説明性の高い自然言語理解ベンチマークの構築

2022-03-08

菅原 朔（国立情報学研究所）

IBISML 研究会

研究テーマ：自然言語理解

- **目標**
  - 自然言語の理解を計算論的にモデル化する

→ 人間の言語理解を追究 ＆ 人間の言語活動の支援

（様々な応用技術の基礎の基礎）

対話システム

→言われてることちゃんとわかってる？

情報抽出・
ファクトチェック

→書いてあること
ちゃんとわかってる？

説明性の高い自然言語理解のベンチマーク

Fact and Mythmaking Blend in Ukraine's Information War

Experts say stories like the Ghost of Kyiv and Snake Island, both of questionable veracity, are propaganda or morale boosters, or perhaps both.

The Dnieper River embankment in the Podil neighborhood of Kyiv last Thursday. Brendan Hoffman for The New York Times

March 3, 2022

Just days into the Russian invasion of Ukraine, a pilot with a mysterious nickname was quickly becoming the conflict's first wartime hero. Named the Ghost of Kyiv, the ace fighter had apparently single-handedly shot down several Russian fighter jets.

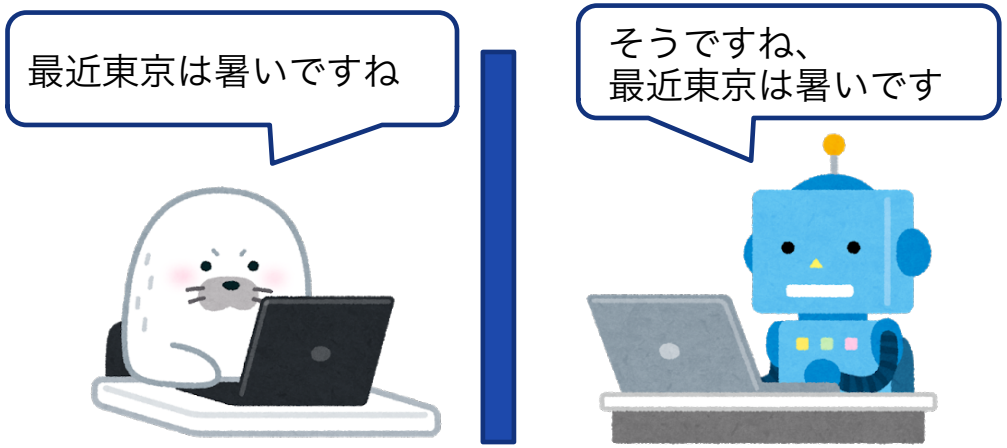
The story was shared by the official Ukraine Twitter account on Sunday in a thrilling montage video set to thumping music, showing the fighter swooping through the Ukrainian skies as enemy planes exploded around him. The Security Service of Ukraine, the country's main security agency, also relayed the tale on its official Telegram channel, which has over 700,000 subscribers.

source: NY Times

評価タスクの歴史

- チューリングテスト (1950-)
- 質問応答 (1960s-)
- 含意関係認識 (2005-)
- Winograd Schema Challenge (2011-)

- 機械による読解 (2013-)
  (machine reading comprehension)



チューリングテスト

前提: A woman selling bamboo sticks talking to two men on a loading dock.
仮説: There are at least three people on a loading dock.
含意: Yes

含意関係認識

The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.

Q: Who [feared/advocated] violence?
A: the city councilmen / the demonstrators

Winograd Schema Challenge

機械読解タスクとは

含意関係は単文対だったが、
読解は複数文にわたる文脈的な理解を問える

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 good ways.
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

照応解析　　常識推論　　時間関係の認識

→ 言語理解タスクの上で様々な技術を（**独立でなく同時に**）評価できる

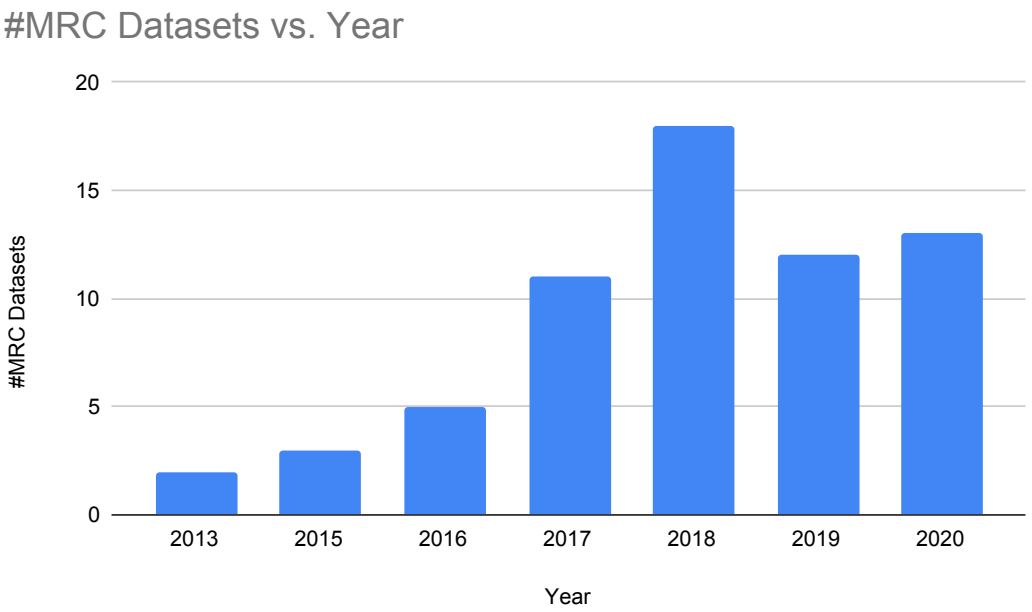
主要なデータセット：SQuAD (2016) [paper]

- 与えられた文章に関する質問に答える。回答は文章中の連続した単語列から抜き出し

Passage:	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".
Question:	What causes precipitation to fall?
Answer:	**gravity**
Question:	What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
Answer:	**graupel**
Question:	Where do water droplets collide with ice crystals to form precipitation?
Answer:	**within a cloud**

- Stanford Question Answering Dataset
- Passage は Wikipedia から
  - CC-BY で使いやすいデータとして Wikipedia が使われることが多い
- 質問はクラウドソーシングで収集
- ひとつの passage が複数の質問をもつ
- 2つのバージョン
  - v1.1 (2016): 100k examples
  - v2.0 (2018): +50k "unanswerable" examples

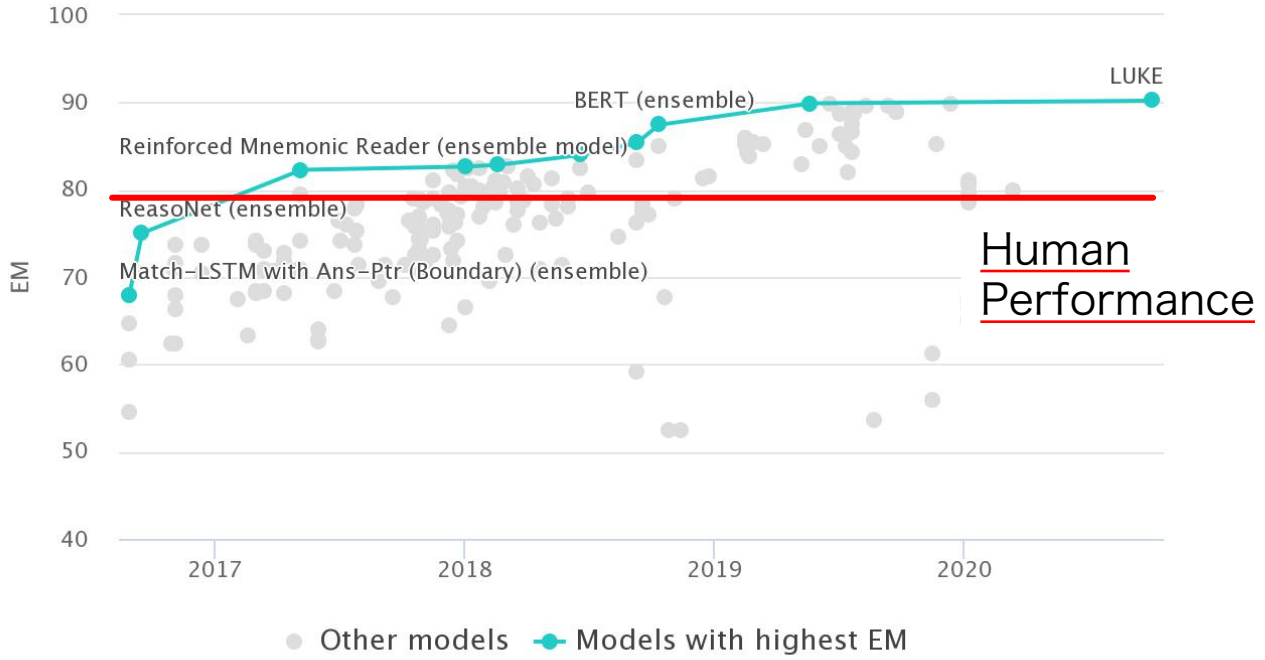
機械読解タスクの流行とシステムの発展



データセットの増加

- 2013年ごろから 70 以上
- 様々な問題形式、ドメイン、言語
- どれも大規模（>10k が普通）

SQuAD v1.1
paperwithcode



システムの性能向上

- SQuAD v1.1 で人間を上回る (2018)

  これに限らず、多くのベンチマークで人間に肉薄

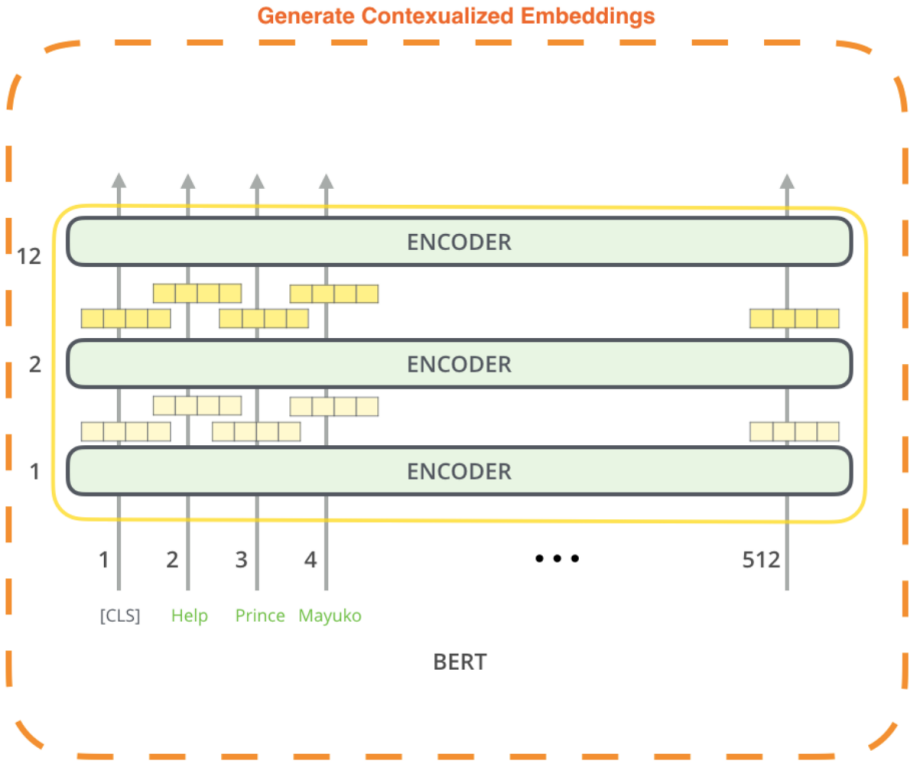
さまざまなデータセットがあります

- 文脈情報
  - 100-200単語の 1 パラグラフ、複数パラグラフの集合、長い文章、（情報検索的な）文書集合 etc
  - 画像や構造情報（テーブルや DB など）を含んでいるものも（マルチモーダル）
- 質問形式
  - 自然な質問文、穴埋め（自動的に作りやすい）、検索クエリ（の流用）
- 回答形式
  - 文章から抜き出し、選択肢（Yes/No 含む）、生成
- 質問・回答の収集方法
  - 研究者・専門家の執筆、クラウドソーシング、自動生成、試験問題等のリソース
- ドメイン・言語
  - Wikipedia、ニュース記事、小説（Project Gutenberg など）、医療系・バイオ系特化など
  - 英語が最も多いが、中国語も増えてきた　それ以外の言語はあまりないかも（multilingual でひとまとめ）

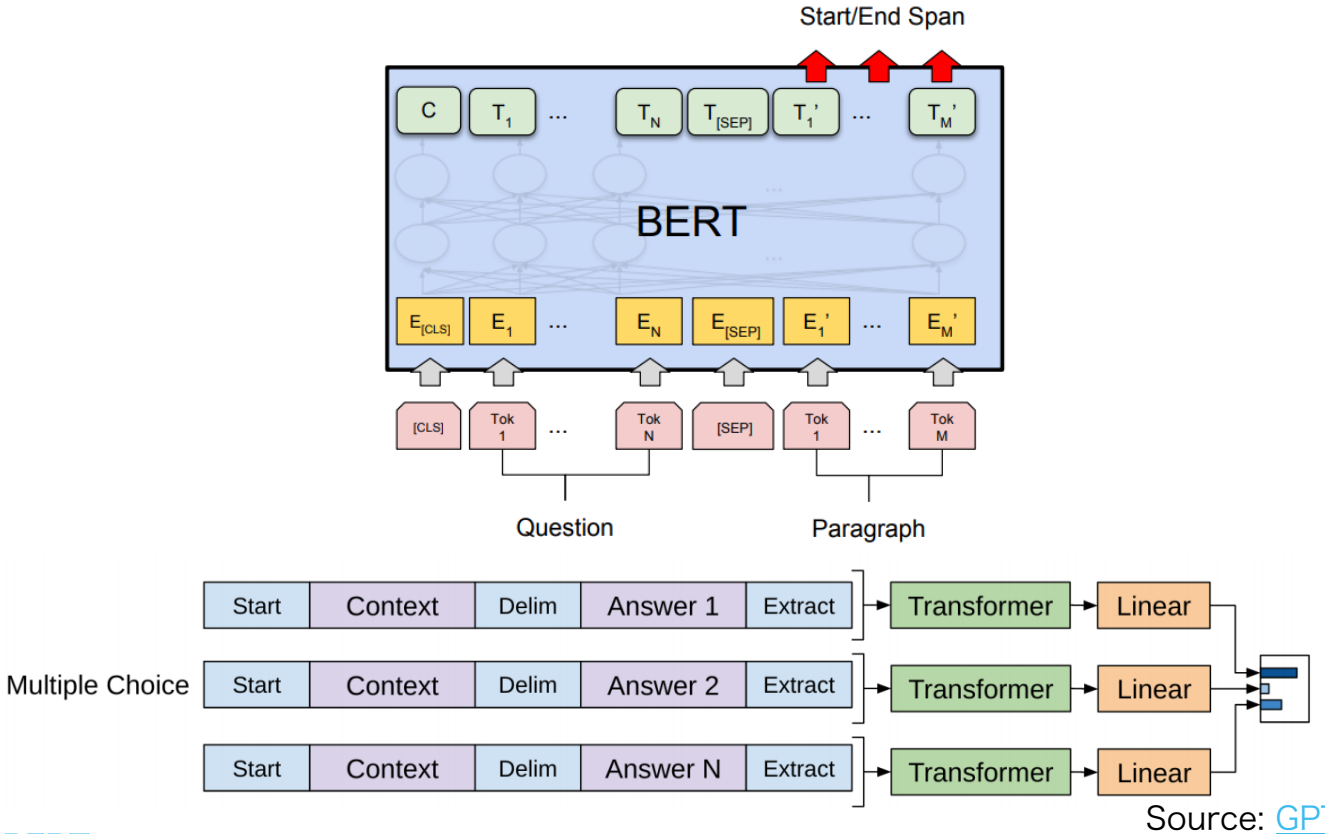
最近よかったサーベイは
Rogers+ (2021) [link]など

システムはどうやって解いているか

- Transformer が連なったモデルを大規模コーパスで事前訓練してエンコーダとして利用（長さは維持）
- エンコードされた情報を入力にして文章から抜き出し・選択肢ごとの確率を出力・テキスト生成など



Source: illustrated BERT



Source: GPT

課題1. 評価指標の単純さ

データセット全体で
正答との一致精度のみ

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 May 21, 2019	**XLNet (single model)** *Google Brain & CMU*	**89.898**	**95.080**
2 Oct 05, 2018	**BERT (ensemble)** *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
3 May 14, 2019	ATB (single model) *Anonymous*	86.940	92.641
4	KT-NET (single model)	85.944	92.425

← 人間のスコア

Transformer 系
深層学習モデル

単一の正答率だけ用いることの何が問題か

- 機械読解の長所は「さまざまな技術の総合タスク」になりうることだが……

	System X
Q1	✓
Q2	×
Q3	✓
⋮	⋮
Q1000	×
Acc.	75.0%

照応解析？
常識推論？
時間関係認識？
何が得意/苦手？

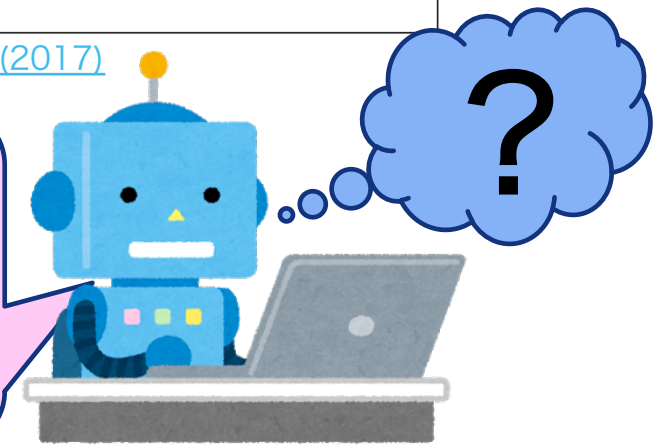
問題を解くためにどのような技術が必要なのか特定されておらず、
指標として精緻な情報が得られない（＝言語理解的な仮説・説明項の欠如）

課題２. データセットの品質保証

- 敵対的事例が容易に構築でき、高度な理解をしなくても多くの質問に正答できる (Jia and Liang, 2017)
- 質問文・文脈文の大半を読まずに解けるような過度に簡単な質問が存在 (Kaushik+, 2018; Sugawara+, 2018)
- データセットが意図している能力が実際は要求されない場合が多い (Gururangan+, 2018; Min+, 2019)
  - Annotation artifact や shortcut と呼ばれたりします

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

Source: Jia and Liang (2017)



意図した高度な理解や解法を必要とする品質でない可能性があり、
言語理解を正しく評価できているか不明瞭
（＝テスト理論的な真正性・説明の忠実性の欠如）

入力に編集を加えても解けることがある (Sugawara+, AAAI 2020)

課題文の語順シャッフル

C: Chris Ulmer, the 26-year-old teacher in Jacksonville starts his class by calling up each student individually to give them much admiration and a high-five. I couldn't help but be reminded of Syona's teacher and how she supports each kid in a very similar way.
Q: What can we learn about Chris Ulmer?
A: He praises his students one by one (multiple choice)

C: his help a in calling class but Syona's starts each 26-year-old similar individually Ulmer, and Chris admiration way. Jacksonville kid much I by couldn't them the a to supports of in student and teacher each be teacher reminded give how she high-five. up very
Q: What can we learn about Chris Ulmer?
A: He praises his s

POS タグ + ID による語彙の匿名化

C: Immediately behind the basilica is the Grotto, a Marian place of prayer. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858.
Q: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?
A: Saint Bernadette Soubirous (text span selection)

C: @adverb1 @prep5 @other0 @noun17 @verb2 @other0 @noun20 […] @other0 @noun20 @prep6 @noun25 @punct0 @noun26 @wh0 @other0 @noun7 @noun8 @adverb3 @verb4 **@prep4** @noun27 @noun28 @noun29 **@prep2** **@number0** @period0
Q: **@prep4** @wh2 @verb6 @other0 @noun7 @noun8 @adverb4 @verb4 **@prep2** **@number0** @prep2 @noun25 @noun26
election)

解けているのはすごいけど、何が評価できたのかわからない？

現在の研究：評価用データセットの構築

発表者のこれまでの研究

- 読解に必要な能力の提案（AAAI 2017）
- 能力と読みやすさの関係の分析（ACL 2017）

どう説明するか

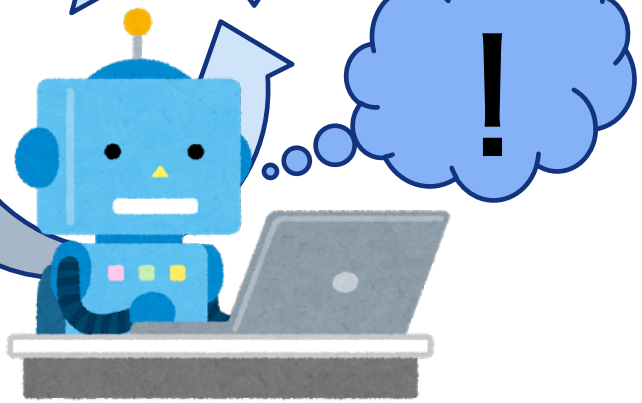
評価指標に貢献

- 問題の難易度の自動分類（EMNLP 2018）
- 能力の必要性の自動分析（AAAI 2020）

その説明の裏付け

品質保証に貢献

精緻・高品質なベンチマーク



既存のデータセットで不明瞭だった「これが解けることで何ができるようになったか」に説明を与えるベンチマークを目指す

ベンチマークが満たすべき性質 (Bowman and Dahl, 2021) [paper]

SNLI や GLUE の人

1. 内容的な妥当性を上げるために言語的な内容や振る舞いの**多様性を高める**
   - どのようにその多様性を測るかは難しい
2. アノテーションにおける**エラーや曖昧さ**をできる限り**排除**する
   - 本質的な曖昧さ、つまり人間においても揺れるものはそのまま取り込むべきという方向性もある。ここで指摘しているのはあくまでアノテーション由来のもの
3. 識別性・信頼度を上げるため**難易度を上げる・サンプル数を増やす**
   - 近年のベンチマークにおける性能はほとんど飽和している
4. 社会的**バイアスを助長しない**ようなデータになっている
   - バイアスに関する評価指標も必要

評価用データセットの構築：まず素材集め

最近のシステムは賢いのでそもそも
パターンパッチで解けないものを
集める時点で大変

① 素材となる
データの収集

② 人手による能力分析・品質保証

③ データセットの整備・公開

評価用データセットの構築：データの集め方

- クラウドソーシング (Amazon Mechanical Turk など) を利用したデータ収集
- API を利用して大規模化・インタラクティブに収集できるようにしています

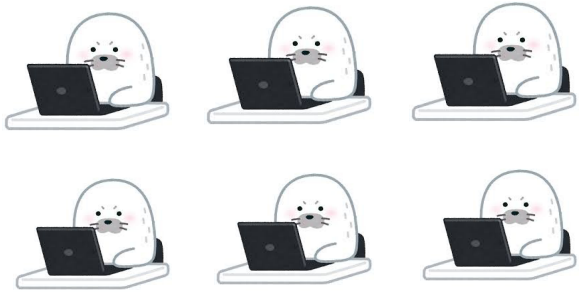
依頼者

Web サーバ

③ やりとりののち
作業結果提出

④ 承認報告

⑤ 報酬受取



作業者

単なるラベリングではなく
創造的で多様なデータが必要：
大規模に収集したい！

① タスク投稿

サーバへのリンクだけ投稿して
動的にタスクを生成できます

Crowdsourcing
platform

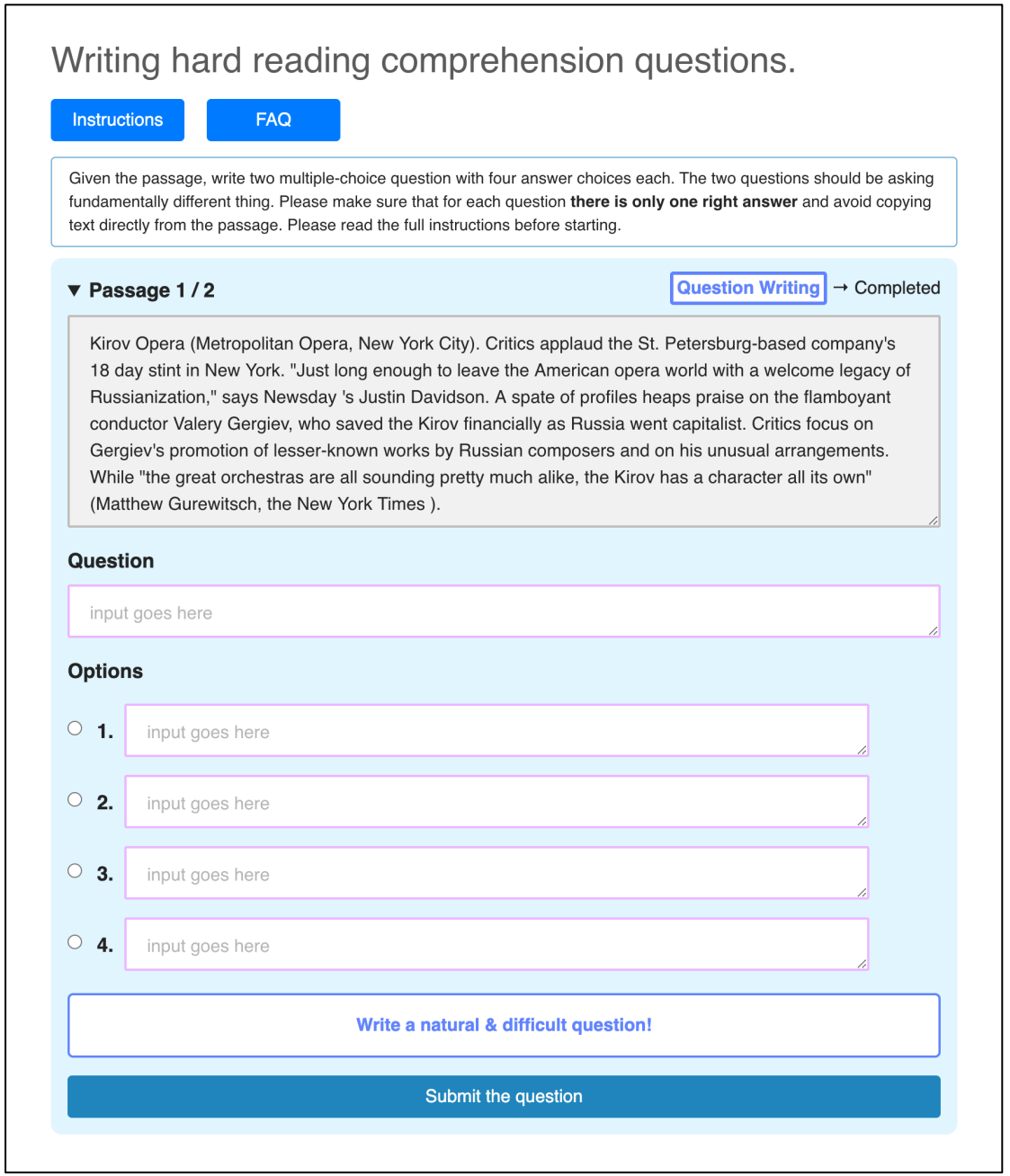
② タスク受注

受注する人に制約をかけることもできる
（これまでの実績や居住地など）

Mturk API: boto3
Web server: FastAPI

読解問題の作問画面

- Instruction には詳細な説明、良い例や悪い例、書き方のヒント、ボーナス情報
- FAQ にはタスクはいつ承認されるか・何回まで受注できるか、など周辺情報
- Submit したあとも最終確定までは戻って編集可能にしている
  - 複数のステップを踏むタスクも作れる
- 空入力や語数、コピペ等の制約はクライアント側に実装
- 自由度の高さから四択問題を収集



クラウドソーシングで大変なこと

楽な作業で高い報酬を得たい作業者

vs.

質の良いデータを安価に集めたい依頼者

- ある程度難しい質問を書いてもらいたい → 例示はできるが「程度」がわかりづらく難しい
- データの多様性のために語数・語彙などの制約を課す → 自然な文章になるとは限らない
- 作業が複雑 & 低報酬 → 誰も参加してくれない
- それなりの長さの文章を読んで質問を書いてもらうのは時間がかかる
  - 時給$10-15を保証しようとすると $2-3/問は必要（依頼者の reputation も見られる）
  - 回答可能性確認のためのアノテーションまで含めるとかなりの額に……（数千問ほどの規模）

最近は fair payment であることも研究の ethics statement で重要

よいデータを集めるためには
クラウドソーシングのノウハウの蓄積が必要！

最近の研究：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、有意に難しい）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
- 疑問
  - 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？
  - どのような文章（ジャンル・読みやすさ etc.）を素材にするのがよいか？
  - 敵対的な収集は有効か？　データが偏らないか？ (cf. Kaushik+ 2021)

どちらも NYU との共同研究です

Nangia & Sugawara+ (ACL 2021)

Sugawara+ (ACL 2022, to appear)

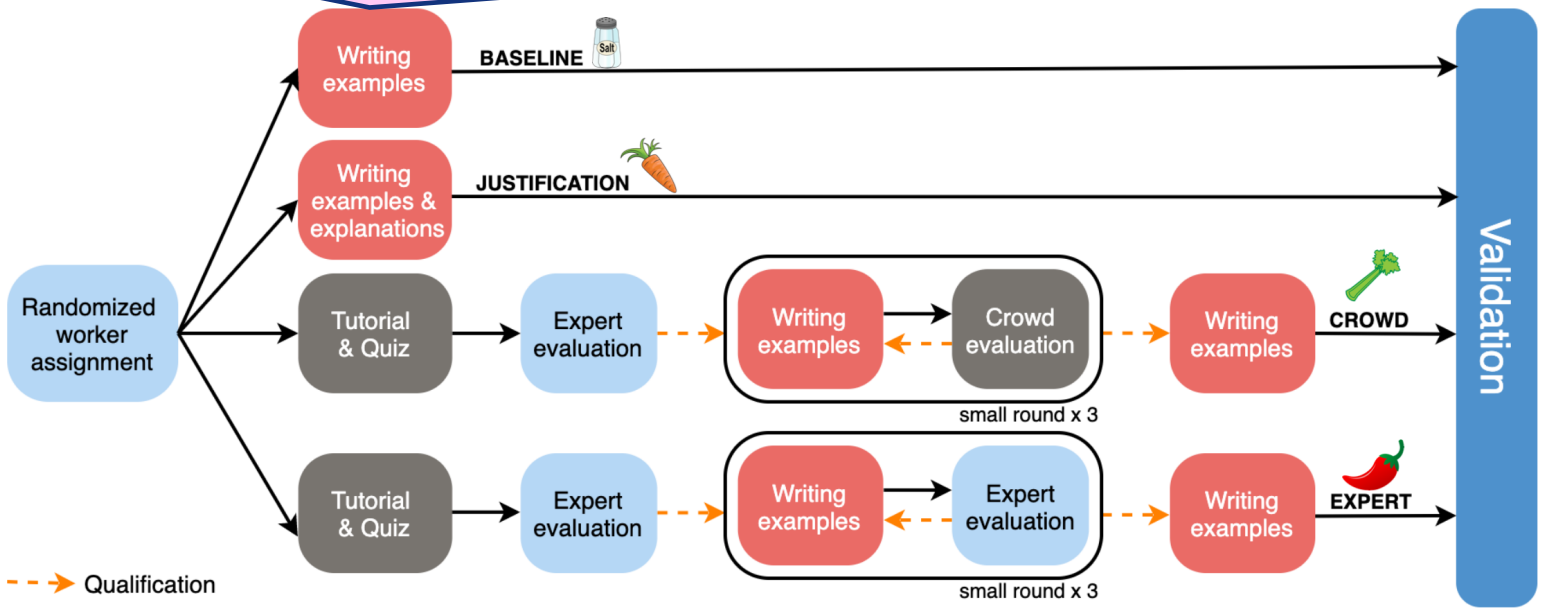
データ収集のプロトコルを比較する

(Nangia & Sugawara+, ACL 2021)

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？

手法：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

指示だけで集める or 正当化の説明を求める or 事前に選抜・フィードバック（by 作業者 or 専門家）



質問の難易度
＝ 人間の精度 ー システムの精度

データ収集のプロトコルを比較する

(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 ➡ 作業者同士 (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

疑問

- 高難易度と言ってもシステム精度 90% vs 80% ほどの差 😢
  - どのような**文章**なら難しくなる？（この研究ではドメインひとつだけ）
  - **敵対的**にモデルが解けないものを集めるとどうなる？

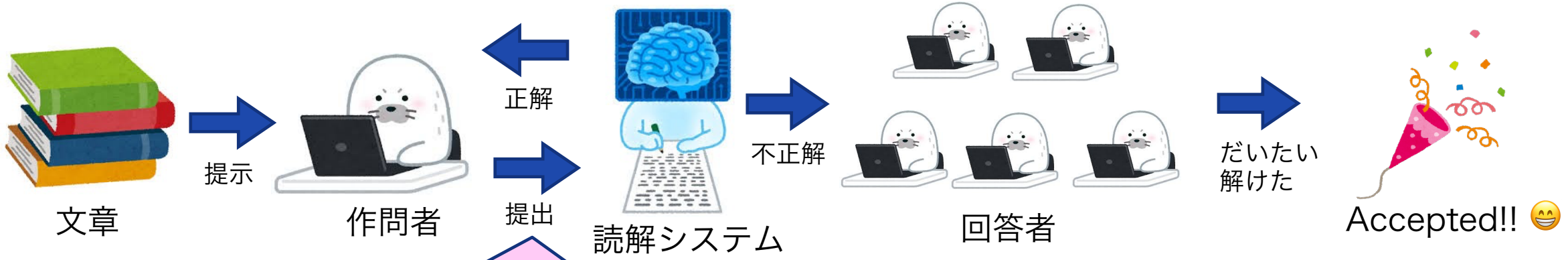
集まる質問は文章に応じてどう変わるか？

(Sugawara+ ACL 2022)

どういった文章なら高難易度かつ多様な質問を集めることができるか？

手法：異なる読みやすさ・ドメインの文章について質問を収集して比較分析

- 物語文（子供/大人向）、専門的な記事、Wikipedia（理/人文系）、試験問題（中高/GMAT）
- 作業者の最初のフィルタリングが重要かつ効率的だったため、質問応答と執筆の二段階で選抜
- 提出をそのまま受け取る / モデルが解けないものだけ受け取る（敵対的） の2通り試す



集まる質問は文章に応じてどう変わるか？

(Sugawara+ ACL 2022)

観察

- 異なる読みやすさや長さの文章 ➡ 質問の難易度に変化なし ❌
- 専門的な文章では論理的推論が必要な難しい質問が相対的に多く集まったが数は少ない ✅❌
- 多様な性質の質問を集めたいのであれば、**多様なドメインの文章**を使ったほうがよさそう
- 敵対的に質問を書いてもらう ➡ 四則演算など tricky な質問が多くなってしまう ❌
- 敵対的収集は難しい問題の効率的な収集に**便利**かもしれないが、分布が歪まないように**工夫が必要**

(a) All questions (b) Standard collection (c) Adversarial collection

今後の展開と課題：何をどう測るか？ (Sugawara+ EACL 2021)

- **「何を測るか」**の良さをどう説明すべき？
  - そもそも人間の言語理解はどのように理論化されているのか？
  - その理論のなかで NLP のタスクはどのように位置づけられるか？

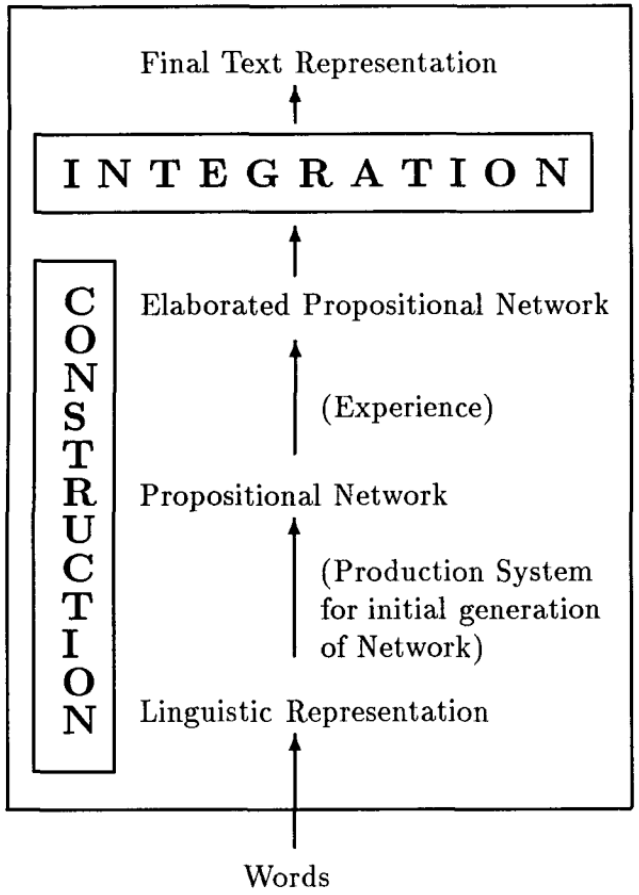
- **「どう測るか」**の良さをどう説明すべき？
  - タスクにおける測定の妥当性をどのように高めていくか？
  - 事例ごとの良さを項目応答理論で測ったりできないか？

何を測るか：心理学における Text Comprehension

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

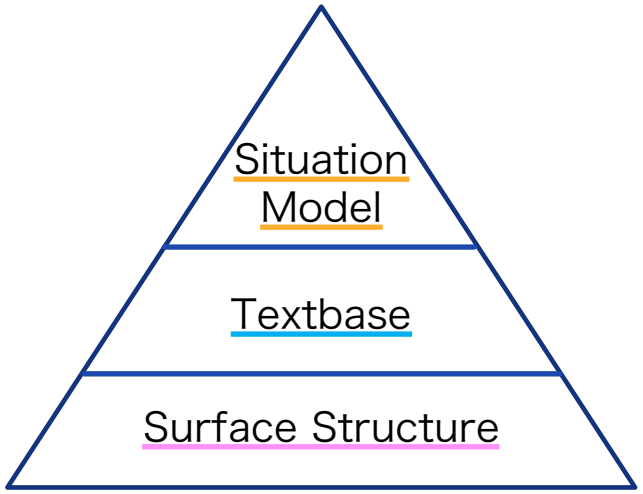
1. Construction
   - 与えられたテキスト情報について、命題のネットワークを構築する。およそ隣接している文同士が結びついている状態
2. Integration
   - 命題集合を用いて一貫した心的表象（状況モデル）を構築する。命題はグローバルに組織化されており、必要に応じて他のモーダルに grounding されている

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)

表現のレベルと言語処理の基礎タスク



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 -> Recent trend?
   - Global structure of propositions
   - Skills: creating a coherent representation and grounding it to other media (sound, image, ...)

将来的な方向性：状況モデルの視点

状況モデルの評価（例）

- 文脈に依存的な状況設定
  - Representations are constructed distinctively depending on the given context
  - Defeasibility: if-then reasoning (Sap+ 2019), abductive reasoning (Bhagavatula+ 2020)
  - 共通基盤（エージェント同士の協同的状況理解）：OneCommon (Udagawa & Aizawa 2021)
  - Novelty: StrategyQA (Geva+ 2021), SituatedQA (Zhgang and Choi 2021)

- 非テキスト情報への grounding
  - 画像： MaRVL (Liu+ 2021), Visual MRC (Tanaka+ 2021), Visual Commonsense Reasoning (Zellers+ 2019), Science textbooks (Kembhavi+ 2017), FigureQA (Kahou+ 2018)
  - 構造的データ： HybridQA (tabular) (Chen+ 2020), Knowledge Base (many...)

どう測るか： psychometrics における構成概念妥当性

構成概念妥当性 (Messick 1995)

- 心理学実験から得られた結果は測りたいものが測れたと解釈するに妥当か

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

例: vocabulary

これらを備えた評価基準 (rubric) を作りましょう

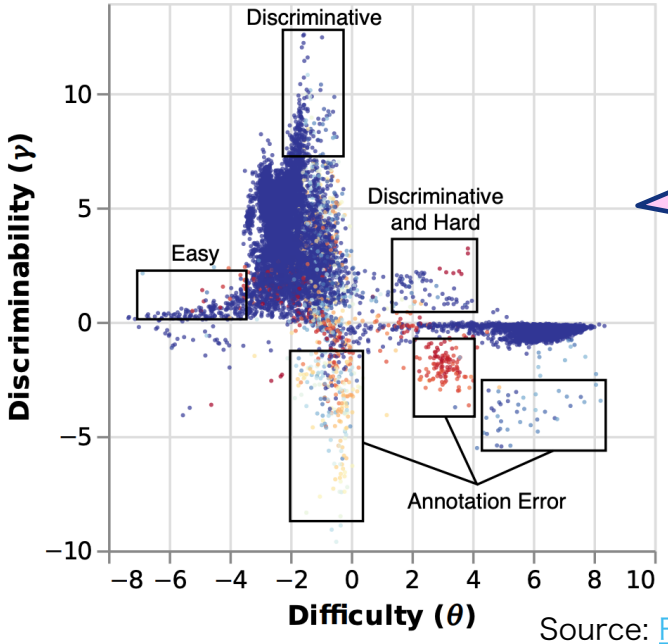
構成概念妥当性の6つの側面 （かなり単純化しています）

1. 内容 (content)
   - 評価すべき対象を評価しているか
2. 本質 (substantive)
   - 過程なども含めて評価できているか
3. 構造 (structural)
   - 評価指標同士が適切に構造化されているか
4. 一般化可能性 (generalizability)
   - 評価指標が信頼できるか
5. 外的指標との関わり (external)
   - 外的な情報と適切に一貫しているか
6. 結果 (consequential)
   - 評価や解釈の誤りにリスクはないか

事例ごとの良さを測れないか？

- 項目応答理論（item response theory）というものがあります：
  - 問題の難易度・識別力をモデリングし、受験者の能力に合った難易度かつ識別力の高い問題を特定してテストを作れるようにする
- 言語理解系のデータセットで IRT を使ってみる研究が徐々に出てきた
  - Lalor+ (2016), Lalor+ (2018), Rodriguez+ (2021) など

能力のモデリングをどう細かくするかなど、multidimensional IRT の方向性はまだこれから



Source: Rodriguez+ (2021)

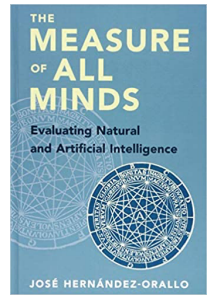
Discriminative かつ Hard な事例に注目すればよい？

項目応答理論を用いることでベンチマーク的に良い質問を特定することはできるが、入力情報中の何が識別性・難易度に寄与したかがわからなければ説明性は低いまま……（人間と同じように解いているとは限らない）
→ 解釈手法と組み合わせる必要がある？

まとめ

「理解」そのものについての哲学的な議論に興味がある方は
最近だと信原 (2020)「強いAI」[link] が導入におすすめです

- **背景**
  - 機械の自然言語理解を評価することは人間の言語理解の探究や応用技術の発展の基盤として重要
- **課題**
  - 従来の言語理解ベンチマークは「これが解けると何ができるようになったと言えるか」が不明瞭
  - 何を評価するかの理論的基礎づけや、データ作りの方法論、妥当性の向上のため何をすべきか難しい
- **現在の研究**
  - 細かな評価指標を備え、意図した解法について品質の保証がなされたベンチマークの構築
  - クラウドソーシングで大規模に効率的にデータを収集する手法の開発
- **展望**
  - 何を測るか：状況的な文脈をより広く・テキストだけでなくマルチモーダルに？
  - どう測るか：心理学的な「測定」の知見と NLP の解釈手法の合流など



[amazon]

[amazon]