Benchmarking Natural Language Understanding: A Psychological and Philosophical Perspective?

Saku Sugawara (National Institute of Informatics)

February 9, 2022 @ NLP Colloquium

saku@nii.ac.jp

本日のおはなし

1. 言語理解はいかにして benchmark の対象になるか [WIP]
   - 言語の理解って結局なんなの？
   - それを benchmark するってどういうこと？

哲学の人との共同研究です

2. じゃあなにをどうやって benchmark するのがよさそうか [EACL 2021]
   - あくまでひとつの考え方としての心理学的理論
   - 妥当性を向上させるための方法論

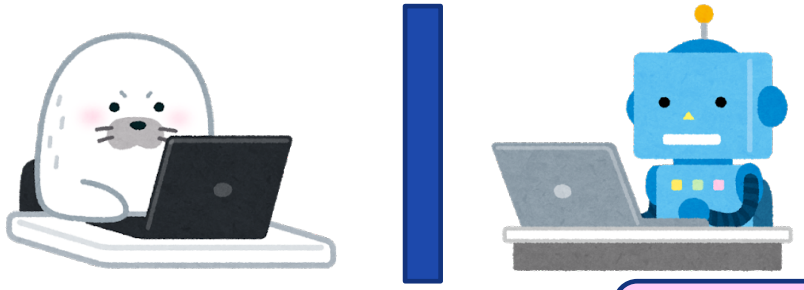
3. データセット作りのためのあれこれ [ACL 2021]
   - じわじわ進めているのでちょっとだけ

NYU の人との共同研究です

おことわり：全体的に議論が雑です＆文章が多くてすみません

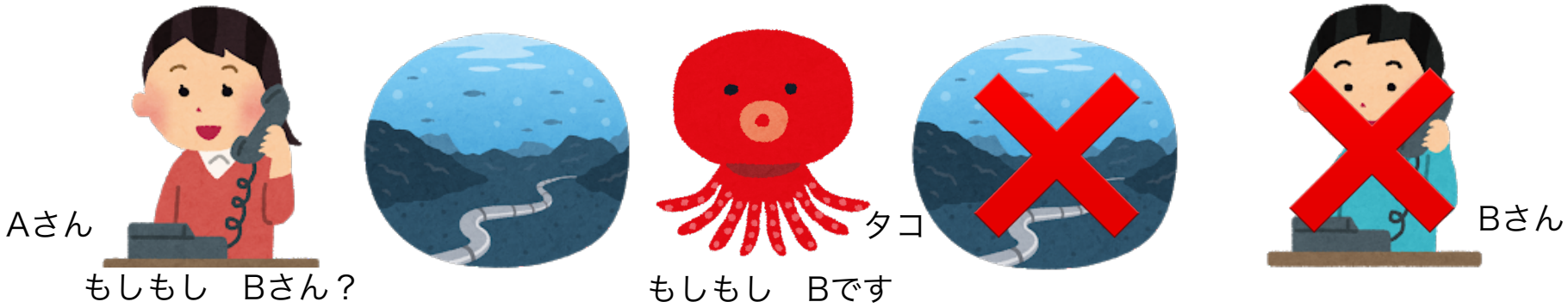
Turing Test と Octopus Test

- Turing Test (Turing 1950)
  - 相手から見えない状態で機械だとバレないように会話を続けられたら「知性」



- Octopus Test (Bender & Koller 2020)
  - Grounding ができないタコは相手を欺けないので言語理解には grounding が必須

言語モデル的な学習・入出力しかできない例として
海底にいるタコで Turing Test をするという設定



言語理解の反応依存性

- Turing Test の解釈 (Proudfoot 2020)
  - 原論文の中で Turing は「何かが知的に振る舞っているとどの程度みなされるのかは観測する我々の心的状態によって決まるよね」と述べている
  - Proudfoot はこれをもとに知性は response-dependent （反応依存的）な性質を持つと述べ、次のように定式化した：

    X is intelligent if, in an unrestricted computer-imitates-human game, X appears intelligent to an average interrogator.

- とくに重要なのはこれが「十分条件である」という点で、「Turing Test にパスできないから知性を持たない」とは言えないことに留意する必要がある
  - 人間もまた Turing Test にパスできないことが普通にありえる
- あくまで擬人的な振る舞いができるかどうかの欺きのテストなので実用的ではないよねという指摘がこれまで様々なされてきました (e.g., Hayes & Ford 1995)

言語理解の定式化？

- 同様に考えると、言語理解もまたおそらく反応依存的な振る舞いだと考えられそう：

X *understands language* if and only if, in a test, X appears to *understand language* to an average observer.

Turing Test ではない何らかの behavioral test のもとでの必要十分条件をひとまず考えたい

- 反応依存性をもうすこし書き下すと：

X understands language if and only if an average observer's subjective probability for the hypothesis "X understands language" is higher than a certain threshold where the observer can use the result of a test as evidence to support the hypothesis.

テスト結果はあくまで反応依存性に影響する手がかりでしかない

この定式化で何もかも大丈夫……？
→ 実用の上では科学的・社会的に合意が取れてほしいけど、このままだと難しい！

なぜ benchmark をするのか

- 反応依存性は個人的すぎる
  - 人間の集合がなるべく合意できそうなものに拡張する必要がある

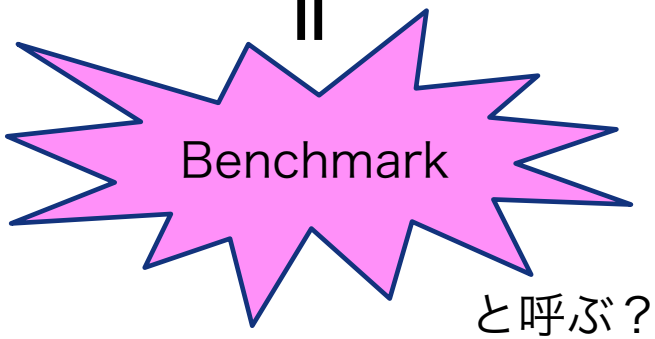
    an average observes → “average observers”

- 「言語理解」が広すぎる（多様な振る舞いが理解だとみなされうる）
  - タスクやドメインを限定する必要がある

    linguistic behavior in “some domain”
  - ある程度の人間ならだいたいできるよね、という基準に近い

両者を満たして反証可能・
説明可能な証拠を得る手段
としての言語理解のテスト

=



と呼ぶ？

MRC for Benchmarking Language Understanding

- 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

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)

Function word dropping (Sugawara+ 2020)

Assumptions, Goal, and Research Questions

合意としての理解に至るために情報を増やしたい！という感じ

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

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

Goal

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

Research Questions

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


What

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

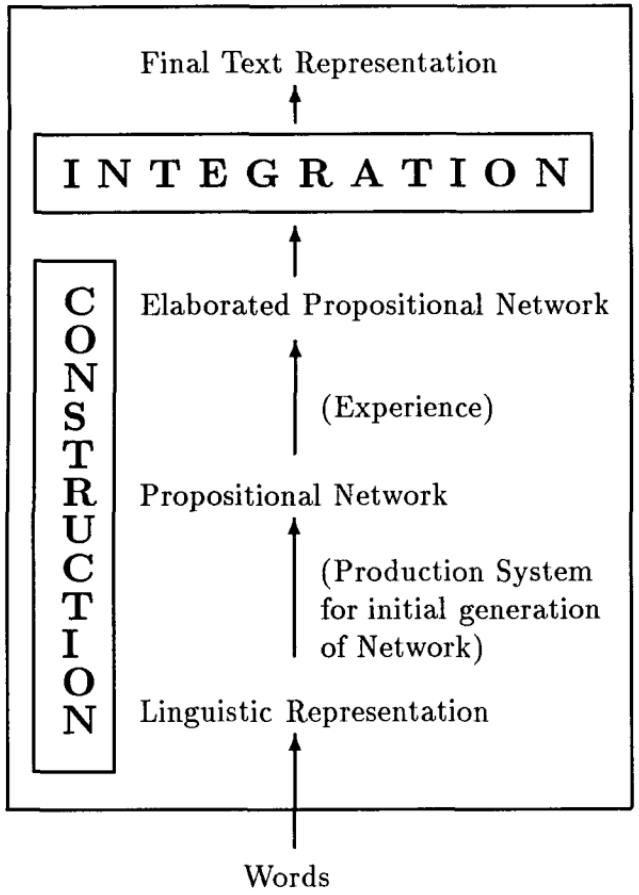

How

What Question: Text Comprehension in Psychology

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

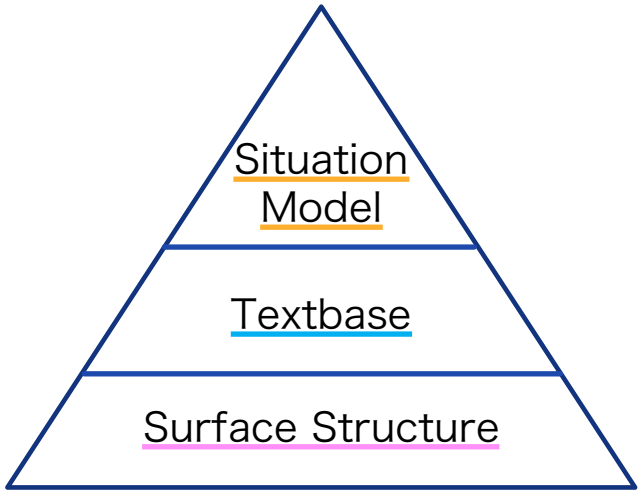
1. Construction
   - Given raw textual input, create a propositional network in which propositions are adjacently connected & elaborated
2. Integration
   - Using the propositions, create a coherent representation (situation model) where propositions are organized globally, sometimes grounded to other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

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?

クラウドソーシング方法論

- 目標
  - 高品質（パターンマッチで解きづらく、差分が見える程度に難しい）な質問を集めたい
  - さまざまな種類の質問を集めたい（評価指標・内容的妥当性）
  - できれば効率よく、少ない費用で多くのデータを集めたい

- 課題
  - どのように「ちゃんと作業してくれる作業者」を見つけるか？
  - 成果物の品質を上げるためにはどのような指示・フィードバックが必要か？

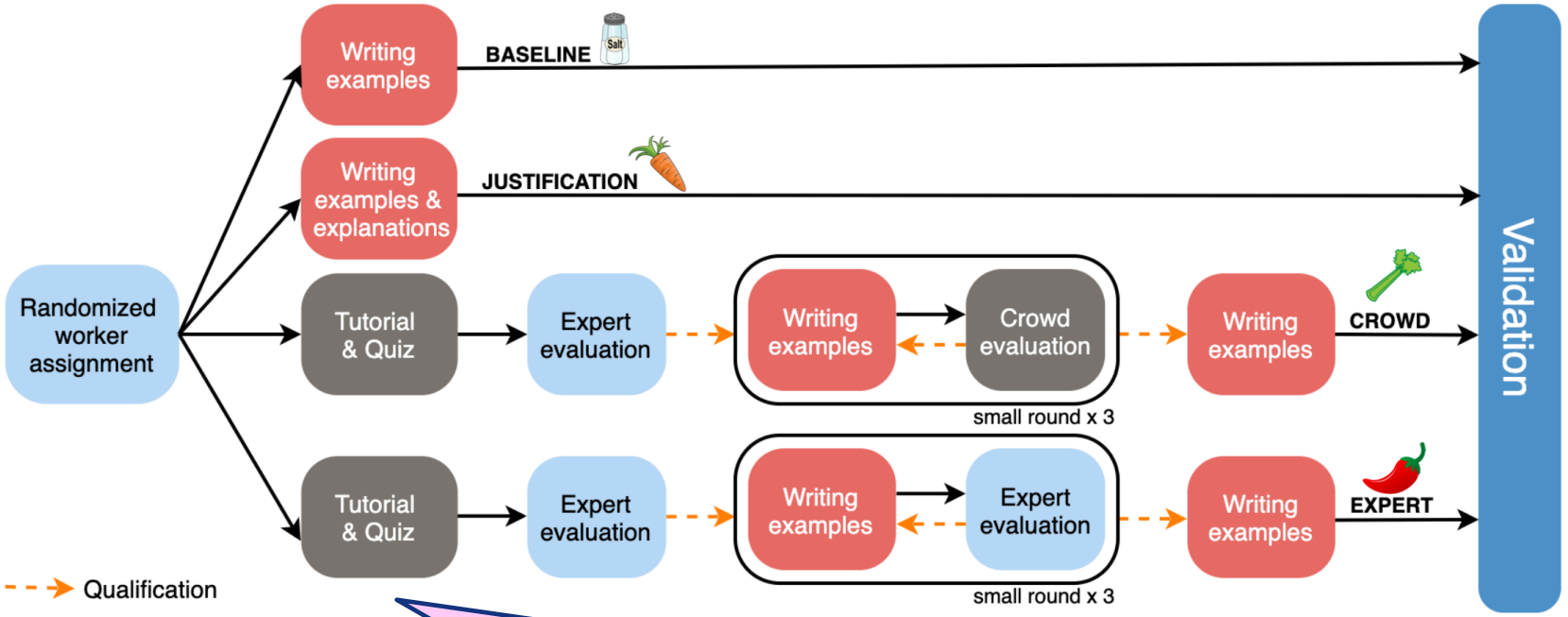
} ACL 2021

クラウドソーシング方法論：プロトコル比較

(Nangia & Sugawara+, ACL 2021)

Q: どういった収集方法なら高難易度かつ一致率の高い質問を集めることができるか？

やったこと：異なるプロトコルで読解問題を収集し、質問の難易度・一致率を比較



質問の難易度
＝ 人間の精度 ー システムの精度

様々な手法を比較しながら
利便性の高いデータ収集
ツールを作成

Tutorial & Quiz を課して filtering

クラウドソーシング方法論：プロトコル比較

(Nangia & Sugawara+, ACL 2021)

観察

- 「なぜ難しいか」の説明を作問と同時に依頼 (justification) ➡ 難易度に変化なし❌
- 作業者の上位2割のみデータ作成に参加 (expert/crowd) ➡ 難易度・一致率が向上 ✅
- 作業内容のフィードバック ➡ 作業者 (crowd) より専門家 (expert) が行うと一致率向上 ✅

Crowdworker でもある程度は質を担保できる

今後の展開

- 高難易度と言ってもシステム精度 95% vs 90% ほどの差 😢 ➡ どのような**文章**なら難しくなる？
- 難しい問題のほうが選択肢が長い傾向 🤔 ➡ より**長い選択肢**の問題を強制すると難しくなる？

本日のおはなしでした

1. 言語理解はいかにして benchmark の対象になるか [WIP]
   - 言語の理解って結局なんなの？ → 観測する人の反応に依存する
   - それを benchmark するってどういうこと？ → 集団で合意が取れるように証拠を得る手段

2. じゃあなにをどうやって benchmark するのがよさそうか [EACL 2021]
   - あくまでひとつの考え方としての心理学的理論 = situation model や repr. level の観点
   - 妥当性を向上させるための方法論 = construct validity の考え方

3. データセット作りのためのあれこれ [ACL 2021]
   - じわじわ進めているのでちょっとだけでした → 手段の違いで得られるデータの比較分析 etc.