発表なしでごめんなさい (´；ω；｀)
短くまとめました

What’s The Meaning Of Superhuman Performance In Today’s NLU?

Simone Tedeschi et al. (ACL 2023)

読み手：菅原 朔 （国立情報学研究所）

最先端 NLP 勉強会 2023

まとめ

Q （訓練済みの）言語モデルが様々なベンチマークで人間超えの精度を出しているけど、本当に人間を超えたと言えるの？

A そんなことはないのできっちり評価しましょう：既存の取り組みサーベイとデータセットの小さな分析、最後に recommendation をまとめています

§2. Popular Leaderboards are Saturated

- 有名なベンチマークは飽和しつつある
- ただしRACE （黄緑、選択式の文章題）でシステム性能が human baseline estimate よりかなり上回っていると主張するのは結構 misleading だと思う
  - Turkers: 73.3%
  - Ceiling (authors): 94.5%
  - Best system: 90% 前後
  - 一般に turkers (amazon mechanical turk の crowdworkers) は真面目に作業してくれず、 RACE の論文はその点の品質保証をあまりしていない

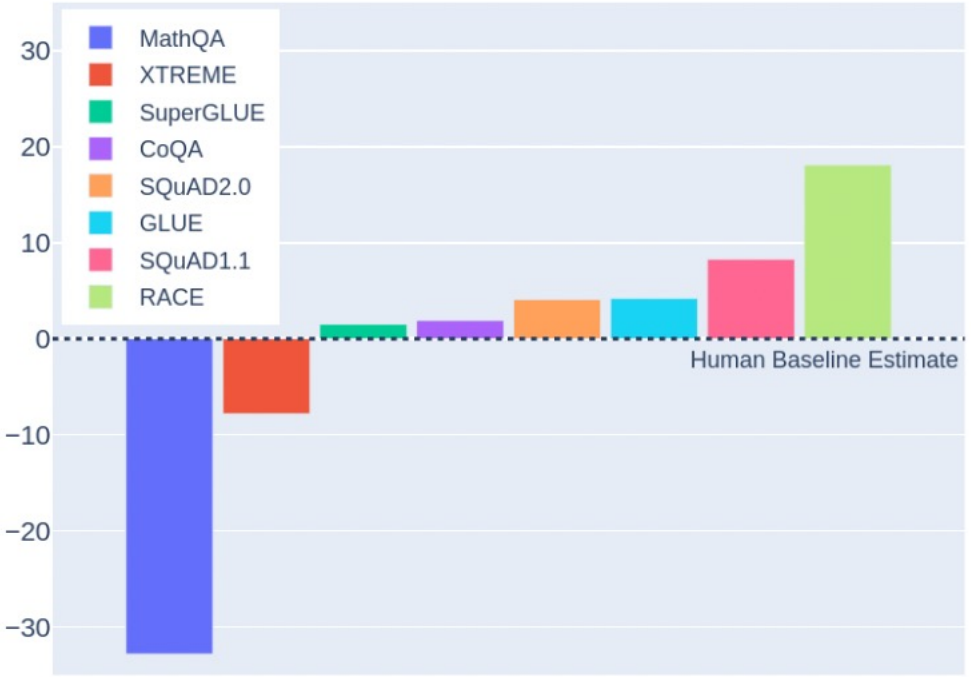

Figure 1: Difference between the scores obtained by the best-performing systems and humans in various popular NLP benchmarks. The systems outperform humans on 6 out of 8 of the reported benchmarks (best seen in color).

§3. Human Baselines are not Reliable

- SuperGLUE では多くのタスクにおいて人間の精度を出すときにテストデータの一部しか使っていない
- 人間のアノテーションが複数あるとき、最終的に何をラベルとするかが統一されていない。多数決をとるか平均をとるか、誰を正答にするか等
  - Validation と response の区別をつけるのが難しかったりする（個人的経験）
- アノテータがどれくらい真面目かわからない：報酬が低かったりすればクオリティが下がるのは当たり前
- アノテーションのガイドラインやアノテータの選定方法が明らかにされていない

§4. Setups Favor Misleading Comparison

- Train-test split が固定的だったり、システムが掴めるようなバイアスが含まれていたりすると、システムの性能を過大評価しうる
- データセットの性質に多様性がないと、一般性のない結果が出る
- 自動評価指標で測れるものには限界がある
- 人間のアノテーションの（ノイズではない）揺れを考慮した評価ができていない

§5. Humans can Explain their Answers

- 人間は回答の根拠を説明できるのだから、システムにもそれを求めては？
- もちろん評価は難しいけど……

§6. Recommendations

1. 機械を人間よりもひいきしないようにしよう
   - Train/test で素材になる文章を変える、問題の難易度をバランスする、テストセットを定期的に新しくする、人間のラベルも効果的に使う、人間のラベルの品質を上げる
2. 人間の性能評価の透明性・再現性を上げよう
   - アノテータ・アノテーションの詳細を記述する、個別のラベルを公開する
3. アノテーションの説明性を上げよう
   - アノテータにラベル付けの根拠を説明させ、システムにも説明を出力させよう

動機が近いので自分の論文の紹介（すみません）

言語理解タスクの結果の解釈の妥当性を高めるためにはこういうことに気をつけるとよさそうです、というのを心理測定学における validity argument という枠組みを参考にしながらリスト化しました (Sugawara & Tsugita, ACL 2023 Findings)

昨年どこかで話したスライドの4.2節あたりも参考になるかも……