EACL 2021 #661

saku@nii.ac.jp

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

1. Background and Motivation

Assumptions, Goal, and Research Questions

Assumptions

- To benchmark NLU, we need to explain how the task is accomplished
- Interpreting models may be insufficient for the explainability of tasks

Goal

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

- Investigate a theoretical foundation for better benchmarking of MRC

Research Questions

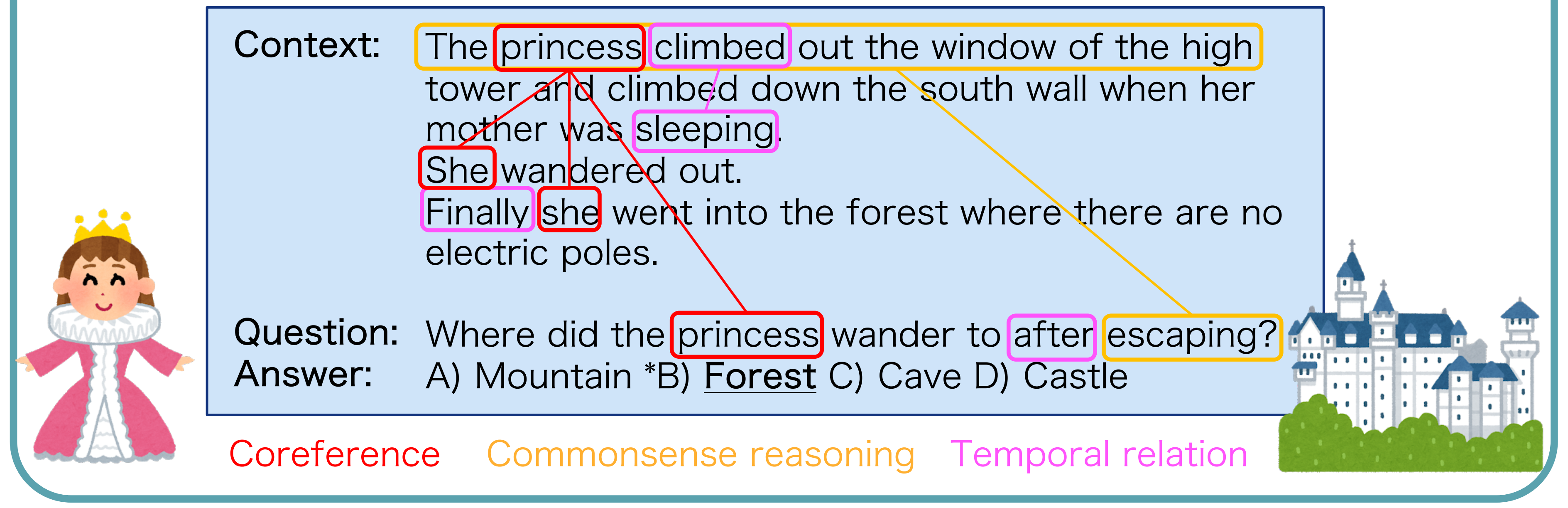
- Q: What does reading comprehension involve?
  → Computational model of reading comprehension in psychology
- Q: How can we evaluate reading comprehension?
  → Validity of interpreting measurements in psychometrics

What

How

Machine Reading Comprehension Task

- Machine reading comprehension is one of NLU tasks with a general form, which could ask various NLP tasks to answer questions.



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

Benchmarking Issues: Analytic Studies

- Models for SQuAD are **easily fooled by manually injected distracting sentences** (Jia & Liang 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)
- Questions are **solvable even after shuffling context words or dropping content words** (Sugawara+ 2020)

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

2. What is Reading Comprehension?

What

Text Comprehension in Psychology

Construction-Integration Model (Kinsch 1986)

1. Construction: create a propositional network from given raw text
2. Integration: create a globally coherent representation

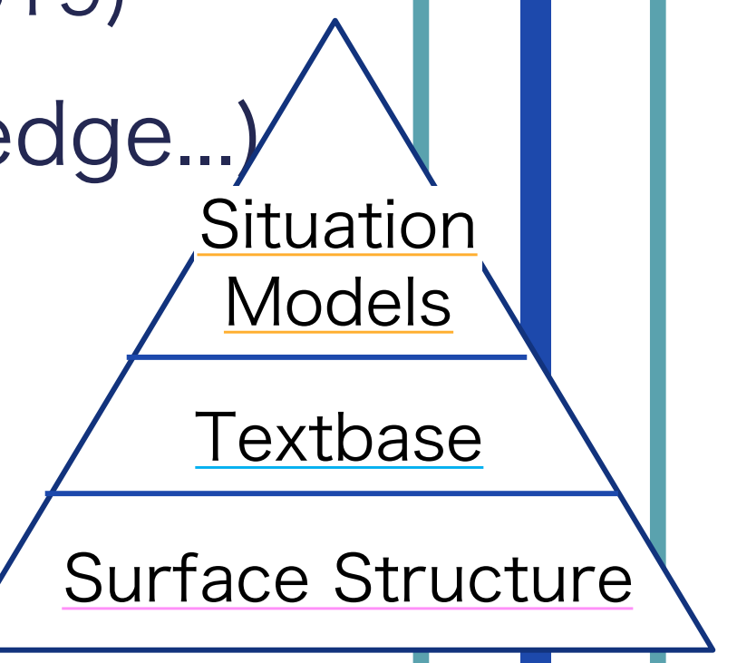
Situation Models (Zwaan & Fadvansky 1998)

- Integrated mental representations of a described state of affairs

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

Representation Levels and NLP Tasks

1. Surface Structure Level -> "checklist" approach (e.g., Ribeiro+ 2020)
   - Linguistic propositions from the textual input (parsing, tagging...)
   - Q: *Who climbed out of the high tower?*
2. Textbase Level -> "skill set" approach (e.g., Rogers+ 2020, Wang+ 2019)
   - Local relations of propositions (sentence relation, factual knowledge...)
   - Q: *Where did the princess wander to after escaping?*
3. Situation Model Level -> Future directions
   - Global structure of propositions (situation model, grounding...)
   - Q: *What would happen if her mother was not sleeping?* A: she would be caught...



3. How Can We Evaluate It?

How

Construct Validity in Psychometrics

Construct Validity (Messick1986)

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

Six Aspects of Validity in MRC

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

A Rubric Matters!

Rubric

- A scoring guide used for assessments in education (Popham 1997)

Ideally, a rubric for MRC needs to cover...

- (1) Content aspect
  - Does the task have sufficient coverage of linguistic phenomena?
- (2) Substantive and (3) structural aspects
  - Do questions evaluate the internal process and have corresponding metrics?
- (4) Generalizability and (5) external aspects
  - Are models performing well on your dataset good at out-of-domain datasets?

4. Future Directions: Situation Models and Substantive Validity

Situation Models

Context-dependent situations

- Representations are constructed 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 (and so on...)

Substantive Validity

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^{4}C (Inoue+ 2020) asks about "derivations" for answering questions—not only supporting sentences (e.g., HotpotQA)
- Creating question dependency: ProPara (Dalvi+ 2018), CoQA (Reddy+ 2019)