Item Response Theory for Natural Language Processing

John P. Lalor,[1] **Pedro Rodriguez,**[2] **João Sedoc,**[3,4] **Jose Hernandez-Orallo**[5]
[1] IT, Analytics, and Operations, University of Notre Dame
[2] Meta FAIR, Seattle
[3] Technology, Operations and Statistics, New York University
[4] Center for Data Science, New York University
[5] Universitat Politècnica de València
john.lalor@nd.edu, me@pedro.ai, jsedoc@stern.nyu.edu, jorallo@upv.es

1 Description

This tutorial will introduce the NLP community to Item Response Theory (IRT; Baker, 2001). IRT is a method from the field of psychometrics for model and dataset assessment. IRT has been used for decades to build test sets for human subjects and estimate latent characteristics of dataset examples. Recently, there has been an uptick in work applying IRT to tasks in NLP. It is our goal to introduce the wider NLP community to IRT and show its benefits for a number of NLP tasks. From this tutorial, we hope to encourage wider adoption of IRT among NLP researchers.

As NLP models improve in performance and increase in complexity, new methods for evaluation are needed to appropriately evaluate performance improvements. In addition, data quality continues to be important. Models exploitation of annotation artifacts, annotation errors, and a misalignment between models and dataset difficulty can hinder an appropriate assessment of model performance. As models reach and exceed human performance on certain tasks, it gets more difficult to distinguish between improvements and innovations and changes in scores due to chance. In this **three-hour, introductory** tutorial, we will review the current state of evaluation in NLP, then introduce IRT as a tool for NLP researchers to use when evaluating their data and models. We will also introduce and demonstrate the py-irt Python package for IRT model-fitting to help encourage adoption and facilitate IRT use.

We believe that this should be a tutorial instead of a specialized workshop since the tutorial will aid in exposing a larger NLP audience to IRT. While this methodology has been applied successfully to NLP applications, further community exposure specifically for graduate students may provide a new methodological perspective. We aim to make the tutorial interactive with hands-on Jupyter notebooks which will give concrete simple examples. Tutorial materials are available online.[1]

2 Target Audience/Prerequisites

The tutorial content will be self-contained so that a broad target audience of *CL conference attendees (researchers, PhD students, industry professionals, etc.) can take away information on incorporating IRT in their workflow. In terms of prerequisites, we expect the audience to have basic knowledge of probability and statistics. We also expect audience members to have experience with Python is useful for py-irt.

3 Outline

1. Evaluation in NLP (30 minutes)
2. Introduction to IRT (1 hour)
    - Defining IRT Models
    - IRT Model Fitting
    - Introduction to py-irt
        - This section will include tutorial content and live demonstration of the py-irt package.
3. IRT in NLP (45 minutes)
    - Building Test Sets
        - Model Evaluation
        - Chatbot Evaluation
    - Training Dynamics
        - Example Mining
        - Curriculum Learning
    - Model and Data Evaluation
        - Rethinking Leaderboards
        - Features Related to Difficulty
4. Advanced Topics and Opportunities for Future Work (45 minutes)

[1] https://eacl2024irt.github.io/

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Tutorial Abstracts, pages 9–13

3.1 Evaluation in NLP

Today more than ever evaluation of generative AI and datasets has become more important than ever. We will start with a brief introduction to evaluation in NLP, covering the state of the field over the years (Church and Hestness, 2019). We will cover traditional classification metrics, the rise of leaderboards (Ethayarajh and Jurafsky, 2020), and issues with incremental improvement on summary statistics (Blum and Hardt, 2015).

3.2 Introduction to IRT

We will then move to an introduction of IRT (Baker, 2001; Carlson and von Davier, 2013). IRT is a psychometric method for estimating latent characteristics of test takers and test examples (typically called "items"). IRT has a rich history in the psychometric literature, and is used to construct tests of subject competency (Carlson and von Davier, 2013), mental health screeners (Cole et al., 2011), and health literacy tests (Lalor et al., 2018a), among others.

As IRT is most likely new to the NLP audience, we will spend time discussing the motivation for IRT and the mathematical foundations which make the building blocks of IRT models. We will introduce IRT, highlight some of the important use cases from the literature, and introduce the relevant IRT models.

Specifically, we will introduce models that are used when there is a known correct answer, e.g., an NLP classification task. Such models take a binarized data input and estimate the latent ability ("skill") of the subject and the latent parameters (such as difficulty) of the dataset items.

We will describe how these models are fit, and highlight issues with traditional methods when considering NLP datasets. Traditionally, sampling methods have been use to fit IRT models, but they are computationally expensive on today's large-scale datasets (Wu et al., 2020). We will then introduce variational-inference methods (VI) for IRT model fitting and show how they can alleviate some of the prior concerns (Natesan et al., 2016; Lalor et al., 2019; Wu et al., 2020).

Lastly, we will introduce the py-irt package for fitting IRT models in Python (Lalor and Rodriguez, 2022) and demonstrate how the tool is used using Jupyter notebooks. While IRT has shown promise in NLP, existing software for fitting models are limited by human-data sized constraints. The py-irt package leverages variational-inference (VI) methods to fit IRT models fast and with large data sets. This section of the tutorial will cover the methods built into py-irt and also include a demo with Jupyter notebooks of using py-irt for different NLP evaluation tasks.

3.3 IRT for NLP

We will next discuss how IRT can and has been incorporated into NLP. Prior work has looked at building new test sets with IRT, conducting human-machine comparisons, reevaluating leaderboards, and evaluating chatbot outputs, among other tasks.

3.3.1 IRT for NLP: Dataset Construction and Evaluation

We will first look at IRT for NLP dataset construction and analysis (Lalor et al., 2016; Martínez-Plumed et al., 2019; Sedoc and Ungar, 2020). Specifically, how can one use IRT to build a test set with a variety of examples included that can measure a range of model ability. We will show how IRT can complement traditional evaluation metrics while also revealing new information about both models and test data (Vania et al., 2021; Amidei et al., 2020).

3.3.2 IRT for NLP: Training Dynamics

Next, we will show how IRT can be used to improve the model training process. For example, by filtering datasets to exclude outliers (e.g., those examples that are too easy or too hard) or by using IRT to build a curriculum learning pipeline (Lalor and Yu, 2020), model training can be done more effectively and with better results.

3.3.3 IRT for NLP: Model Evaluation

Finally, we will discuss how IRT can help us to reimagine model evaluation (Otani et al., 2016; Sedoc and Ungar, 2020). We will show how incorporating IRT into leaderboards can give us much more information on model performance (Rodriguez et al., 2021). We will also show how targeted model probing using IRT can lead to new insights about model behavior (Lalor et al., 2018b; Laverghetta Jr. et al., 2021). Finally, we will compare IRT to other methods such as Elo-Ranking, TrueSkill, and other methods.

3.3.4 Advanced Topics

Lastly, we will discuss opportunities for further incorporating IRT into NLP research. This section will discuss more advanced IRT models, as well as ways that NLP research can inform IRT.

For example, what characteristics of examples make them more difficult (Rodriguez et al., 2022)? Also, we will cover IRT extensions and variants to parametrize new instances, such as proxies for difficulty (Martínez-Plumed et al., 2022), or using language models to annotate instance demands, the use of the agent characteristic curves (Martinez-Plumed and Hernandez-Orallo, 2018; Hernández-Orallo et al., 2021) and other ways to use IRT in cases where there is no population of systems.

3.4 Content Breadth

Our goal in this tutorial is to introduce the audience to IRT broadly, and the applications of IRT in NLP specifically. To that end, the content we present will be a mix of foundational IRT research and methods from psychometrics, recent work by the presenters, and work from others in the NLP community who have incorporated IRT into their research.

4 Diversity Considerations

The presenters represent a mix of industry and academic researchers. We also span both Europe and the US. The methods described can be applied to a variety of NLP tasks and languages. The tutorial content will be posted online for wide distribution beyond those able to attend the conference.

5 Ethics Statement

IRT methods can provide fine-grained information about dataset examples and models. With regard to datasets, IRT can potentially surface discrepancies in how groups of examples are handled by NLP models. For example, IRT analyses may show that examples collected from a certain demographic group are systematically more difficult than those examples collected from another demographic group.

6 Pedagogy

We hope that this tutorial can serve as a comprehensive introduction to IRT for an NLP audience and that the content can be reused by others who are not able to attend. To that end, the tutorial will include a combination of presentation slides, demos via Jupyter Notebooks, and interactive sessions in Jupyter notebooks. All content for the tutorial will be hosted online and made publicly available for future use and dissemination.

7 Presenters

John P. Lalor is an Assistant Professor of IT, Analytics, and Operations at the University of Notre Dame. His research interests include model evaluation, curriculum learning, fairness, and BioNLP. Prior to Notre Dame, John received his PhD in Computer Science from the University of Massachusetts, Amherst (advised by Hong Yu) in 2020. John has presented a tutorial on Evaluation and Interpretability in Deep Neural Networks to the 2018 American Medical Informatics Association (AMIA) Annual Symposium with Abhyuday Jagannatha and Hong Yu. Website: https://jplalor.github.io/.

Pedro Rodriguez is a researcher at Meta AI – FAIR. His research interests include question answering, information retrieval, and evaluation. Before joining Meta, Rodriguez completed his PhD at the University of Maryland, advised by Jordan Boyd-Graber. He has reviewed for ACL conferences and workshops, area chaired for COLING, was an organizer of the Dynamic Adversarial Data Collection Workshop at NAACL 2022, and an organizer of a question answering challenge at NeurIPS 2017. Website: https://www.pedro.ai/.

João Sedoc is an Assistant Professor in the department of Technology, Operations and Statistics at New York University Stern School of Business. He is also affiliated with the Center for Data Science at New York University and one of the co-PIs of the Machine Learning for Language (ML^2) group. João's research areas are at the intersection of machine learning and natural language processing. His interests include conversational agents, model evaluation, deep learning, crowdsourcing, spectral clustering, and time series analysis. He has organized multiple workshops: Workshop on Insights from Negative Results in NLP (EMNLP 2020-2021, ACL 2022, EACL 2023), the Workshop on Chatbots and Conversational Agent Technologies & Dialogue Breakdown Detection Challenge (DBDC) (IWSDS 2019, 2020, 2021), Workshop on Neural Conversational AI (ICLR 2021), Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (2021-3), Dialog System Technology Challenge Tracks (AAAI 2021, SIGDIAL 2023), GEM workshop (EMNLP 2023), HumEval workshop 2023 (RANNLP 2023) Website: https://www.stern.nyu.edu/faculty/bio/joao-sedoc.

Jose Hernandez-Orallo is Professor at the Universitat Politècnica de València and Senior Research Fellow at the Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK. His academic and research activities have spanned several areas of AI, machine learning, data science and intelligence measurement, with a focus on a more insightful analysis of the capabilities, generality, progress, impact and risks of AI. He has published five books and more than two hundred journal articles and conference papers on these topics. His research in the area of machine intelligence evaluation has been covered by several popular outlets, such as The Economist, New Scientist and Nature. For a couple of decades, he has vindicated a more integrated view of the evaluation of natural and artificial intelligence, a position represented by his book "The Measure of All Minds" (Cambridge University Press, 2017, PROSE Award 2018) and by multiple papers and events, using IRT, extensions and techniques from some other disciplines to evaluate general-purpose AI such as LLMs. He is a member of AAAI, CLAIRE and ELLIS, and a EurAI Fellow. Website: https://josephorallo.webs.upv.es/

8 Estimate Audience Size

We expect between 50 to 150 attendees. This is based on previous experience at *CL tutorials as well as interest from others to learn about IRT methods.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2020. Identifying annotator bias: A new IRT-based method for bias identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4787–4797, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Avrim Blum and Moritz Hardt. 2015. The ladder: A reliable leaderboard for machine learning competitions. PMLR.

James E Carlson and Matthias von Davier. 2013. Item response theory. *ETS Research Report Series*, 2013(2):i–69.

Kenneth Ward Church and Joel Hestness. 2019. A survey of 25 years of evaluation. *Natural Language Engineering*, 25(6):753–767.

David A Cole, Li Cai, Nina C Martin, Robert L Findling, Eric A Youngstrom, Judy Garber, John F Curry, Janet S Hyde, Marilyn J Essex, Bruce E Compas, et al. 2011. Structure and measurement of depression in youths: applying item response theory to clinical data. *Psychological assessment*, 23(4):819.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. Association for Computational Linguistics.

José Hernández-Orallo, Bao Sheng Loe, Lucy Cheke, Fernando Martínez-Plumed, and Seán Ó hÉigeartaigh. 2021. General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific reports*, 11(1):22822.

John P. Lalor and Pedro Rodriguez. 2022. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*.

John P. Lalor, Hao Wu, Li Chen, Kathleen M. Mazor, and Hong Yu. 2018a. ComprehENotes, an Instrument to Assess Patient Reading Comprehension of Electronic Health Record Notes: Development and Validation. *Journal of Medical Internet Research*, 20(4):e139.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018b. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716, Brussels, Belgium. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259, Hong Kong, China. Association for Computational Linguistics.

John P. Lalor and Hong Yu. 2020. Dynamic data selection for curriculum learning via ability estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 545–555, Online. Association for Computational Linguistics.

Antonio Laverghetta Jr., Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato. 2021. Can transformer language models predict psychometric properties? In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 12–25, Online. Association for Computational Linguistics.

Fernando Martínez-Plumed, David Castellano, Carlos Monserrat-Aranda, and José Hernández-Orallo. 2022. When ai difficulty is easy: The explanatory power of predicting irt difficulty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7719–7727.

Fernando Martinez-Plumed and Jose Hernandez-Orallo. 2018. Dual indicators to analyze ai benchmarks: Difficulty, discrimination, ability, and generality. *IEEE Transactions on Games*, 12(2):121–131.

Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271:18–42.

Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. 2016. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422.

Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2016. IRT-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, Austin, Texas. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

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.

João Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1141–1158, Online. Association for Computational Linguistics.

M Wu, R Davis, B Domingue, C Piech, and Noah D Goodman. 2020. Variational item response theory: Fast, accurate, and expressive.