Benchmarking Machine Reading Comprehension: A Psychological Perspective

Saku Sugawara[1], Pontus Stenetorp[2], Akiko Aizawa[1]

[1]National Institute of Informatics, [2]University College London

saku@nii.ac.jp

MRC for Research on NLU

- Natural language understanding (NLU) research aims to model a machine that can understand natural language (e.g., WSC (Levesque 2012), RTE (Dagan 2005))
- Machine reading comprehension (MRC) is one of NLU tasks, with a general form

Context: The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping.
She wandered out.
Finally she went into the forest where there are no electric poles.

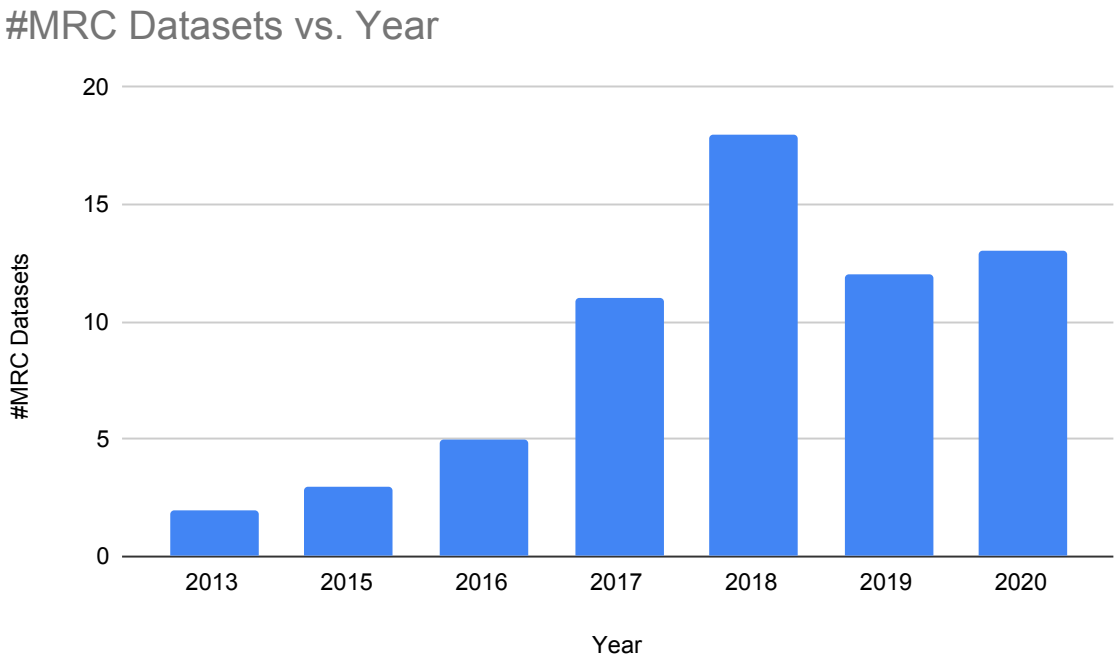
Question: Where did the princess wander to after escaping?
Answer: A) Mountain *B) **Forest** C) Cave D) Castle

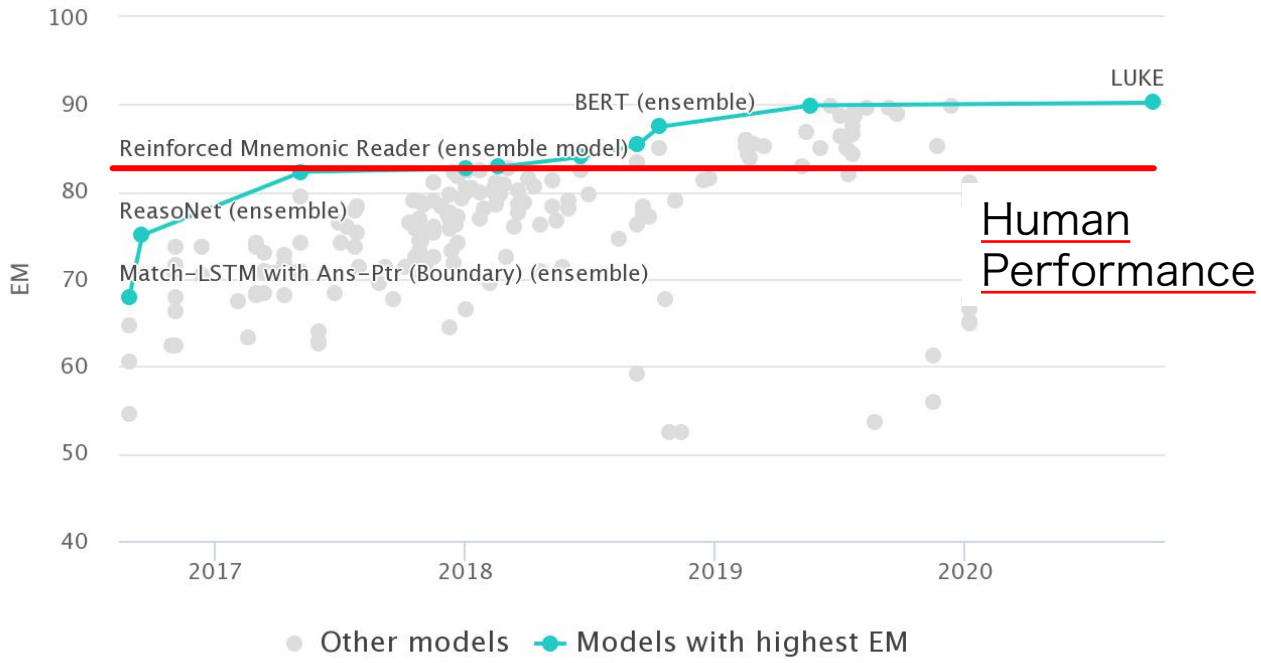
Coreference

Commonsense reasoning

Temporal relation

Trends and Progress of MRC





- ＞70 datasets (personal estimate)
- Various formats, domains, skills…

- Outperforms human in SQuAD v1.1

  = MRC is “solved”?

Benchmarking Issues: Analytic Studies

- Models for SQuAD are easily fooled by manually injected distracting sentences (Jia & Liang 2017)
- Questions are solvable even after shuffling context words or dropping content words (Sugawara+ 2020)
- >90% of the questions in SQuAD v1.1 require reading only one sentence in passage (Min+ 2017)
- Questions are solvable only with a few question tokens (or none) (Sugawara+ 2018, Feng+ 2018, Mudrakarta+ 2018, Kaushik & Lipton 2018)
- Multi-hop reasoning datasets do not necessitate multi-hop reasoning (Min+ 2019, Chen & Durrett 2019)

Q: What understanding is required by the datasets and is actually achieved by models?

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

Adversarial example (Jia & Liang 2017)

Paragraph: [redacted] many persons [redacted] cannot afford [redacted] buy books, [redacted] usually go [redacted] libraries and spend hours reading something [redacted] interests [redacted] lot. [redacted] point [redacted] view, literature [redacted] important [redacted] life. [redacted] example, reading [redacted] means [redacted] gaining culture [redacted] enriching [redacted] knowledge [redacted] different areas .

Q: People who are fond of literature are those that ____ .
A: have much interest in reading (multiple choice)

Content word dropping (Sugawara+ 2020)

Assumptions, Goal, and Research Questions

Assumptions

- To benchmark NLU, we need to ensure the explainability of tasks in human terms
- Interpreting models may be insufficient for explaining how the task is accomplished

Goal

From a top-down perspective (Bender & Koller 2020)

- Investigate a theoretical foundation for better benchmarking of MRC (or NLU)

Research Questions


What

- Q: **What** does reading comprehension involve?
  → Computational model of reading comprehension in psychology


How

- Q: **How** can we evaluate reading comprehension?
  → Validity of interpreting measurements in psychometrics

Future Directions

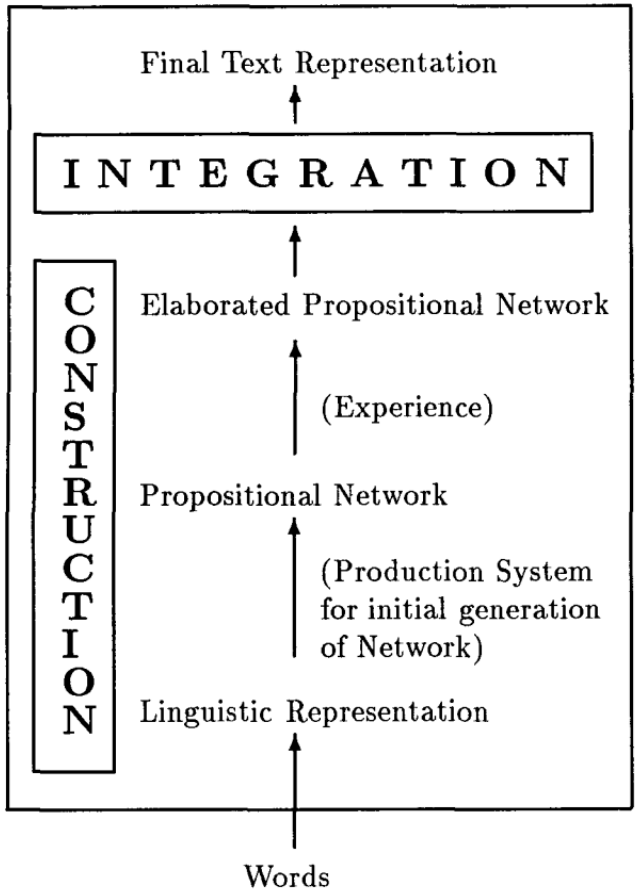

Future

What Question: Text Comprehension in Psychology

Construction-Integration Model (Kintsch 1986) / Situation Model (Zwaan & Radvansky 1998)

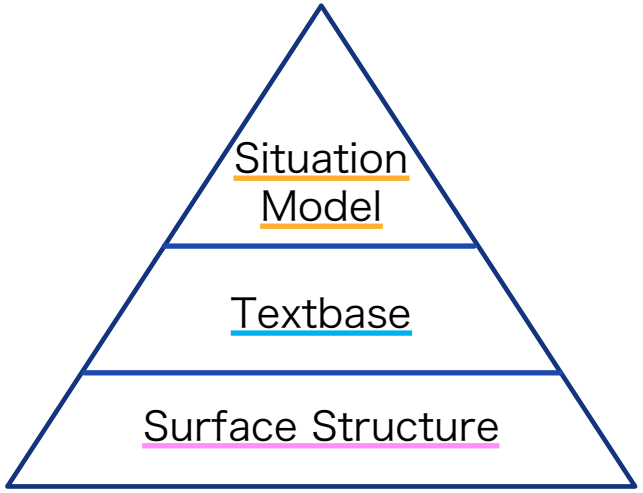
1. Construction
   - Given raw textual input, create a propositional network in which propositions are adjacently connected & elaborated
2. Integration
   - Using the propositions, create a coherent representation (situation model) where propositions are organized globally, sometimes grounded to other media

Hernández-Orallo (2017): (successful) comprehension is the process of searching for a situation model that best explains the given text and the reader's background knowledge



From Wharton & Kintsch (1991)

Representation Levels and NLP Tasks



1. Surface Structure Level -> “checklist” approach (Ribeiro+ 2020)
   - Linguistic propositions from the textual input
   - Skills: syntactic parsing, POS tagging, SRL, and NER
2. Textbase Level -> “skill set” approach (e.g., Rogers+ 2020, Wang+ 2019)
   - Local relations of propositions
   - Skills: recognizing relations between entities and sentences such as coreference, factual knowledge, and discourse relations
3. Situation Model Level -> Future directions (discuss later)
   - Global structure of propositions
   - Skills: creating a coherent representation and grounding it to other media (sound, image, ...)

Representation Levels: Example

Passage: The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping. She wandered out. Finally, she went into the forest where there are no electric poles.

Q1. Who climbed out of the high tower?
A1. *Princess*

Q2. Where did the princess wander after escaping?
A2. *Forest*

Q3. What would happen if her mother was not sleeping?
A3. *The princess would be found soon* (multiple choice)

Q1. Surface structure level

- Understanding the subject of the first event.

Q2. Textbase level

- Understanding of relations among described entities and events:
  - Coreference: “she” = “princess”
  - Commonsense: “escaping” = the first event
- Solvable only by looking for a place specified by *where* ? -> validation needed!

Q3. Situation model level

- Imagining a different situation

How Question: Construct Validity in Psychometrics

Construct (psychology): an abstract concept used to facilitate understanding of human behavior

Construct Validity (Messick 1995)

- Evidence (or criteria) that is necessary to validate the interpretation of outcomes of psychological experiments.

Six Aspects of Validity in MRC (or NLU?)

1. Content aspect
   - Wide coverage of representations
2. Substantive aspect
   - Evaluation of the internal process
3. Structural aspect
   - Structured evaluation metrics
4. Generalizability aspect
   - Reliability of evaluation metrics
5. External aspect
   - Consistency with external variables
6. Consequential aspect
   - Robustness to adversarial attacks and reducing social biases

A Rubric Matters!

What is a rubric?

- A scoring guide used for assessments in education (Popham 1997)
- Including evaluation criteria, quality definitions, and scoring strategy

Ideally, a rubric needs to cover... (for example)

- (1) Content aspect
  - Does the task have sufficient coverage of linguistic phenomena over the representation levels?
- (2) Substantive and (3) structural aspects
  - Are questions ensured to evaluate the internal process and/or have corresponding metrics?
- (4) Generalizability and (5) external aspects
  - Are models performing well on your dataset good at out-of-domain datasets and other NLU tasks?
- (6) Consequential aspect
  - Does the task check models' robustness to adversarial inputs and their unintended biases?

Future Directions (1) Situation Model

1. Evaluating Situation Model

- Context-dependent situations
  - Representations are constructed distinctively depending on the given context
  - Defeasibility: if-then reasoning (Sap+ 2019), abductive reasoning (Bhagavatula+ 2020)
  - Novelty: StrategyQA (Geva+ 2021)

- Grounding in non-textual information
  - Images: Visual MRC (Tanaka+ 2021), Visual Commonsense Reasoning (Zellers+ 2019), Science textbooks (Kembhavi+ 2019), FigureQA (Kahou+ 2018), (but more focus on text?)
  - Structured data: HybridQA (tabular) (Chen+ 2020), Knowledge Base (many...)

Future Directions (2) Substantive Validity

2. Assuring Substantive Validity: Better Evaluation of Internal Processes

- Collecting high-quality, shortcut-free questions
  - Removing shortcuts (Geirhos+ 2020) by post-processing (e.g., in ReClor)
  - Alleviating dataset-specific and word-association biases (Sakaguchi+ 2020)
- Formulating an explanation-by-design task
  - Introspective explanation: R^4C (Inoue+ 2020), which consists of "derivation" for answering questions—not only supporting sentences (e.g., HotpotQA)
  - Creating question dependency: ProPara (Dalvi+ 2018), CoQA (Reddy+ 2019)

Summary

Background & Motivation

- MRC could be a good benchmarking task for natural language understanding
- What understanding is required by the datasets and is achieved by models?

Research Questions and Tentative Answers

What

- Q: What does reading comprehension involve?
- A: Process of creating a situation model that best explains context; Computational model of human text comprehension may be useful

How

- Q: How can we evaluate reading comprehension?
- A: Designing a rubric that covers various aspects of construct validity