and “ambiguous/hard” cases account for 25%



Barrier for training models with an accuracy above 75%

Only two examples require examination of multiple sentences for inference



A lower rate of challenging questions



The inference is based upon identifying the most relevant sentence.

Per-category Performance

Category	Classifier	Neural net
Exact match	13 (100.0%)	13 (100.0%)
Paraphrasing	32 (78.1%)	39 (95.1%)
Partial clue	14 (73.7%)	17 (89.5%)
Multiple sentences	1 (50.0%)	1 (50.0%)
Coreference errors	4 (50.0%)	3 (37.5%)
Ambiguous / hard	2 (11.8%)	1 (5.9%)
All	66 (66.0%)	74 (74.0%)

- 1) The exact-match cases are quite simple and both systems get 100% correct.
- 2) Both of systems perform poorly for the ambiguous/hard and entity-linking-error cases.
- 3) The two systems mainly differ in paraphrasing cases and “partial clue” cases. This shows how neural networks are better capable of learning semantic matches.
- 4) The neural-net system already achieves near-optimal performance on all the single-sentence and unambiguous cases.

Authors' conclusion

- I. This dataset is easier than previously realized.
- II. Straightforward, conventional NLP systems can do much better on it than previously suggested.
- III. Deep learning systems are very effective at recognizing paraphrases.
- IV. Presented are close to the ceiling of performance for single-sentence and unambiguous cases of this dataset.
- V. It is hard to get final 20% of questions correct, since most of them had issues in the data preparation which decreases the chances of answering the question

References

- 1) Chen, D., Bolton, J., & Manning, C. D. (2016). A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

- 2) Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems* (pp. 1693-1701).

Appendix 1.1: Two models of Hermann et al. (2015) for comparison

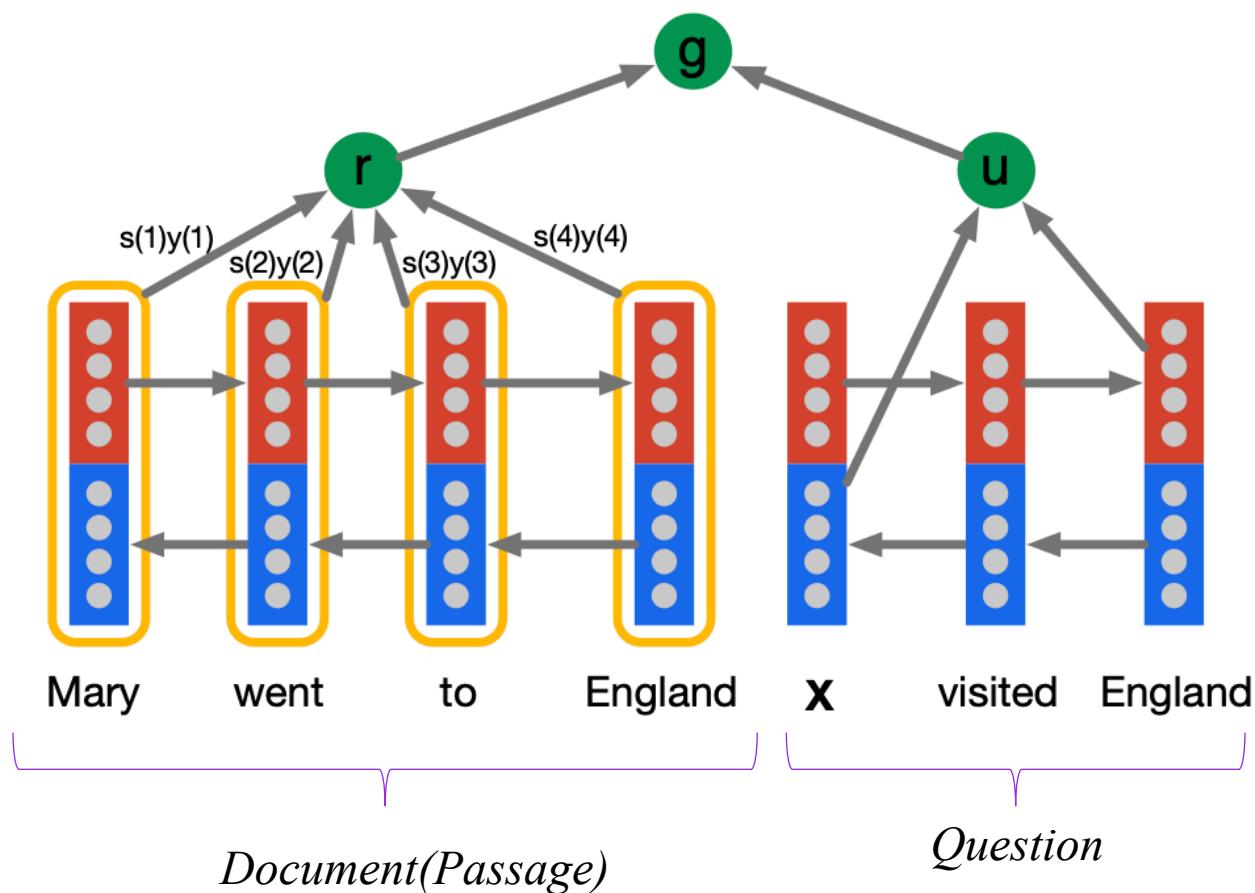
- Frame-Sematic Parsing
- Attentive Reader

Appendix 1.2: Frame-Sematic Parsing by Hermann et al.

Strategy	Pattern $\in q$	Pattern $\in d$	Example (Cloze / Context)
1 Exact match	(p, V, y)	(\mathbf{x}, V, y)	X loves Suse / Kim loves Suse
2 be.01.V match	$(p, \text{be.01.V}, y)$	$(\mathbf{x}, \text{be.01.V}, y)$	X is president / Mike is president
3 Correct frame	(p, V, y)	(\mathbf{x}, V, z)	X won Oscar / Tom won Academy Award
4 Permutated frame	(p, V, y)	(y, V, \mathbf{x})	X met Suse / Suse met Tom
5 Matching entity	(p, V, y)	(\mathbf{x}, Z, y)	X likes candy / Tom loves candy
6 Back-off strategy	<i>Pick the most frequent entity from the context that doesn't appear in the query</i>		

Extracting entity-predicate triples—denoted as (e_1, V, e_2) —from both the query q and context document d , Hermann et al. (2015) attempt to resolve queries using a number of rules with an increasing recall/precision trade-off.

Appendix 1.3: Attentive Reader by Hermann et al.



Authors denote the outputs of the forward and backward LSTMs as $\overrightarrow{y(t)}$ and $\overleftarrow{y(t)}$ respectively.

Encoding vector of question:

$$\mathbf{u} = \overrightarrow{y_q(|q|)} \parallel \overleftarrow{y_q(1)}$$

For the document, the output for each token at t:

$$y_d(t) = \overrightarrow{y_d(t)} \parallel \overleftarrow{y_d(t)}$$

The representation r of the document d is formed by a weighted sum of these output vectors:

$$m(t) = \tanh (W_{ym} y_d(t) + W_{um} u),$$

$$s(t) \propto \exp (w_{ms}^\top m(t)),$$

$$\mathbf{r} = \mathbf{y}_d s$$

The model is completed with the definition of the joint document and query embedding via a non-linear combination:

$$g^{\text{AR}}(d, q) = \tanh (W_{rg} r + W_{ug} u).$$

Appendix 1.4: Differences between two neural models

- **Essential:**
 - Using of a bilinear term, instead of a tanh layer to compute attention between question and contextual embeddings.
- **Simplification of a model:**
 - After obtaining the weighted contextual embeddings \mathbf{o} , authors use \mathbf{o} for direct prediction. In contrast, the original model in Hermann et al. (2015) combined \mathbf{o} and the question embedding \mathbf{q} via another non-linear layer before making final predictions.
 - The original model considers all the words from the vocabulary V in making predictions. Chen et al(205) only predict among entities which appear in the passage.



SQuAD: 100,000+ Questions For Machine Comprehension Of Text

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang

Stanford University

Presented By: Keval Morabia (morabia2)

OUTLINE



QA TASK



EXISTING QA
DATASETS



SQuAD
COLLECTION
PROCESS



SQuAD
STATISTICS



METHODS



EXPERIMENTS

1. THE QUESTION ANSWERING TASK

- Types of Answers:
 - Multiple Choice
 - Selecting a span of text
- Challenges:
 - Understanding Natural Language
 - Knowledge about the world

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

2. EXISTING QA DATASETS

- Reading Comprehension QA datasets
- Open-domain QA datasets
 - Answer a question from a large collection of docs
- Cloze datasets
 - Predict missing word (often a named entity) in a passage
 - Performance almost saturated

Dataset	Question source	Formulation	Size
SQuAD	crowdsourced	RC, spans in passage	100K
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855
WikiQA (Yang et al., 2015)	query logs	IR, sentence selection	3047
TREC-QA (Voorhees and Tice, 2000)	query logs + human editor	IR, free form	1479
CNN/Daily Mail (Hermann et al., 2015)	summary cloze	RC, fill in single entity	1.4M
CBT (Hill et al., 2015)	cloze	RC, fill in single word	688K

Table 1: A survey of several reading comprehension and question answering datasets. SQuAD is much larger than all datasets except the semi-synthetic cloze-style datasets, and it is similar to TREC-QA in the open-endedness of the answers.

3. SQuAD COLLECTION PROCESS

3.1 Passage Curation

- Sample 536 articles from top 10k Wikipedia articles
- Extract individual paragraphs (with >500 characters) from each article
- Finally, 23k paragraphs (8:1:1 split)

3. SQuAD COLLECTION PROCESS

3.2 Question-answer collection

Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O₂.

2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words

Select Answer

Ask a question here. Try using your own words

Select Answer

Figure 2: The crowd-facing web interface used to collect the dataset encourages crowdworkers to use their own words while asking questions.

3. SQuAD COLLECTION PROCESS

3.3 Additional answers collection

- For robust evaluation
- 2 additional answers for each question in dev/test set
- 2.6% unanswerable

4. SQuAD STATISTICS – DEV SET

4.1 Diversity in answers

- Non-numerical answers categorized by
 - Constituency parsers
 - POS tags
- Proper nouns categorized by NER tags

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Table 2: We automatically partition our answers into the following categories. Our dataset consists of large number of answers beyond proper noun entities.

4. SQuAD STATISTICS – DEV SET

4.2 Reasoning required to answer

- Sample 4 questions from each article
- Manually label into one or more of the below categories
 - Lexical Variation [42%]
 - Syntactic Variation [64%]
 - Multiple Sequence Reasoning [14%]
 - Ambiguous [6%]

4. SQuAD STATISTICS – DEV SET

4.3 Syntactic divergence

- Edit distance b/w unlexicalized dependency paths in the question (Q) and the sentence containing the answer (S)

Q: What department store is thought to be the first in the world?
S: Bainbridge's is often cited as the world's first department store.

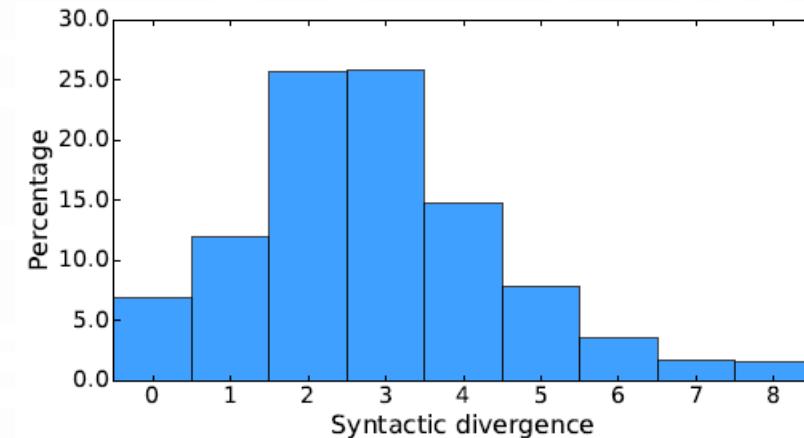
Path:

```
graph LR; Q[first] -- "xcomp" --> S[store]; Q -- "nsubjpass" --> S; Q -- "det" --> S[what]; S -- "delete" --> Q[first]; S -- "substitute" --> Q[first]; S -- "amod" --> S[store]; S -- "nmod" --> S[cited]; S -- "nsubjpass" --> S[Bainbridge's]
```

Edit cost:

1 +2 +1=4

Figure 3: An example walking through the computation of the syntactic divergence between the question Q and answer sentence S.



(a) Histogram of syntactic divergence.

5. METHODS FOR QA

- Candidate answer generation:
 - Instead of $O(L^2)$ spans, consider those which are constituents in the constituency parse generated by Stanford CoreNLP
 - 77.3% answers in dev set are constituents (Upper bound on accuracy of such models)

5. METHODS FOR QA

5.1 Sliding Windows baseline

- For each candidate answer, compute unigram/bigram overlap between question and the sentence containing the answer
- Select the best candidate answer using a Sliding Window approach (Not clear in paper)
- Add distance based extension to consider long-range dependencies

5. METHODS FOR QA

5.2 Logistic Regression

- Discretize continuous feature in 10 equally-sized bins
- Extract 180 million features for each candidate answer
 - Matching unigram/bigram frequency
 - Length feature
 - Constituent label
 - POS tag
 - Lexical features
 - Dependency tree path features

6. EXPERIMENTS

- Evaluation Metrics:
 - **Exact Match** - % of predictions that match any ground truth answer
 - **Macro-averaged F1 score** – measure average overlap b/w prediction and ground truth answer (both considered as bag of tokens)

6. EXPERIMENTS

	Exact Match		F1	
	Dev	Test	Dev	Test
Random Guess	1.1%	1.3%	4.1%	4.3%
Sliding Window	13.2%	12.5%	20.2%	19.7%
Sliding Win. + Dist.	13.3%	13.0%	20.2%	20.0%
Logistic Regression	40.0%	40.4%	51.0%	51.0%
Human	80.3%	77.0%	90.5%	86.8%

Table 5: Performance of various methods and humans. Logistic regression outperforms the baselines, while there is still a significant gap between humans.

	Logistic Regression Dev F1	Human Dev F1
Date	72.1%	93.9%
Other Numeric	62.5%	92.9%
Person	56.2%	95.4%
Location	55.4%	94.1%
Other Entity	52.2%	92.6%
Common Noun Phrase	46.5%	88.3%
Adjective Phrase	37.9%	86.8%
Verb Phrase	31.2%	82.4%
Clause	34.3%	84.5%
Other	34.8%	86.1%

Table 7: Performance stratified by answer types. Logistic regression performs better on certain types of answers, namely numbers and entities. On the other hand, human performance is more uniform.

7. CONCLUSION

- Introduced the Stanford Question Answering Dataset (SQuAD) v1.0 containing 100k questions
- Contains a diverse range of QA types
- Human Performance >> Logistic Regression (Scope of improvement)
- SQuAD v1.1 and v2 created afterwards



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

Authors: Pranav Rajpurkar, Robin Jia, Percy Liang

Presenter: Si Zhang

Machine Reading Comprehension

- Question: about a paragraph or a document
 - Answer: often a span in the document
-

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Previous SQuAD Dataset

- SQuAD 1.1
 - Good performance by context and type-matching
 - Not robust to distracting sentences
- Reason: Guaranteed correct answers exist in the context
- Limitations:
 - Model only needs to select the related span
 - No need to check answers entailed by the text
- Q: How can we make the dataset more challenging?

Generic Solution

- Add unanswerable questions about same paragraph (neg. example)
- Two desiderata for unanswerable questions
 - Relevance:
 - Unanswerable questions appear relevant to context paragraph
 - **Benefit:** simple heuristics can't distinguish answerable & unanswerable
 - Existence of plausible answers:
 - Exists some span whose type matches the type of answer
 - **Benefit:** type-matching can't distinguish answerable & unanswerable

An Example

Article: Endangered Species Act

Paragraph: “*... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*”

Question 1: “Which laws faced significant opposition?”

Plausible Answer: *later laws*

Question 2: “What was the name of the 1937 treaty?”

Plausible Answer: *Bald Eagle Protection Act*

Figure 1: Two unanswerable questions written by crowdworkers, along with plausible (but incorrect) answers. Relevant keywords are shown in blue.

Dataset – Creation

- Employ workers on Daemo crowdsourcing platform.
- Each task consists of an article from SQuAD 1.1.
- Workers are asked to write 5 questions per paragraph

Paragraph 2 of 25

Spend around 7 minutes on the following paragraph to ask 5 **impossible** questions! If you can't ask 5 questions, ask 4, but do your best to ask 5. Select a plausible answer from the paragraph by clicking on 'Select Plausible Answer', and then highlight the smallest segment of the paragraph that is a plausible answer to the question.

In the 1960s, a series of discoveries, the most important of which was seafloor spreading, showed that the Earth's lithosphere, which includes the crust and rigid uppermost portion of the upper mantle, is separated into a number of tectonic plates that move across the plastically deforming, solid, upper mantle, which is called the asthenosphere. There is an intimate coupling between the movement of the plates on the surface and the convection of the mantle: oceanic plate motions and mantle convection currents always move in the same direction, because the oceanic lithosphere is the rigid upper thermal boundary layer of the convecting mantle. This coupling between rigid plates moving on the surface of the Earth and the convecting mantle is called plate tectonics.

Ask a question here. Use your own words, instead of copying from paragraph

Select Plausible Answer

Ask a question here. Use your own words, instead of copying from paragraph

Select Plausible Answer

Ask a question here. Use your own words, instead of copying from paragraph

Select Plausible Answer

Ask a question here. Use your own words, instead of copying from paragraph

Select Plausible Answer

Ask a question here. Use your own words, instead of copying from paragraph

Select Plausible Answer

Questions for inspiration

What was the most important discovery that led to the understanding that Earth's lithosphere is separated into tectonic plates?

seafloor spreading

Which parts of the Earth are included in the lithosphere?
the crust and rigid uppermost portion of the upper mantle

What is another word for the Earth's upper mantle?
asthenosphere

Plate tectonics can be seen as the intimate coupling between rigid plates on the surface of the Earth and what?
the convecting mantle

In what decade was seafloor spreading discovered?
the 1960s

Dataset – Human Accuracy

- Dataset statistics

	SQuAD 1.1	SQuAD 2.0
Train		
Total examples	87,599	130,319
Negative examples	0	43,498
Total articles	442	442
Articles with negatives	0	285
Development		
Total examples	10,570	11,873
Negative examples	0	5,945
Total articles	48	35
Articles with negatives	0	35
Test		
Total examples	9,533	8,862
Negative examples	0	4,332
Total articles	46	28
Articles with negatives	0	28

- Hire workers to answer question in dev. & test sets
- Select final answer by majority voting

Dataset – Analysis

- Goal: to understand the challenges that neg. examples present

Reasoning	Description	Example	Percentage
Negation	Negation word inserted or removed.	Sentence: “Several hospital pharmacies have decided to outsource high risk preparations . . .” Question: “What types of pharmacy functions have never been outsourced?”	9%
Antonym	Antonym used.	S: “the extinction of the dinosaurs. . . allowed the tropical rainforest to spread out across the continent.” Q: “The extinction of what led to the decline of rainforests?”	20%
Entity Swap	Entity, number, or date replaced with other entity, number, or date.	S: “These values are much greater than the 9–88 cm as projected . . . in its Third Assessment Report.” Q: “What was the projection of sea level increases in the fourth assessment report ? ”	21%
Mutual Exclusion	Word or phrase is mutually exclusive with something for which an answer is present.	S: “BSkyB. . . waiv[ed] the charge for subscribers whose package included two or more premium channels.” Q: “What service did BSkyB give away for free unconditionally ? ”	15%
Impossible Condition	Asks for condition that is not satisfied by anything in the paragraph.	S: “Union forces left Jacksonville and confronted a Confederate Army at the Battle of Olustee. . . Union forces then retreated to Jacksonville and held the city for the remainder of the war.” Q: “After what battle did Union forces leave Jacksonville for good ? ”	4%
Other Neutral	Other cases where the paragraph does not imply any answer.	S: “Schuenemann et al. concluded in 2011 that the Black Death. . . was caused by a variant of <i>Y. pestis</i> . . . ” Q: “Who discovered <i>Y. pestis</i> ? ”	24%
Answerable	Question is answerable (i.e. dataset noise).		7%

Table 1: Types of negative examples in SQuAD 2.0 exhibiting a wide range of phenomena.

Experimental Setups

- Baseline models
 - BiDAF-No-Answer (BNA)
 - DocQA w/ ELMo
 - DocQA w/o ELMo
- Metrics
 - Average exact match (EM)
 - F1 scores

Experimental Results

- Main results

System	SQuAD 1.1 test		SQuAD 2.0 dev		SQuAD 2.0 test	
	EM	F1	EM	F1	EM	F1
BNA	68.0	77.3	59.8	62.6	59.2	62.1
DocQA	72.1	81.0	61.9	64.8	59.3	62.3
DocQA + ELMo	78.6	85.8	65.1	67.6	63.4	66.3
Human	82.3	91.2	86.3	89.0	86.9	89.5
Human–Machine Gap	3.7	5.4	21.2	21.4	23.5	23.2

- Observation #1:
 - Best model is 23.2% lower than human accuracy
 - Indicate significant room for model improvement
- Observation #2:
 - Much larger human-machine gap on two datasets → a harder task

Experimental Results

- Comparisons on different neg. example generation
 - Against automatic generation by TFIDF and Rule-Based

System	SQuAD 1.1 + TFIDF		SQuAD 1.1 + RULEBASED		SQuAD 2.0 dev	
	EM	F1	EM	F1	EM	F1
BNA	72.7	76.6	80.1	84.8	59.8	62.6
DocQA	75.6	79.2	80.8	84.8	61.9	64.8
DocQA + ELMo	79.4	83.0	85.7	89.6	65.1	67.6

- Observation:
 - Highest F1 score on SQuAD 2.0 is >15.4% lower than automatic ways
 - Suggesting automatic ways are easier to detect

Conclusion

- A new dataset SQuAD 2.0 with unanswerable questions
- Data creation
 - Crowdsourcing
- Experiments
 - Indicate the data is more challenging than SQuAD 1.1
 - Indicate challenging negative examples compared to automatic generation ways



Thank You!

Q & A

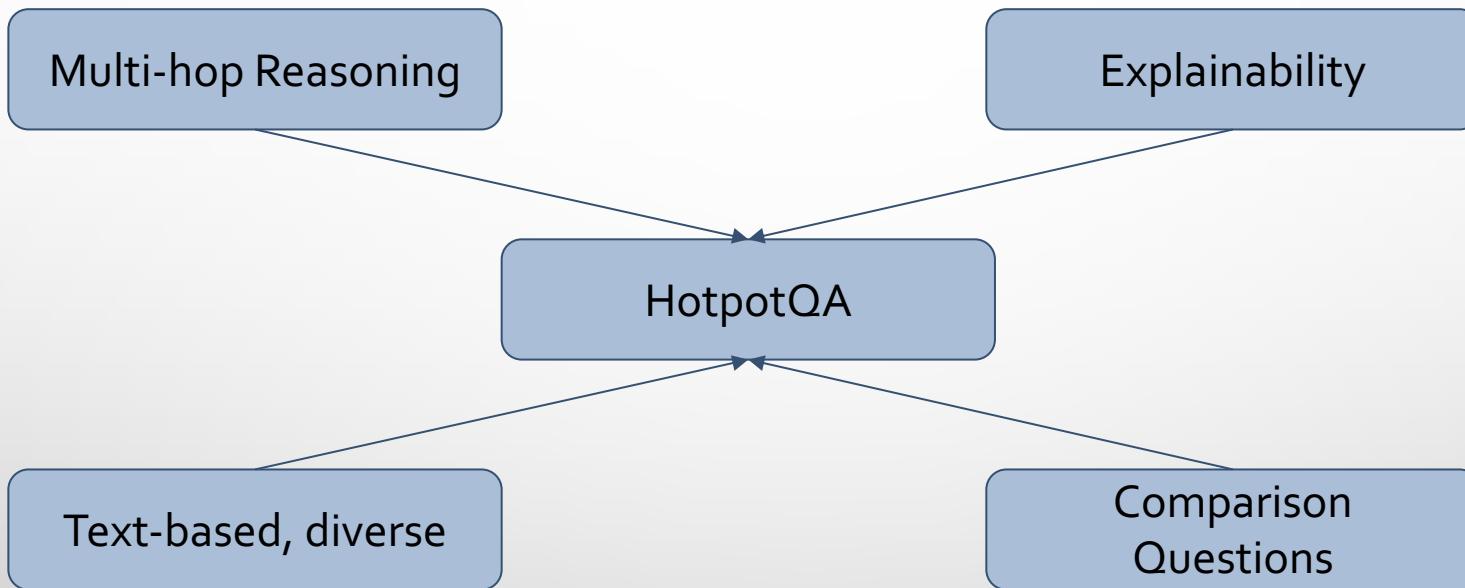
HOTPOTQA: A DATASET FOR DIVERSE, EXPLAINABLE MULTI-HOP QUESTION ANSWERING

ZHUOHAO ZHANG

MOTIVATION AND RESEARCH QUESTION

- Do we really need another QA dataset?
 - How to solve multi-hop reasoning QA?
-
- Simple question that stumped a lot NLP systems:
 - In which city was Facebook first launched
 - Mark Zuckerberg -> Harvard -> Cambridge

FEATURES

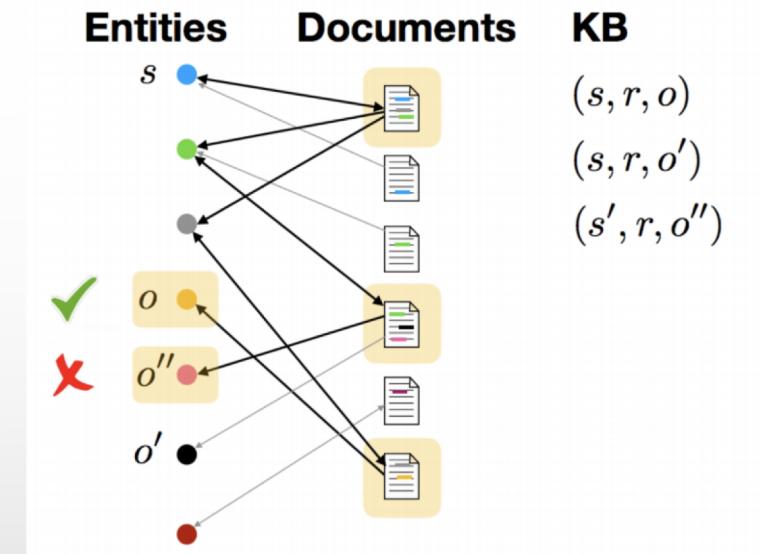
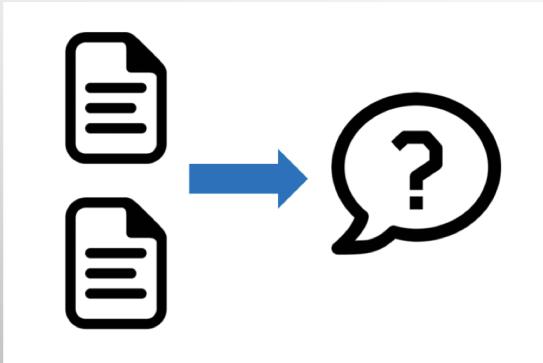


FEATURES: MULTI-HOP REASONING ACROSS DOCUMENTS

- Previous work (SQuAD, TriviaQA, etc): When was **Chris Martin** born?
- Hotpot QA: When was **the lead singer of Coldplay** born?
- Not the first Multi-hop reasoning QA dataset
- Difference between HotpotQA and previous work?
 - Diverse
 - Not pre-defined by other schema
 - Explainable by supporting factors

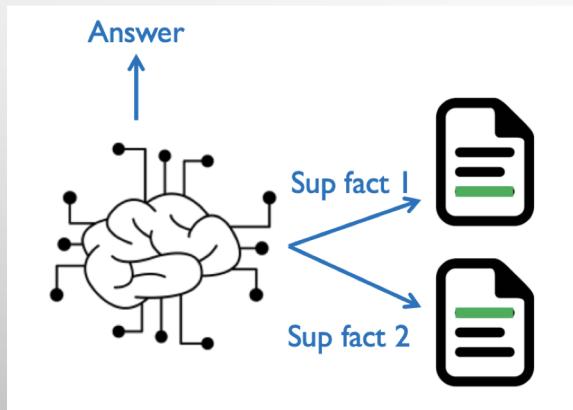
FEATURES: OPEN-DOMAIN TEXT-BASED QS AND AS

- Previous work
 - QAngaroo
 - ComplexWebQuestions
 - ...
- HotpotQA
 - Not relying on Knowledge base



FEATURES: EXPLAINABILITY

- Previous work: Black box
- HotpotQA: Supporting factors



FEATURES - COMPARISON QUESTION

1. Arithmetic comparison & comparing properties
2. Answer could be yes/no

Examples:

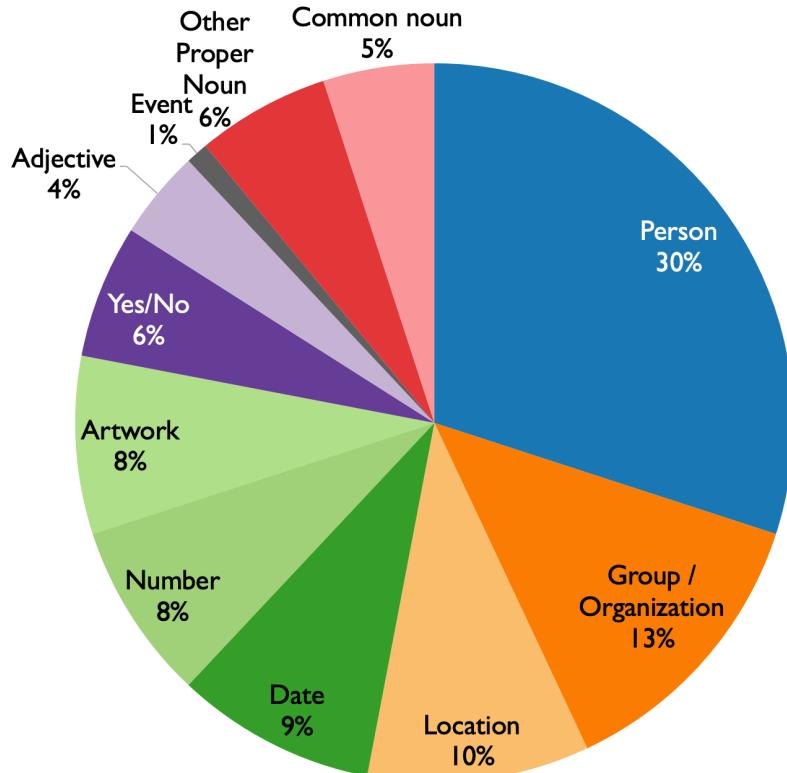
- Who has played for more teams, Michael Jordan or Kobe Bryant?
- Who was born earlier, Yuri Gagarin or Valentina Tereshkova?

DATA COLLECTION

- Hyperlinks -> Entity Graph
- Bridge entity questions
 - Mark Zuckerberg -> Harvard University -> Cambridge, MA
- Comparison Questions
 - Wikipedia: Lists of lists of lists
- Gave these to Turkers to come up with questions

QUESTION DIVERSITY

- Including:
 - Person,
 - Group/Organization,
 - Artwork,
 - Other proper noun



TYPES OF REASONING

Type I: Build a bridge

“Where’s the US President born?”

Type II: Intersection between two paragraphs

“What contains ice and fire”

Type III: More complex

Zuckerberg -> Harvard -> Cambridge

EVALUATION SETTINGS

- Distractor
 - 2 gold paragraphs + 8 from information retrieval (fixed for all models)
- Fullwiki
 - Entire Wikipedia as context
 - (In this work) 10 paragraphs from IR

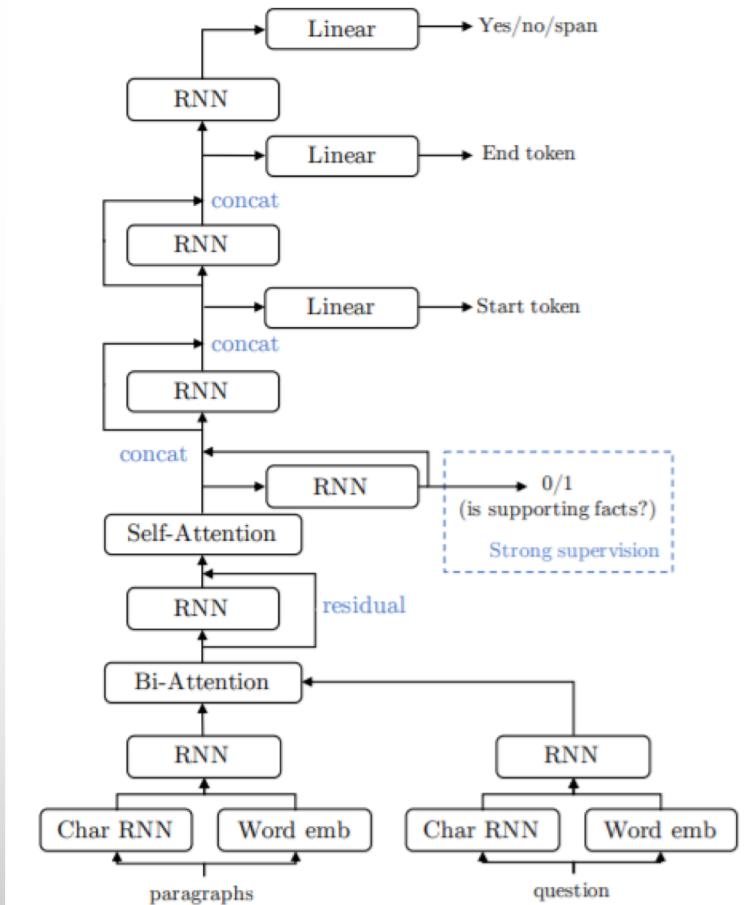
EVALUATION METRICS

1. Accuracy of answer
 2. Supporting factors
-
- Joint metric to combine the two

$$P^{(\text{joint})} = P^{(\text{ans})} P^{(\text{sup})}, \quad R^{(\text{joint})} = R^{(\text{ans})} R^{(\text{sup})},$$

$$\text{Joint F}_1 = \frac{2P^{(\text{joint})}R^{(\text{joint})}}{P^{(\text{joint})} + R^{(\text{joint})}}.$$

- Baseline Model: BiDAF++ w/ S-Norm



BASELINE RESULTS

Setting	Split	Answer		Sup Fact		Joint	
		EM	F ₁	EM	F ₁	EM	F ₁
distractor	dev	44.44	58.28	21.95	66.66	11.56	40.86
distractor	test	45.46	58.99	22.24	66.62	12.04	41.37
full wiki	dev	24.68	34.36	5.28	40.98	2.54	17.73
full wiki	test	25.23	34.40	5.07	40.69	2.63	17.85

ADDING HUMAN EVALUATIONS

Setting	Answer		Sp Fact		Joint	
	EM	F ₁	EM	F ₁	EM	F ₁
gold only	65.87	74.67	59.76	90.41	41.54	68.15
distractor	60.88	68.99	30.99	74.67	20.06	52.37
Human	83.60	91.40	61.50	90.04	52.30	82.55
Human UB	96.80	98.77	87.40	97.56	84.60	96.37

CONCLUSION

1. Multi-hop reasoning QA with diversity and explainability
2. New type of comparison question
3. New baseline model:

HotpotQA baseline model is available at <https://github.com/hotpotqa/hotpot>

Constructing Datasets for Multi-hop Reading Comprehension Across Documents

JOHANNES WELBL, PONTUS STENETORP, SEBASTIAN RIEDEL

PRESENTED BY LU WANG

Multi-Hop Reading Comprehension

Answer of given query can be inferred from information across multiple documents.

The new fact is derived by combining facts via a chain of multiple steps.

The Hanging Gardens, in [Mumbai], also known as Pherozeshah Mehta Gardens, are terraced gardens ... They provide sunset views over the [Arabian Sea] ...

Mumbai (also known as Bombay, the official name until 1995) is the capital city of the Indian state of Maharashtra. It is the most populous city in **India** ...

The **Arabian Sea** is a region of the northern Indian Ocean bounded on the north by **Pakistan** and **Iran**, on the west by northeastern **Somalia** and the Arabian Peninsula, and on the east by **India** ...

Q: (Hanging gardens of Mumbai, country, ?)
Options: {Iran, **India**, Pakistan, Somalia, ...}

Problem Statement

- Most of existing QA system limit on answer question from single source
- This work introduce a method to produce datasets given a collection of query-answer pairs and multiple linked documents

Task Formalization

The model consisting the following

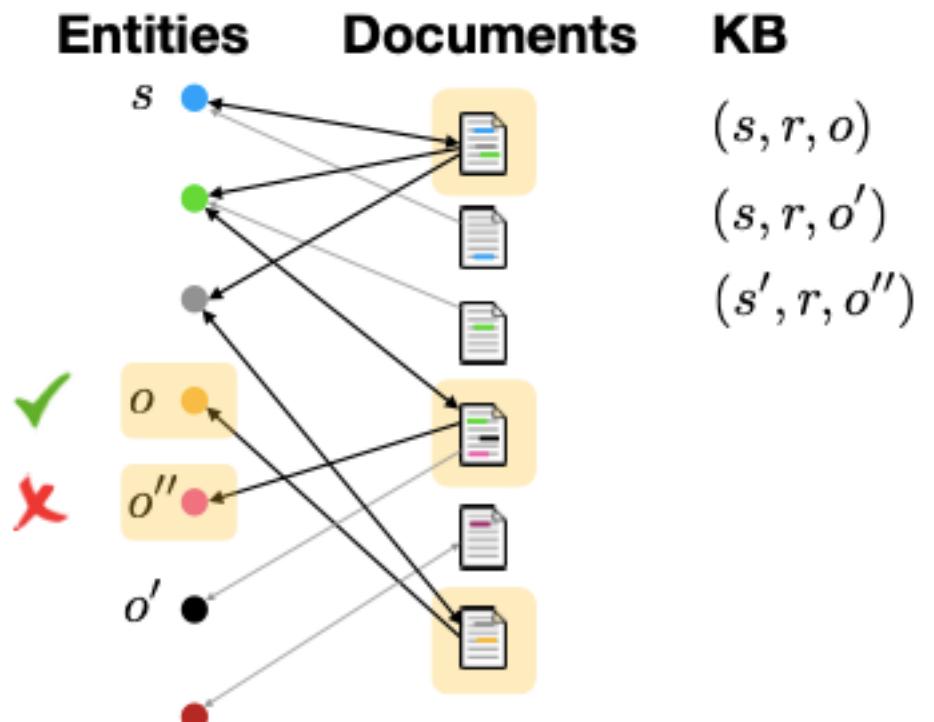
- A query q
- A set of supporting documents S_q
- A set of candidate answers C_q

The goal is to identify the correct answer $a^* \in C_q$

Dataset Construction Method

Use of Knowledge Base

- Assume that there exists a document corpus D , together with a KB containing fact triples (s, r, o)
 - Ex: (Hanging Gardens of Mumbai, country, India)
- Query Answer Pair: $q = (s, r, ?)$ and $a^* = o$
- Start From entity s
- Traverse to find type-consistent candidates



WikiHop

Source

Documents: WIKIPEDIA

Knowledge Base: WIKIDATA

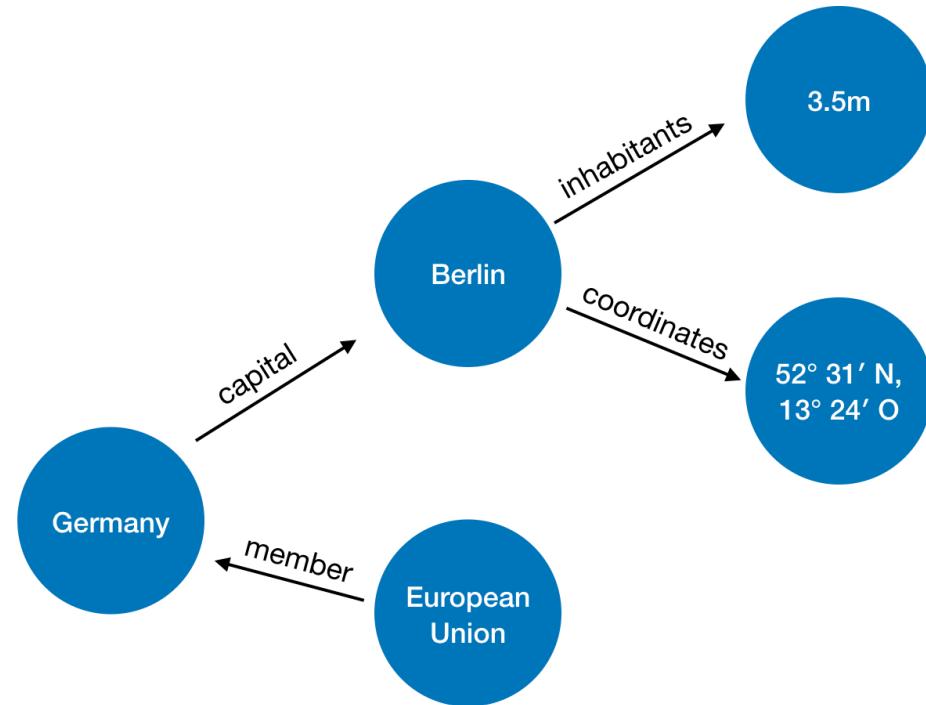
Bipartite Graph Construction

Edge Structure:

- edges from articles to entities: all articles mentioning an entity e are connected to e
- edges from entities to articles: each entity e is only connected to the WIKIPEDIA article about the entity.

Traverse up to a maximum chain length of 3 documents

Remove samples(1%) with more than 64 different support documents or 100 candidates



MedHop

Source

Documents: research paper abstracts from MEDLINE

Knowledge Base: DRUGBANK which is KB containing drug information

Dataset Construction

Interacts_with: The only relation for DRUGBANK connecting pairs of drugs

- Ex: (Leuprolide, interacts with, ?)

Edge Structure:

- Edges from a document to all proteins mentioned in it
- Edges between a document and a drug
- Edges From a protein p to a document mentioning p

Mitigating Dataset Biases

Candidate Frequency Imbalance

- significant bias in the answer distribution of WIKIREADING.
 - Ex: in the majority of the samples the property country has the *United States of America* as the answer.
- Solution: Subsampling so that any answer candidate make up no more than 0.1% of the dataset

Document-Answer Correlations

- certain documents frequently co-occur with the correct answer, independently of the query.
 - Ex: if the article about *London* is present in S_q , the answer is likely to be the *United Kingdom*
- Solution:
 - $cooccurrence(d, c)$: The total count of document d co-occurs with correct answer c in a sample
 - Filter out samples with document-candidate pair (d, c) which $cooccurrence(d, c) > 20$

Large Document Sets

- Entities in MedHop have large support document sets
- Solution: subsample documents until reach the limit of 64 documents

Dataset Analysis

Dataset Size

	Train	Dev	Test	Total
WIKIHOP	43,738	5,129	2,451	51,318
MEDHOP	1,620	342	546	2,508

Number of candidates and documents per sample

	min	max	avg	median
# cand. – WH	2	79	19.8	14
# docs. – WH	3	63	13.7	11
# tok/doc – WH	4	2,046	100.4	91
# cand. – MH	2	9	8.9	9
# docs. – MH	5	64	36.4	29
# tok/doc – MH	5	458	253.9	264

Qualitative Analysis

WikiHop

Unique multi-step answer.	36%
Likely multi-step unique answer.	9%
Multiple plausible answers.	15%
Ambiguity due to hypernymy.	11%
Only single document required.	9%
<hr/>	
Answer does not follow.	12%
WIKIDATA/WIKIPEDIA discrepancy.	8%

MediHop

Considered if the answer to the query “*follows*”, “*is likely*”, or “*does not follow*”

68% of the cases were considered as “*follows*” or as “*is likely*”

Baseline Models

Random Selects a random candidate

Max-mention Predicts the most frequently mentioned candidate in the documents S_q

Majority-candidate-per-query-type Predicts the candidate $c \in C_q$ that was most frequently observed as the true answer in the training set.

TF-IDF Predicts the candidate with the highest TF-IDF similarity score

$$\arg \max_{c \in C_q} [\max_{s \in S_q} (\text{TF-IDF}(q + c, s))]$$

Baseline Models

Document-cue

capture model ability to exploit document-answer co-occurrences. It predicts the candidate with highest score across C_q :

$$\arg \max_{c \in C_q} [\max_{d \in S_q} (\text{cooccurrence}(d, c))]$$

Extractive RC models: FastQA and BiDAF

Two *LSTM*-based extractive QA models are capable to predict answer within a *single* document

Adapt them to a multi-document setting by concatenating all $d \in S_q$ into a super document

Experiment Results

Experimental results for WIKIHOP and MEDHOP with masked setting and unmasked setting

Model	WIKIHOPO				MEDHOP			
	standard		masked		standard		masked	
	test	test*	test	test*	test	test*	test	test*
Random	11.5	12.2	12.2	13.0	13.9	20.4	14.1	22.4
Max-mention	10.6	15.9	13.9	20.1	9.5	16.3	9.2	16.3
Majority-candidate-per-query-type	38.8	44.2	12.0	13.7	58.4	67.3	10.4	6.1
TF-IDF	25.6	36.7	14.4	24.2	9.0	14.3	8.8	14.3
Document-cue	36.7	41.7	7.4	20.3	44.9	53.1	15.2	16.3
FastQA	25.7	27.2	35.8	38.0	23.1	24.5	31.3	30.6
BiDAF	42.9	49.7	54.5	59.8	47.8	61.2	33.7	42.9

Conclusion

- The constructed datasets enable multi-hop reading comprehension model successfully perform task with reasonable accuracy
- There is still room to improve further
- Currently datasets are focused on factoid questions about entities and rely on structured knowledge resources
- Future works can be done to free answer from abstractive form

Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering

Todor Mihaylov¹, Peter Clark², Tushar Khot², Ashish Sabharwal²

¹Allen Institute for Artificial Intelligence, Seattle, WA, U.S.A.

²Research Training Group AIPHES & Heidelberg University, Heidelberg, Germany

Presented by Zhonghao Wang

Contents

- Introduction
- OpenBookQA dataset
- Baseline model
- Baseline performance
- Discussions
- Conclusions

Introduction

- Introduce a new kind of question answering dataset, OpenBookQA, modeled after open book exams for assessing human understanding of a subject.

Question:
Which of these would let the most heat travel through?

A) a new pair of jeans.
B) a steel spoon in a cafeteria.
C) a cotton candy at a store.
D) a calvin klein cotton hat.

Science Fact:
Metal is a thermal conductor.

Common Knowledge:
Steel is made of metal.
Heat travels through a thermal conductor.

knowledge in a book ←

knowledge in memory ←

Figure 1: An example for a question with a given set of choices and supporting facts.

Introduction

- Contributions of this work
 1. Collect a QA dataset requiring multi-hop reasoning with partial context provided by a set of diverse facts.
 2. Conduct early researches including developing attention-based neural baselines and incorporating external knowledge. The accuracy reaches 76% but is still worse than human performance at 92% and therefore urge further studies.

Contents

- Introduction
- OpenBookQA dataset
- Baseline model
- Baseline performance
- Discussions
- Conclusions

OpenbookQA dataset

- Some numbers
 - ~6,000 4-way multiple-choice questions
 - each question associated with 1 core fact
 - 1326 core facts in total
 - ~6,000 additional facts

Question:

Which of these would let the most heat travel through?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

Science Fact:

Metal is a thermal conductor.

Common Knowledge:

Steel is made of metal.

Heat travels through a thermal conductor.

Figure 1: An example for a question with a given set of choices and supporting facts.

OpenbookQA dataset

- The question generation and filtering process

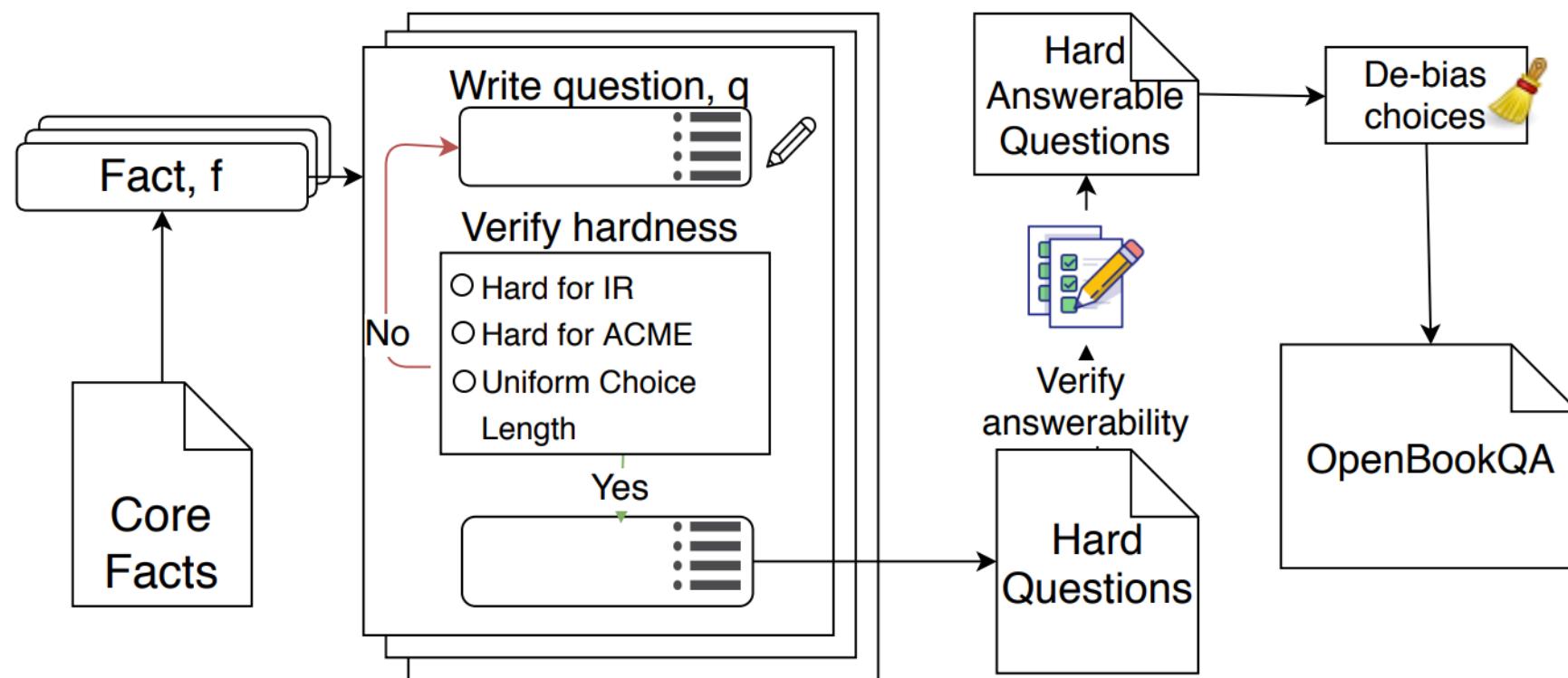


Figure 2: OpenBookQA question generation pipeline

OpenbookQA dataset

- Human performance

The human accuracy on the question set can be estimated by

$$H(Q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|I|} \sum_{i \in I} X_{q,i}$$

where Q represents the question set, I represents the set of human examinees, and

$$X_{q,i} = \begin{cases} 0, & \text{If a wrong answer} \\ 1, & \text{If a correct answer} \end{cases}$$

The result is 92%.

OpenbookQA dataset

- Question Set Analysis

- Statistics

OpenBookQA consists of 5957 questions, with 4957/500/500 in the Train/Dev/Test splits.

OpenBookQA Statistics	
# of questions	5957
# of choices per question	4
Avg. question sentences	1.08 (6)
Avg. question tokens	11.46 (76)
Avg. choice tokens	2.89 (23)
Avg. science fact tokens	9.38 (28)
Vocabulary size (q+c)	11855
Vocabulary size (q+c+f)	12839
Answer is the longest choice	1108 (18.6%)
Answer is the shortest choice	216 (3.6%)

Table 1: Statistics for full OpenBookQA dataset. Parenthetical numbers next to each average are the *max*.

OpenbookQA dataset

- Question Set Analysis
 - Percentage of questions and facts for the five most common type of additional facts

Fact Type	% Questions	% Facts
PROPERTY	29.11%	25.81%
ISA	20.25%	17.20%
BASIC	17.72%	19.35%
DEFINITION	17.72%	15.05%
CAUSAL	11.39%	9.68%
OTHERS	13.92%	12.90%

- Most of questions need simple facts such as isa (instruction set architecture) knowledge and properties of objects, further confirming the need for simple reasoning with common knowledge

Contents

- Introduction
- OpenBookQA dataset
- Baseline model
- Baseline performance
- Discussions
- Conclusions

Baseline models

- No training, external knowledge only
- No training, core facts and external knowledge
- Trained models, no Knowledge
- Trained model with external knowledge

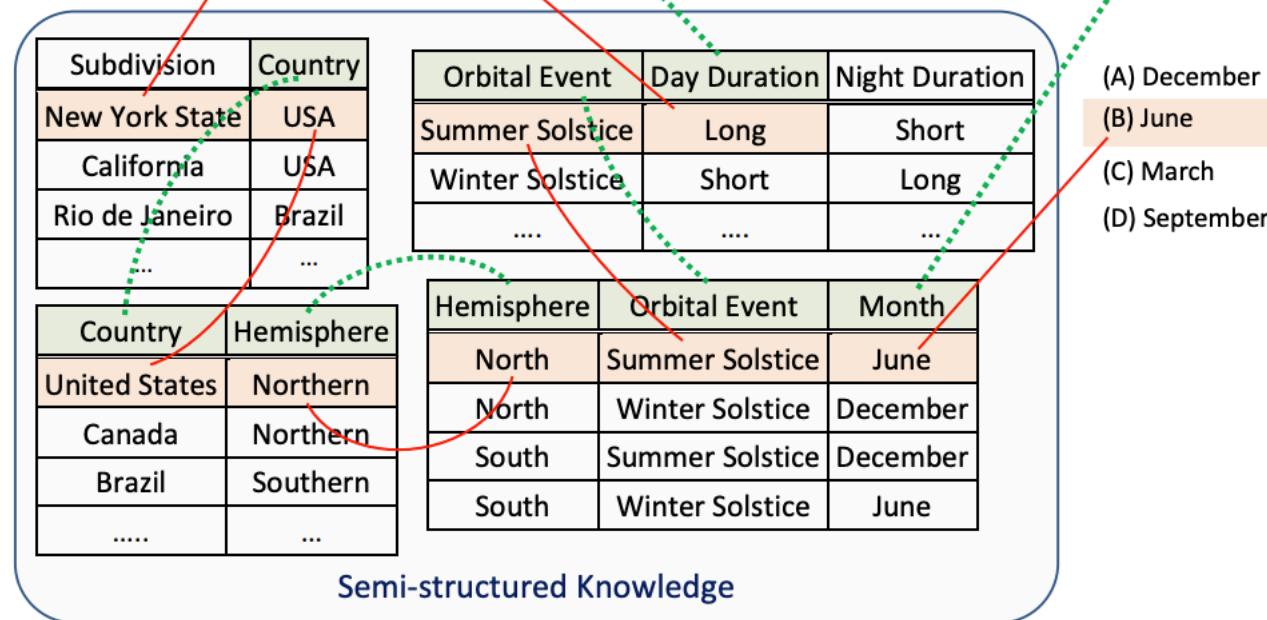
Baseline models

- No training, external knowledge only
 - **PMI** (Clark et al., 2016) uses pointwise mutual information (PMI) to score each answer choice using statistics based on a corpus of 280 GB of plain text.
 - **TableILP** (Khashabi et al., 2016) is an Integer Linear Programming (ILP) based reasoning system. It operates over **semi-structured relational tables** of knowledge. It scores each answer choice based on the optimal “support graph” connecting the question to that answer through table rows.
 - **TupleInference** (Khot et al., 2017), also an ILP-based QA system, uses Open IE tuples (Banko et al., 2007) as its **semi-structured representation**.
 - **DGEM** (Khot et al., 2018) is a neural entailment model that also uses Open IE to produce a **semi-structured representation**.

Baseline models

- No training, external knowledge only

Q: In **New York State**, the **longest period of daylight** occurs during which **month**?



- TableILP searches for the best support graph (chains of reasoning) connecting the question to an answer, in this case June.

Baseline models

- No training, core facts and external knowledge
 - IR solver (Clark et al. 2016).
 - TupleInference solver (Khot et al., 2017).

Baseline models

- Trained models, no Knowledge
 - Embeddings + Similarities as Features.

Experiments with a **logic regression model** that uses centroid vectors r_s^{emb} of the word embeddings of tokens in s , and then computes the **cosine similarities** between the question and each answer choice.

$$r_s^{\text{emb}} = \frac{1}{n_s} \sum_{j=1}^{n_s} e_{s_j} \in \mathbb{R}^d$$

$$r_{q,c_i}^{\text{cos}} = \cos(r_q^{\text{emb}}, r_{c_i}^{\text{emb}}) \in \mathbb{R}^1$$

Baseline models

- Trained models, no Knowledge

- BiLSTM Max-Out Baselines

First encode the question tokens and choice tokens $w_{1..n_s}^s$ independently with a bi-directional context encoder (LSTM) to obtain a context representation. $h_{s_1..n_s}^{ctx} = BiLSTM(e_{1..n_s}^s) \in \mathbb{R}^{n_s \times 2h}$. Then perform an element-wise aggregation operation max on the encoded representations $h_{s_1..n_s}^{ctx}$ to construct a single vector $r_s^{ctx} = \max(h_{s_1..n_s}^{ctx}) \in \mathbb{R}^{2h}$.

Apply solver algorithms utilizing the contextual representations.

- (a) Plausible Answer Detector
- (b) Odd-one-out solver
- (c) Question matching

Baseline models

- Trained model with external knowledge

Implement a two-stage model for incorporating external common knowledge, K .

The first module performs information retrieval on K to select a fixed size subset of potentially relevant facts $K_{Q,C}$ for each instance in the dataset.

The second module is a neural network that takes $(Q, C, K_{Q,C})$ as input to predict the answer to a question Q from the set of choices C .

Contents

- Introduction
- OpenBookQA dataset
- Baseline model
- Baseline performance
- Discussions
- Conclusions

Baseline performances

Solver	Dev	Test
Human solver	89.3*	91.7*
Guess All (“random”)	25.0	25.0
NO TRAINING, KB ONLY (§4.1)		
TupleInference	15.9	17.9
PMI (Waterloo corpus)	19.7	21.2
TableILP	20.0	23.4
DGEM	27.4	24.4
NO TRAINING, KB + \mathcal{F} (§4.2)		
IR with \mathcal{F}	25.5	24.8
TupleInference with \mathcal{F}	23.6	26.6
DGEM with \mathcal{F}	28.2	24.6
TRAINED MODELS, NO \mathcal{F} OR KB (§4.3)		
Embedd+Sim	44.6	41.8
ESIM	53.9±0.4	48.9±1.1
Plausible Answer Detector	54.4±0.7	49.6±0.7
Odd-one-out Solver	56.9±0.5	50.2±1.6
Question Match	54.6±1.2	50.2±0.9
ORACLE MODELS, \mathcal{F} AND/OR KB (§4.4)		
f	63.0±2.3	55.8±2.3
f + WordNet	57.6±1.4	56.3±1.3
f + ConceptNet	57.0±1.6	53.7±1.5
f + k	80.2±1.1	76.9±0.7

Contents

- Introduction
- OpenBookQA dataset
- Baseline model
- Baseline performance
- **Discussions**
- Conclusions

Discussions

- We observe that the best performance is ~76% among the baseline models, which is far behind the human performance at 92%. We can consider two points to improve the model's performance. 1) We need a better retrieval module to provide useful knowledge for a specific question; 2) we need to develop a multi-hop reasoning framework which can reason the concepts hiding in a question.

Contents

- Introduction
- OpenBookQA dataset
- Baseline model
- Baseline performance
- Discussions
- Conclusions

Conclusions

- This paper presents a new dataset, OpenBookQA, of about 6000 questions for open book question answering. This dataset requires simple common knowledge beyond the provided core facts, as well as multi-hop reasoning combining the two. With experiments of baseline models, this paper achieves an accuracy of 76% in answering the questions which is far from the human performance, so further studies are encouraged.