IRT Tutorial
EACL 2024

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

Differences between Examples

Natural language inference (NLI)

Premise	Hypothesis	Label	Difficulty
A little girl eating a sucker	A child eating candy	Entailment	*easy*
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	*hard*
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	*easy*

Sentiment analysis (SA)

Phrase	Label	Difficulty
The stupidest, most insulting movie of 2002's first quarter.	Negative	*easy*
Still, it gets the job done - a sleepy afternoon rental.	Negative	*hard*
An endlessly fascinating, landmark movie that is as bold as anything the cinema has seen in years.	Positive	*easy*
Perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions.	Positive	*hard*

Leaderboards

😊 **Open LLM Leaderboard**

🏆 The 😊 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🏆 Submit a model for automated evaluation on the 😊 GPU cluster on the "Submit" page! The leaderboard's backend runs the great Eleuther AI Language Model Evaluation Harness - read more details in the "About" page!

| 🏆 LLM Benchmark | 📷 Metrics through time | 📋 About | 🧭 Submit here! |

🔍 Search for your model (separate multiple queries with `;`) and press ENTER...

Model types

☑ 🟢 pretrained ☑ ◆ fine-tuned ☑ ⭕ instruction-tuned ☑ ■ RL-tuned ☑ ?

Select columns to show

☑ Average ⃞ ☑ ARC ☑ HellaSwag ☑ MMLU ☑ TruthfulQA ☑ Winogrande

☑ GSM8K ☑ DROP ⃞ Type ⃞ Architecture ⃞ Precision ⃞ Hub License

⃞ #Params (B) ⃞ Hub ❤ ⃞ Available on the hub ⃞ Model sha

Precision

☑ float16 ☑ bfloat16 ☑ 8bit ☑ 4bit ☑ GPTQ ☑ ?

Model sizes (in billions of parameters)

☑ ? ☑ ~1.5 ☑ ~3 ☑ ~7 ☑ ~13 ☑ ~35 ☑ ~60 ☑ 70+

⃞ Show gated/private/deleted models

T	Model	Average ⃞ ▲	ARC ▲	HellaSwag ▲	MMLU ▲	TruthfulQA ▲	Winogrande ▲	GSM8K ▲	DROP ▲
◆	TigerResearch/tigerbot-70b-chat-v2 📄	69.76	87.03	82.83	66	75.4	79.16	46.02	51.9
🟢	bhenrym14/platypus-yi-34b 📄	68.96	68.43	85.21	78.13	54.48	84.06	47.84	64.55
🟢	01-ai/Yi-34B 📄	68.68	64.59	85.69	76.35	56.23	83.03	50.64	64.2
🟢	chargoddard/Yi-34B-Llama 📄	68.4	64.59	85.63	76.31	55.6	82.79	49.51	64.37
⭕	MayaPH/GodziLLa2-70B 📄	67.01	71.42	87.53	69.88	61.54	83.19	43.21	52.31

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Differences in Questions

Differences in Questions

Differences in Questions

Too Hard

Annotation Error

Discriminative Questions

Too Easy

Boyd-Graber J, Börschinger B. What Question Answering can Learn from Trivia Nerds. ACL 2020.
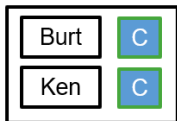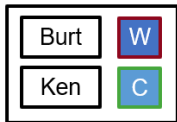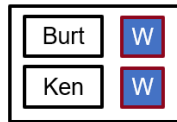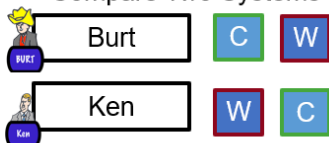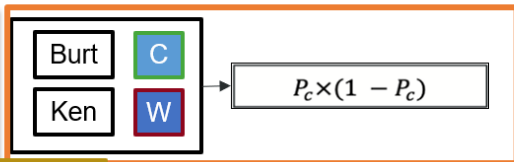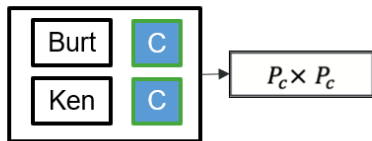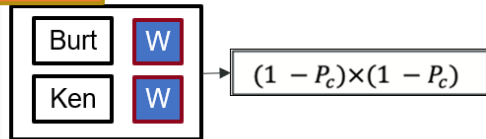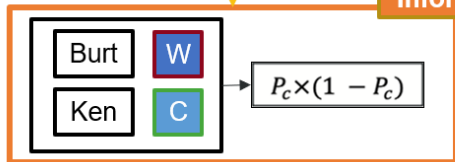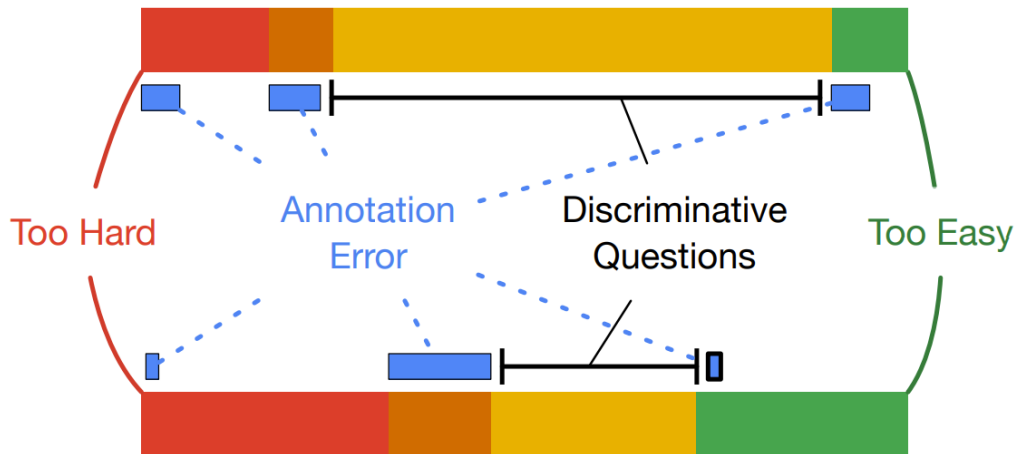
Psychometrics

Psychometrics: study of quantitative measurement practices

- ▶ Building instruments for measurement
- ▶ Development of theoretical approaches to measurement

Item Response Theory (IRT): measure latent traits of test-takers and test questions ("items'')

IRT: 1 Parameter Logistic Model (1PL)

Also known as *Rasch model*

$$p(y_{ij} = 1 | b_i, \theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}}$$

θ_j: latent ability
b_i: difficulty
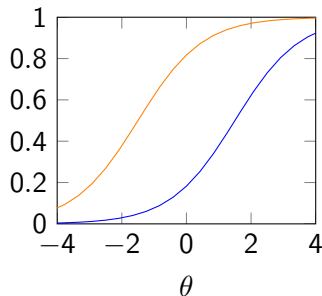
Parameter Estimation

$$p(y_{ij} = 1|b_i, \theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$

$$p(y_{ij} = 0|b_i, \theta_j) = 1 - p(y_{ij} = 1|b_i, \theta_j)$$

$$L = \prod_{j=1}^{J} \prod_{i=1}^{I} p(Y_{ij} = y_{ij}|b_i, \theta_j)$$

$$q(\Theta, B) = \prod_j \pi_j^\theta(\theta_j) \prod_i \pi_i^b(b_i)$$

Natesan, Prathiba, et al. "Bayesian prior choice in IRT estimation using MCMC and variational Bayes.'' Frontiers in psychology 2016

Evaluating DNN Performance with IRT

Premise	Hypothesis	Label	Difficulty
A little girl eating a sucker	A child eating candy	Entailment	-2.74
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	0.51
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	-1.92
Nine men wearing tuxedos sing	Nine women wearing dresses sing	Contradiction	0.08

Phrase	Label	Difficulty
The stupidest, most insulting movie of 2002's first quarter.	Negative	-2.46
Still, it gets the job done - a sleepy afternoon rental.	Negative	1.78
An endlessly fascinating, landmark movie that is as bold as anything the cinema has seen in years.	Positive	-2.27
Perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions.	Positive	2.05

IRT for Leaderboards (SQuAD)

▶ 1.9 million subject-item pairs

IRT for SQuAD
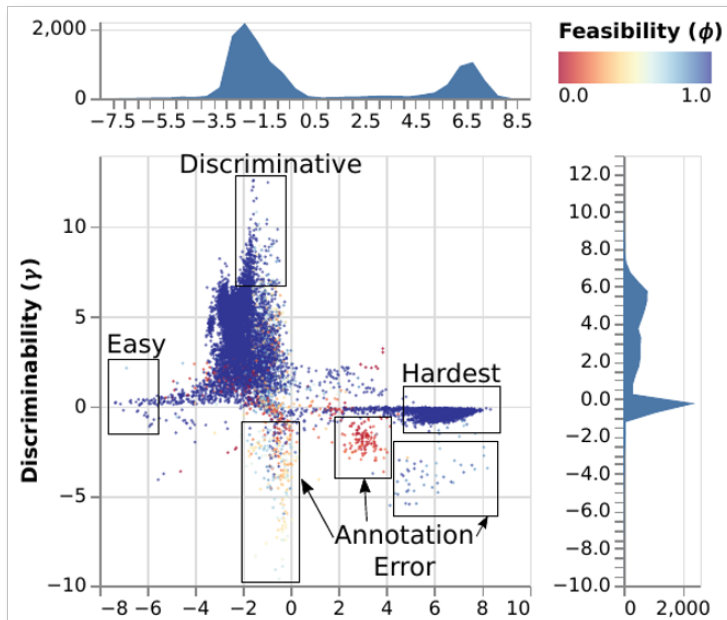
Predicting Correct Responses
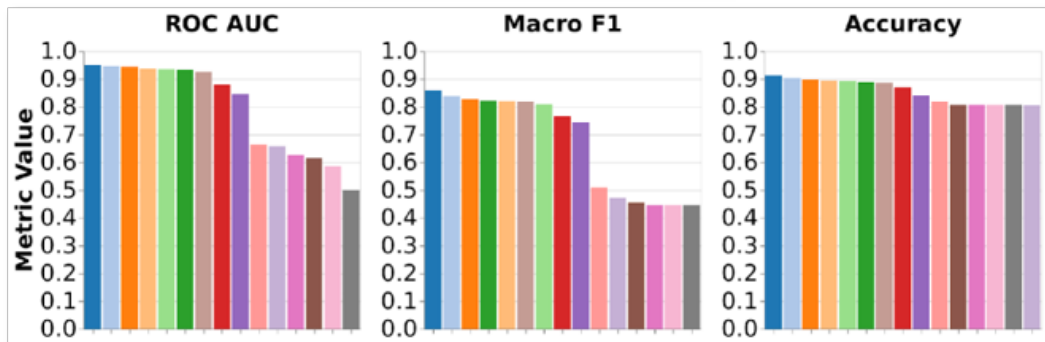
Ranking Performance
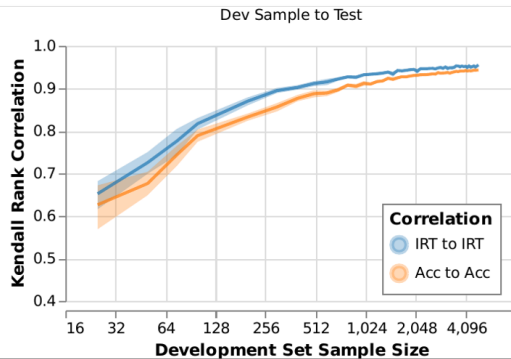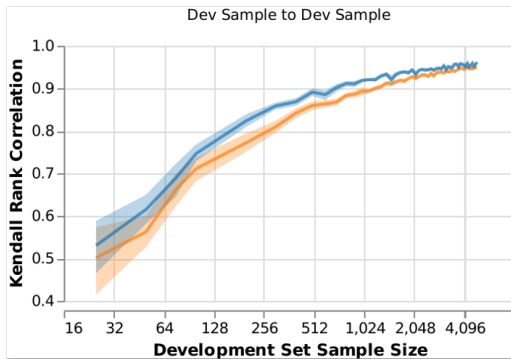
Traditional Curriculum Learning

$$t_e = f(e)$$

- ▶ Example difficulty based on heuristics
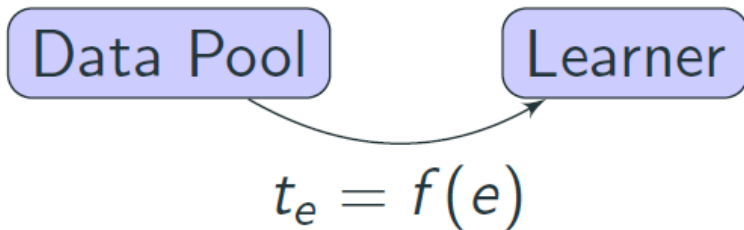 - ▶ Replace heuristic with IRT difficulty
- ▶ Strategy is static
- ▶ Competence-based CL: $t_e = f(e, c_0)$

Platanios, E., et al. "Competence-based Curriculum Learning for Neural Machine Translation.'' NAACL 2019.

Dynamic Data Selection

$$\hat{\theta}_e$$

Data Pool   Learner

$$t_e = f(\hat{\theta}_e)$$
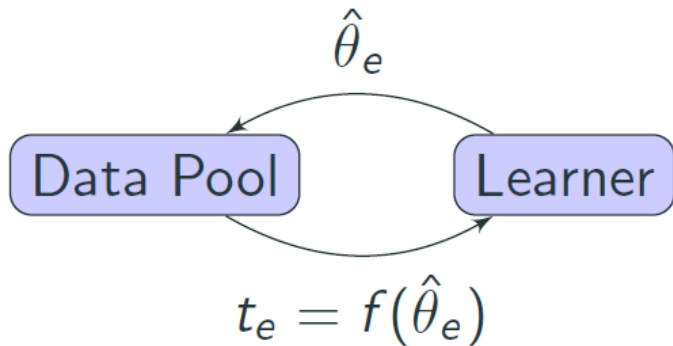
- ▶ Example difficulty is learned
- ▶ Training set *dynamically selected* as a function of model ability

IRT in Python: py-irt

```
{"subject_id": "pedro",    "responses": {"q1": 1, "q2": 0, "q3": 1, "q4": 0}}
{"subject_id": "pinguino", "responses": {"q1": 1, "q2": 1, "q3": 0, "q4": 0}}
{"subject_id": "ken",      "responses": {"q1": 1, "q2": 1, "q3": 1, "q4": 1}}
{"subject_id": "burt",     "responses": {"q1": 0, "q2": 0, "q3": 0, "q4": 0}}
```
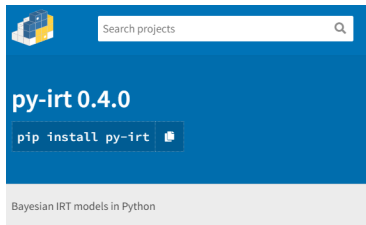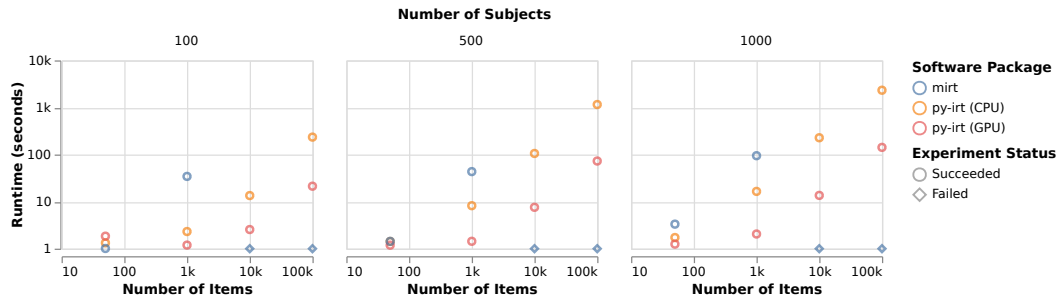
```
py-irt train 1pl data/data.jsonlines output/1pl/
```

```
{                                        "irt_model": "1pl",
  "ability": [                           "item_ids": {
    -1.7251124382019043,                   "0": "q2",
    -0.06789101660251617,                  "1": "q4",
    1.6059941053390503,                    "2": "q1",
    -0.20248053967952728                   "3": "q3"
  ],                                     },
  "diff": [                              "subject_ids": {
    0.008014608174562454,                  "0": "burt",
    9.654741287231445,                     "1": "pinguino",
    -5.5452165603637695,                   "2": "ken",
    -0.2792229950428009                    "3": "pedro"
  ],                                     }
                                       }
```

IRT in Python: py-irt

https://github.com/nd-ball/py-irt