Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences

Daniel Khashabi[1], Snigdha Chaturvedi[2], Michael Roth[3], Shyam Upadhyay[1], Dan Roth[1]
[1]University of Pennsylvania, [2]University of California, Santa Cruz, [3]Saarland University

Motivation

- Machine Comprehension is a fundamental challenge in AI.
- Many challenge datasets in recent years: SQuAD, CNN/Daily Mail, NewsQA, etc.
- Current QA systems do not have abilities comparable to human, as evident from their brittleness.
- We believe that this is partly due to the absence of challenging datasets.
- We propose to address this shortcoming by developing a reading comprehension challenge that requires reasoning over multiple sentences.

- ~9,000 questions (6k are multi-sentence)
- Extracted from +800 paragraphs
- From 8 domains (fictions, news, science, social articles, Wikipedia, ...)

Website: https://cogcomp.org/multirc

Design Principles

Multi-Sentence-ness.
Questions in our challenge require models to use information from multiple sentences

Open-endedness.
Not restricted to verbatim answers in a paragraph. Answer-options can represent information that is not explicitly stated.

Answers to be judged independently.
The number of correct answer options is not pre-specified. Therefore, guessing an answer by elimination or choosing the top candidate is not sufficient.

Variability.
Our paragraphs are from multiple domains, leading to linguistically diverse questions and answers.

Example

S1: Most young mammals, including humans, play.
S2: Play is how they learn the skills that they will need as adults.
S6: Big cats also play.
S8: At the same time, they also practice their hunting skills.
S11: Human children learn by playing as well.
S12: For example, playing games and sports can help them learn to follow rules.
S13: They also learn to work together

Number of correct answers not specified

What do human children learn by playing games and sports?
A)* They learn to follow rules and work together
B) hunting skills
C)* skills that they will need as adult

Requires multiple sentences.

Dataset-collection Pipeline

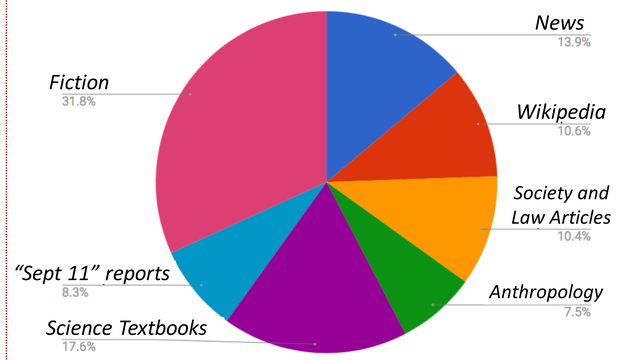
Generate Multi-sentence Questions → Ask 5 people to write 3-5 questions such that the question is answerable from the passage

Multi-Sentence-ness Verification → Ask turkers if they could answer the question, given a single sentence from the paragraph

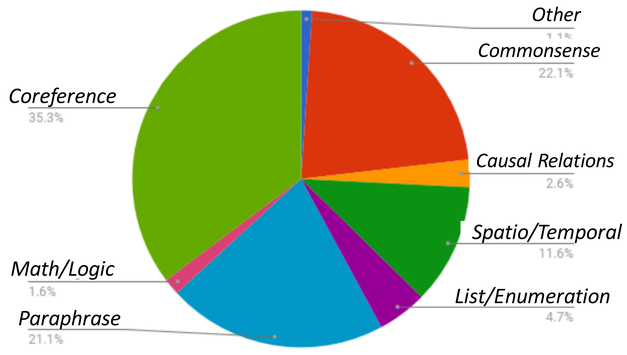
Generating Answer-Options → Ask 3 turkers to create as many correct/incorrect answer-option as they can, for each question.

Quality Verification → Ask turkers to indicate any errors (grammatical or otherwise) among questions and answer-options.

Domains used in the dataset



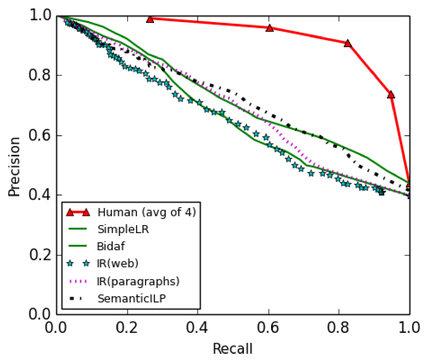
Distribution of phenomena in the dataset



Quantitative Evaluation

- **Random:** Expected score upon random selection answers.
- **IR:** Information retrieval baseline; each answer a is scored by sending a query $question+a$, to a Lucene engine that has indexed all the paragraphs in the training data.
- **SurfaceLR:** Logistic Regression classifier, with features modelling word-overlap and other shallow features.
- **Human:** Average of 4 human annotators

	Dev (F1)	Test (F1)
Random	44.3	47.1
IR	64.3	54.8
SurfaceLR	66.1	66.7
Human	86.4	84.3



PR curve for each of the baselines.

There is a considerable gap between the performances of baselines and humans.

Acknowledgements

Partly based on research sponsored by
- German Research Foundation (DFG EXC 284 and RO 4848/1-1)
- DARPA under contracts FA8750-13-2- 0008 and HR0011-15-2-0025
- Allen Institute for Artificial Intelligence (allenai.org)
- Google
- NSF grant BCS-1348522; and by NIH grant R01-HD054448.