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

IRT for NLP

Overview of IRT Applications:

- Dataset Construction
- Model Training
- Evaluation

Assumptions for IRT + NLP

Basic assumptions of the data and parameterization we have:

- A dataset with items indexed by i.
- A set of subjects indexed by j.
- Responses r_{ij} from graded responses of subjects to each item.
- An IRT parameterization, e.g., one with item difficulty β_i, discriminability γ_i, and skill θ_j might assume:

$$p(r_{ij} = 1 | \beta_i, \theta_j) = \frac{1}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

What IRT Yields

Given the previous information, IRT will yield estimates for chosen parameters, i.e.: item difficulty β_i, discriminability γ_i, and skill θ_j.

Consider two scenarios:

- What if the dataset is the training data?
- What if the dataset is a test set?

IRT from Improving Model Training

TODO: John fill some of this?

Citations to Incorporate

- Rodriguez et al. (2021), (Vania et al., 2021)

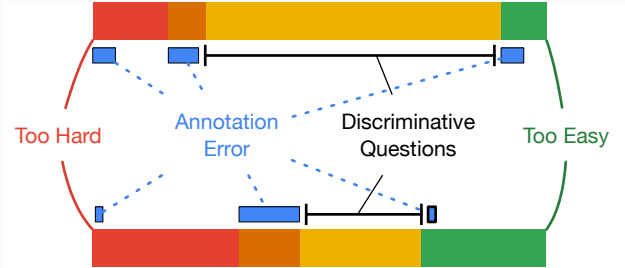
IRT Applications in NLP Evaluations

- Alternate Evaluation Metrics, e.g., Subject skill θ_j (Lalor et al., 2018; Rodriguez et al., 2021)
- Find Bad Evaluation Items (Rodriguez et al., 2021)
- Estimate Longevity of Tasks (Vania et al., 2021)

IRT Applications: Alternate Evaluation Metrics

IRT Applications: Finding Annotation Error

Test examples can be: too hard, discriminative, too easy, or erroneous [1]



How can we use IRT to identify each example type?

[1] Boyd-Graber and Börschinger (2020)

IRT Applications: Finding Annotation Error

- Too Hard $\rightarrow$ Bad examples?:
  - High item difficulty β_i
  - Low discriminability γ_i

IRT Parameters

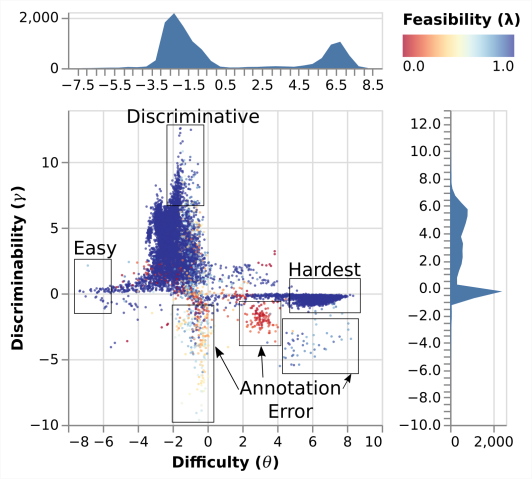
- Item Difficulty: β_i
- Item Discriminability: γ_i
- Subject Skill θ_j

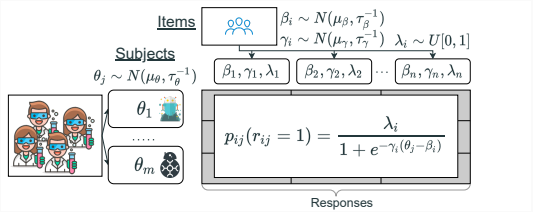
IRT Model

$$p(r_{ij} = 1 | \beta_i, \theta_j) = \frac{1}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

IRT Applications: Finding Annotation Error

In Rodriguez et al. (2021), we used a slightly different model to explicitly model this:





Building IRT Test Sets

TODO: screenshot of older figure, or redo

IRT for NLP: Human Annotations

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 NLP: Human Annotations

Item Set	Theta Score	Percentile	Test Acc.
5GS			
Entailment	-0.133	44.83%	96.5%
Contradiction	1.539	93.82%	87.9%
Neutral	0.423	66.28%	88%
4GS			
Contradiction	1.777	96.25%	78.9%
Neutral	0.441	67%	83%

Human Bottleneck

- Gathering human response patterns is expensive
- Can we use ensembles of models to gather response patterns?
- Even if we can, should we?

Building IRT Models with Artificial Crowds

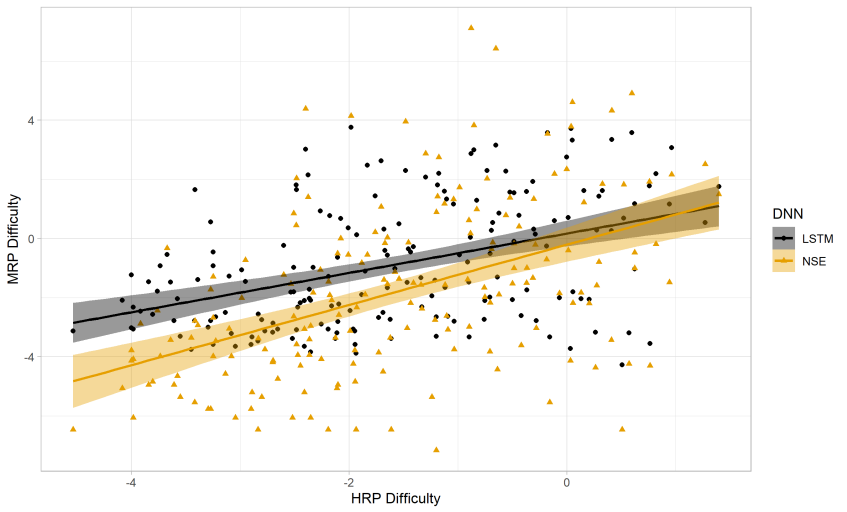
Artificial Crowd Construction

TODO: screenshot of older figure, or redo

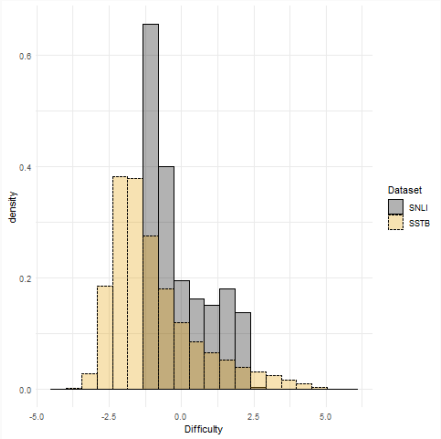
Experiments

- Parameter comparison between models fit with human and machine response patterns
- Downstream use-case: training set filtering
- Qualitative evaluation: how do they look?

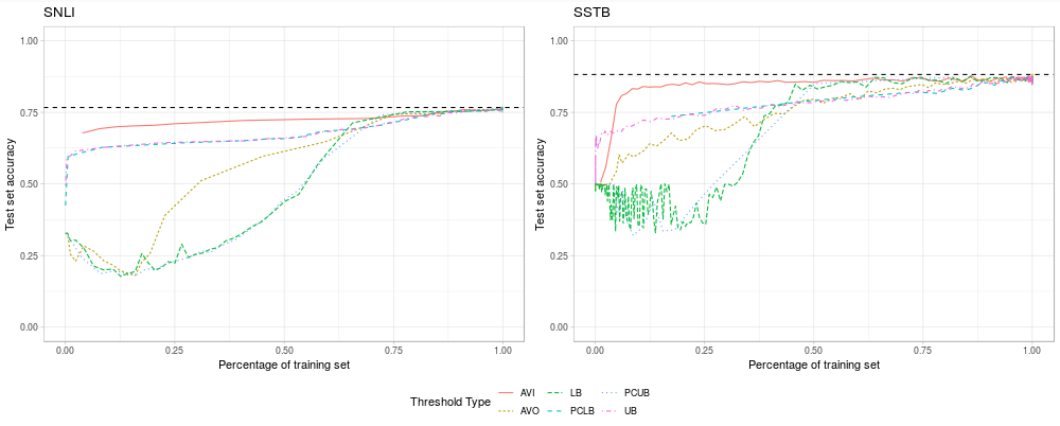
Human-Machine Correlation



Difficulty Distribution



Data set filtering



- AVI: $|b_i| < \tau$
- UB: $b_i < \tau$
- PCUB: $pc_i < \tau$

- AVO: $|b_i| > \tau$
- LB: $b_i > \tau$
- PCLB: $pc_i > \tau$

MT-DNN Results

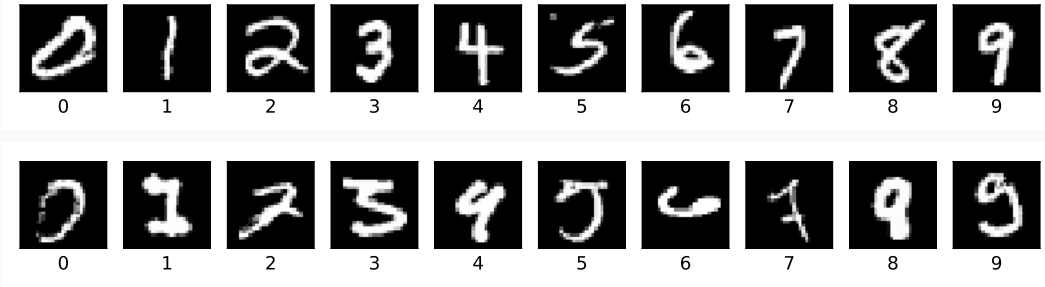
Strategy	% of Training Data		
	0.1%	1%	10%
Random (reported)	82.1	85.2	**88.4**
Random (small batch)	81.79	84.90	88.32
Lower-bound	43.68	41.56	39.89
Upper-bound	81.62	80.46	79.06
AVI	**82.44**	**85.44**	86.73
AVO	43.60	42.05	40.81

Biggest Differences

Task	Label	Item Text	Difficulty ranking		
			Humans	LSTM	NSE
SNLI	Con.	*P:* Two dogs playing in snow. *H:* A cat sleeps on floor	168	1	5
	Ent.	*P:* A girl in a newspaper hat with a bow is unwrapping an item. *H:* The girl is going to find out what is under the wrapping paper.	55	172	176
SSTB	Pos.	Only two words will tell you what you know when deciding to see it: Anthony. Hopkins.	9	103	110
	Neg.	...are of course stultifyingly contrived and too stylized by half. Still, it gets the job done–a sleepy afternoon rental.	128	46	41

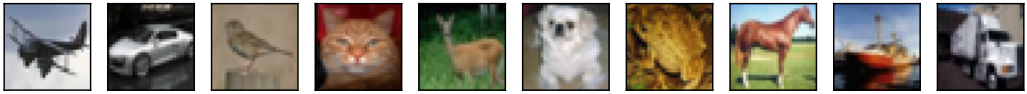
Examples: MNIST

MNIST Test Set



Examples: CIFAR

CIFAR Test Set \



Remarks

- Correlation between parameters between human and machine IRT models
- Downstream effectiveness of difficulty
- Qualitative check of learned parameters
- What about θ?

Dynamic Data Selection for Curriculum Learning via Ability Estimation (DDaCLAE)

Traditional CL

TODO: screenshot of older figure, or redo

- Example difficulty based on heuristics
- Strategy is static

Dynamic Data Selection

TODO: screenshot of older figure, or redo

- Example difficulty is learned
- Training set dynamically selected as a function of model ability

Estimating θ

Gather responses from model j for items with known difficulties

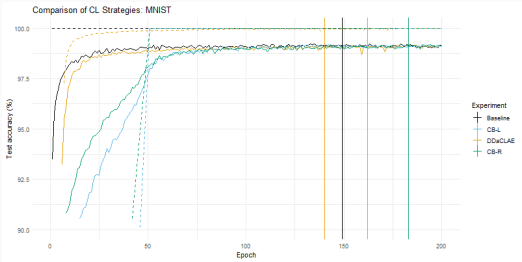
$$
\begin{aligned}
Z_j &= \forall_{y \in Y} \mathbf{I}[y_i = \hat{y}_i] \\
L(\theta_j | Z_j) &= p(Z_j | \theta_j) \\
\hat{\theta}_j &= \arg\max_{\theta_j} \prod_{i=1}^{I} p(z_{ij} = y_{ij} | \theta_j)
\end{aligned}
$$

DDaCLAE

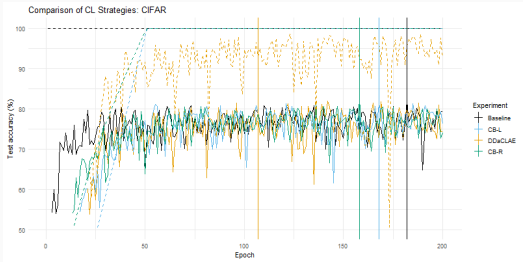
Dynamic Data selection for Curriculum Learning via Ability Estimation

- At each epoch e:
  - Label all data: $\hat{Y}$
  - Estimate $\hat{\theta}_e$: $score(Y, \hat{Y}, B)$
  - Select training data: $b_i \leq \hat{\theta}_e$

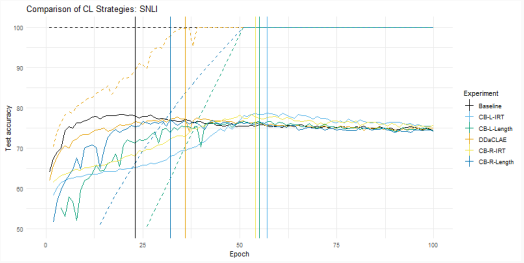
Results



(a) MNIST



(b) CIFAR



Results

Metric	Experiment	MNIST	CIFAR	SSTB	SNLI
%Δ Train Size	Baseline	0	0	0	**0**
	DDaCLAE	**-9.37**	**-53.71**	**-88.68**	33.51
	CB Lin	-8.22	-21.56	-73.17	38.07
	CB Root	11.29	-22.63	10.23	60.08
%Δ Accuracy	Baseline	**0**	0	0	0
	DDaCLAE	-0.17	**0.66**	**0.45**	-1.08
	CB Lin	-0.01	-0.90	-0.18	**0.69**
	CB Root	-0.06	0.13	-0.38	-0.37

Results

Label	Review	Δ_d
Pos	Heart	67342
Pos	The year's greatest adventure, and Jackson's limited but enthusiastic adaptation has made literature literal without killing its soul – a feat any thinking person is bound to appreciate.	67334
Pos	Hip	67332
Neg	Exit	67346
Neg	There's an admirable rigor to Jimmy's relentless anger, and to the script's refusal of a happy ending, but as those monologues stretch on and on, you realize there's no place for this story to go but down.	67330

Results

Label	Premise	Hypothesis	Δ_d
Con.	Two men in a jogging race on a black top street, one man wearing a black top and pants and the other is dressed as a nun with bright red tennis shoes, while onlookers stand in a grassy area and watch from behind a waist high metal railing.	There is no metal railing.	549179
Ent.	Two dogs in the water.	They are swimming	549180
Neut.	Male musicians are playing a gig with one on the drums and the other on the guitar, with a backdrop of purple graphics apart of the light show.	Male musicians with long hair are playing a gig with one on the drums and the other on the guitar, with a backdrop of purple graphics apart of the light show.	549184
Neut.	A dog in a lake.	A dog is swimming.	549183

References

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *EMNLP*.

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 *ACL*.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021