Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

John P. Lalor, Pedro Rodriguez, João Sedoc, Jose Hernandez-Orallo

https://eacl2024irt.github.io/

Conclusion, Recent Work, and Future Directions

Concluding Remarks and Summary

1. Learned about IRT models
2. How to implement IRT models and/or use py-irt
3. Showed ways to apply IRT to specific NLP problems
    3.1 Annotation Error
    3.2 Evaluation
    3.3 Training
4. Classical IRT is a starting point, but the range of IRT methods is much larger

Future Directions

1. Classical IRT is a starting point, but the range of IRT methods is much larger
2. Future Directions
    2.1 LLMs?
    2.2 Multidimensional IRT and Big Benchmarks?
    2.3 Predictability?

Google Group

Join our IRT in NLP Google group! TODO-link

Recent Work

Do great minds think alike? Investigating Human-AI Complementarity for Question Answering

- Skill/difficulty should be multidimensional, but making it work is difficult (Rodriguez et al., 2022)
- Idea: use BERT-informed embeddings to inform multidim difficulty, etc.
- Compare different proficiencies of humans versus models
- Gor et al. (2024) made it work!

Do great minds think alike?
Investigating Human-AI Complementarity for Question Answering

Maharshi Gor[1] **Tianyi Zhou**[1] **Hal Daumé III**[1,2] **Jordan Boyd-Graber**[1]

[1]University of Maryland [2]Microsoft Research
mgor@cs.umd.edu

Abstract

This study examines question-answering (QA) abilities across human and AI agents. Our framework CAIMIRA addresses limitations in traditional item response theory, by incorporating multidimensional analysis, identifiability, and content awareness, enabling nuanced comparison of QA agents. Analyzing responses from ~ 30 AI systems and 155 humans over thousands of questions, we identify distinct knowledge domains and reasoning skills where these agents demonstrate differential proficiencies. Humans outperform AI systems in scientific reasoning and understanding nuanced language, while large-scale LLMs like GPT-4 and LLAMA-2-70B excel in retrieving specific factual information. The study identifies key areas for future QA tasks and model development, emphasizing the importance of semantic understanding and scientific reasoning in creating more effective and discriminating benchmarks.

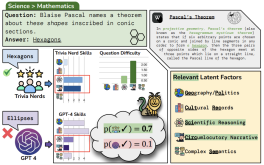

Figure 1: Response Correctness prediction using Agent skills and Question difficulty over relevant latent factors. We list the five latent factors that CAIMIRA discovers, and highlight the relevant ones (green), which contribute to estimating whether an agent will respond to the example question correctly. The agent skills over these relevant factors are highlighted in red boxes.

tion answering, particularly with the new panoply

Related Work

1. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information (Ethayarajh et al., 2022)
2. IRT in Recommender System Benchmarking (Liu et al., 2023)

References

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Maharshi Gor, Tianyi Zhou, III Daumé, Hal, and Jordan Boyd-Graber. 2024. Do great minds think alike? investigating human-ai complementarity for question answering.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2023. What we evaluate when we evaluate recommender systems: Understanding recommender systems' performance using item response theory. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 658–670, New York, NY, USA. Association for Computing Machinery.

Pedro Rodriguez, Phu Mon Htut, John Lalor, and João Sedoc. 2022. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Acknowledgements

Headshots of people we want to thank: Jordan, Hong, Phu, etc.