Building Evaluation Scales for NLP using Item Response Theory

John Lalor

CICS, UMass Amherst

Joint work with Hao Wu (BC) and Hong Yu (UMMS)

Motivation

Evaluation metrics for NLP have been mostly unchanged

- Obtain gold standard test set (usually human-annotated labels)
- Compare model results to test set
  - Accuracy, Recall, Precision, F scores, etc.

Implicit assumption is that each example in test set is equally important

However, not all test set examples are created equal.

- Some easy, some hard, many in between

Existing evaluation metrics do not consider characteristics of examples

Our Approach

We propose using Item Response Theory (IRT) from psychometrics to describe characteristics of individual examples (such as difficulty and discriminating power)

IRT accounts for these characteristics in its estimation of ability for an NLP task

- The evaluation considers how easy correctly answered examples are, and how well the system performs with respect to the human population

Outline of Talk

- Item Response Theory
- Recognizing Textual Entailment
- Data Collection and Model Fitting
- Results and Conclusions
- Ongoing and Future Work

Item Response Theory

IRT Introduction

Psychometric methodology for scale construction and evaluation

Jointly models individual ability and item characteristics

IRT jargon

- "item": single example
- "response patterns:" set of responses to all items
- "evaluation scale:" calibrated set of items to be administered
- "ability score" aka θ: Score assigned to an individual based on her responses to the evaluation scale items

Baker, Frank B., and Seock-Ho Kim, eds. Item response theory: Parameter estimation techniques. CRC Press, 2004.

IRT Introduction

Widely used in educational testing: construction, evaluation, and scoring of standardized tests (e.g. TOEFL, GRE)

By fitting an IRT model for an NLP task using human labels, we can score an NLP system according to IRT

- Plus, the IRT score can place an NLP system performance in the context of a human population

IRT Assumptions

- Individuals differ from each other on an unobserved latent trait dimension (“ability”)
- The probability of correctly answering an item is a function of the person’s ability.
- Responses to different items are independent of each other for a given ability level of the person (“local independence assumption”)
- Responses from different individuals are independent of each other

Baker, Frank B., and Seock-Ho Kim, eds. Item response theory: Parameter estimation techniques. CRC Press, 2004.

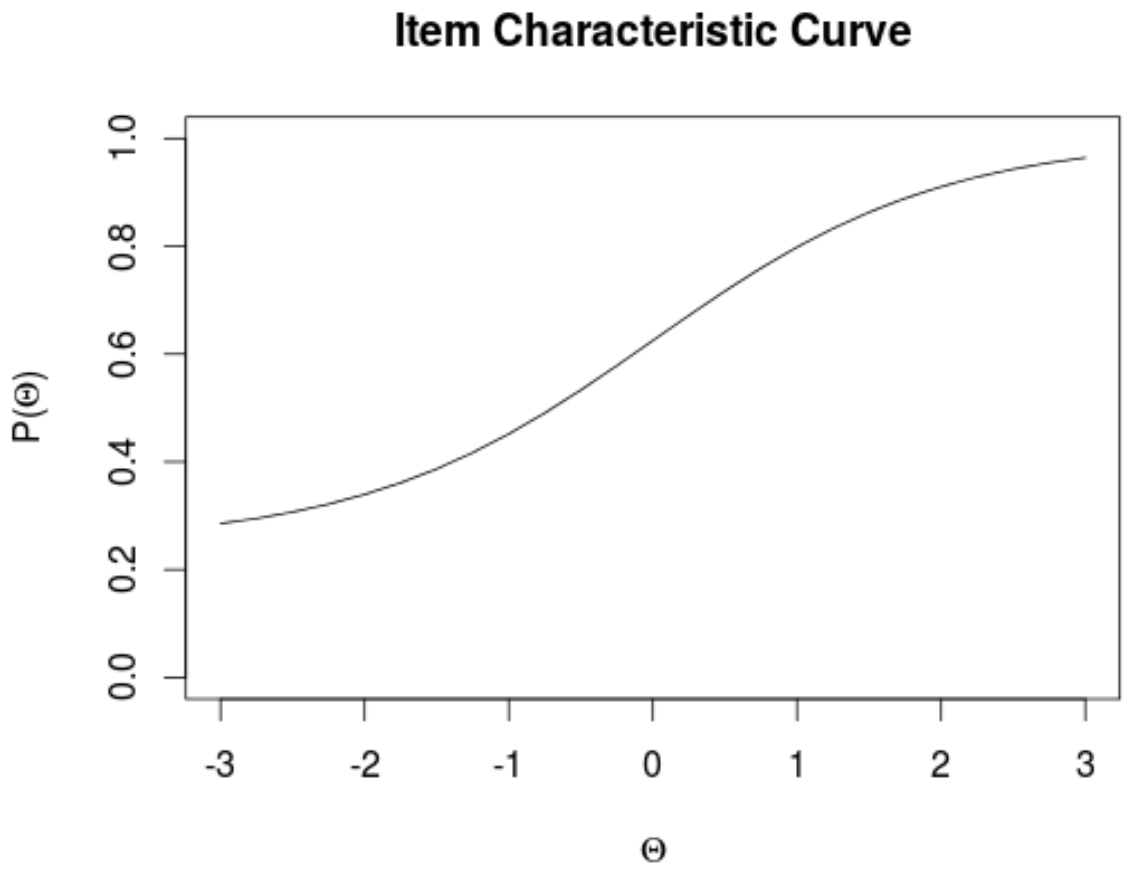
Three Parameter Logistic (3PL) Model

$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

p_{ij}: Probability of individual *j* answering item *i* correctly
θ_j: Ability estimate of individual *j*
a_i: Discrimination parameter for item *i*
b_i: Difficulty parameter for item *i*
c_i: Guessing parameter for item *i*

Baker, Frank B., and Seock-Ho Kim, eds. Item response theory: Parameter estimation techniques. CRC Press, 2004.

Item Characteristic Curve (ICC)



Fitting the Model

Maximize probability of observing current response patterns as a function of the item parameters (Integrate out human ability parameters)

- Given item parameters, estimate individual ability according to normal distribution

Retain/Remove Decision for Individual Items

- Test local independence assumption
- Test the fit of the Item Characteristic Curve for each item

Bock, R. Darrell, and Murray Aitkin. "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm." Psychometrika 46.4 (1981): 443-459.
Chalmers, R. Philip. "mirt: A multidimensional item response theory package for the R environment." Journal of Statistical Software 48.6 (2012): 1-29.

ICC: Varying Difficulty

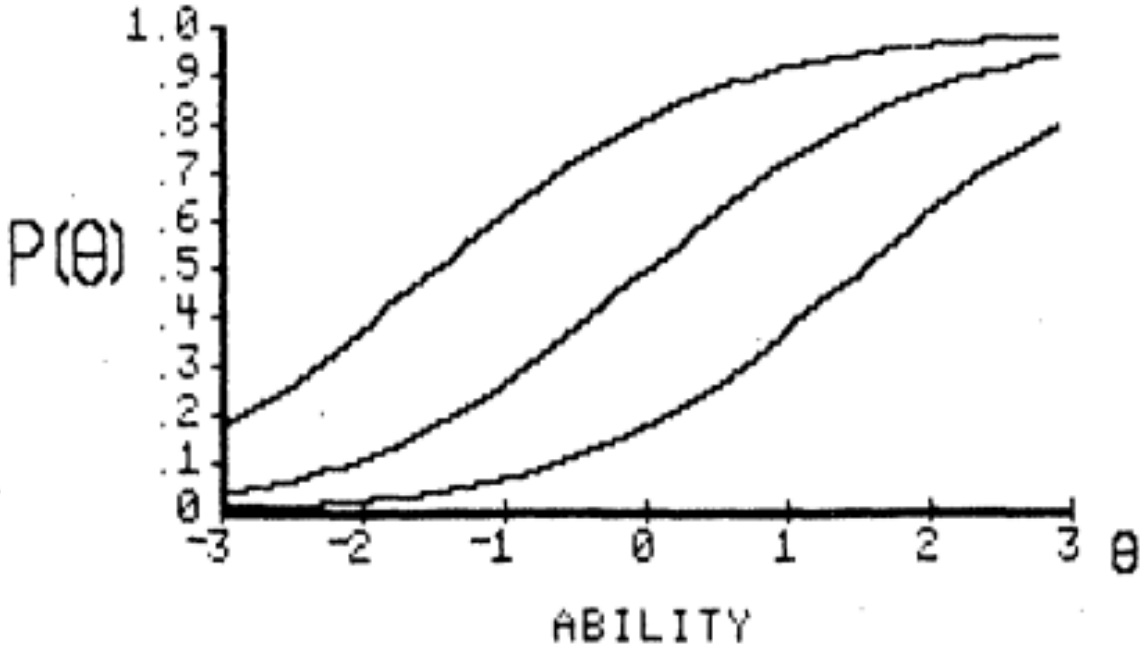


FIGURE 1-2. Three item characteristic curves with the same discrimination but different levels of difficulty

Baker, F. B. The Basics of Item Response Theory. University of Maryland, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation (2001).

ICC: Varying Discrimination

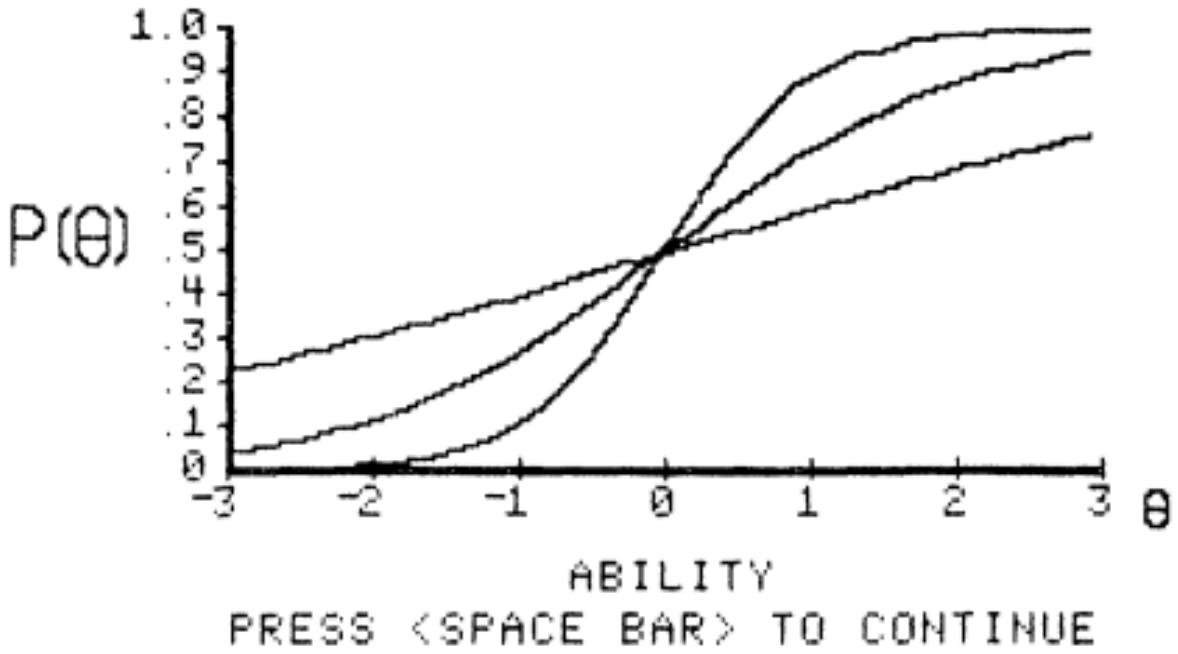


FIGURE 1-3. Three item characteristic curves with the same difficulty but with different levels of discrimination

Baker, F. B. The Basics of Item Response Theory. University of Maryland, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation (2001).

ICC: Perfect Discrimination

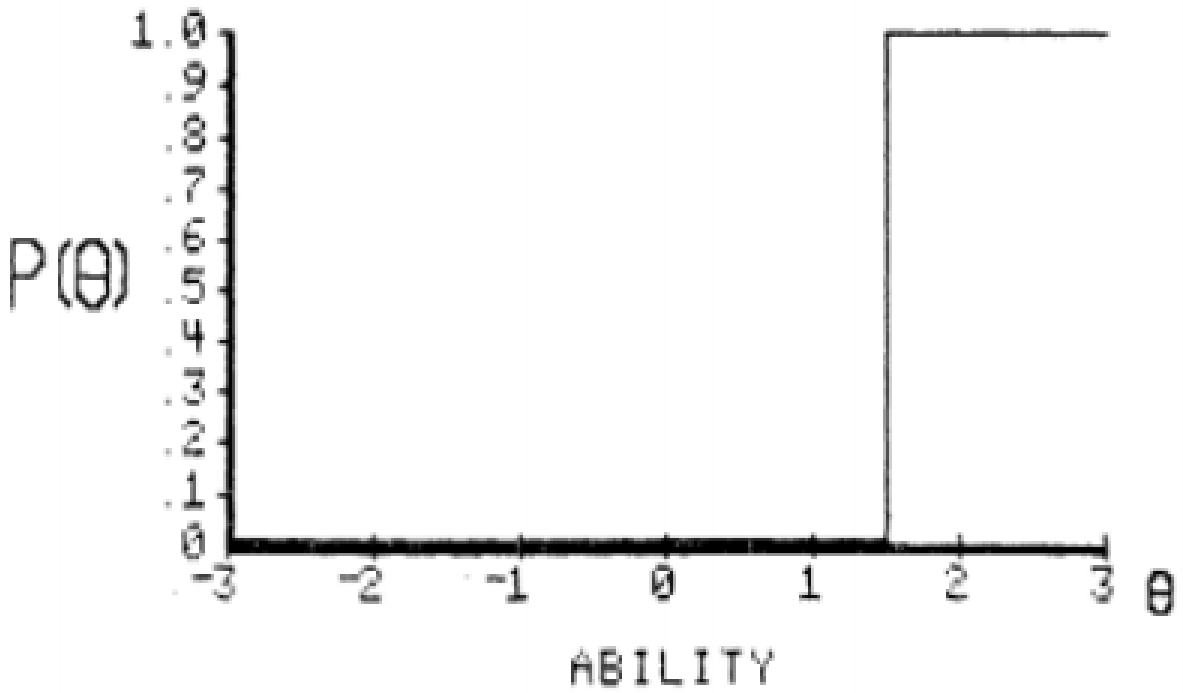


FIGURE 1-4. An item that discriminates perfectly at $\theta = 1.5$

Baker, F. B. The Basics of Item Response Theory. University of Maryland, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation (2001).

Recognizing Textual Entailment

Recognizing Textual Entailment (RTE)

Given premise P and hypothesis H, if P is true what can we infer about H?

- Entailment: P is true implies H is true
- Contradiction: P is true implies H is false
- Neutral: P and H are unrelated

RTE Examples

Entailment

- P: A woman is kneeling on the ground taking a photograph
- H: A picture is being snapped

Contradiction

- P: People were watching the tournament in the stadium
- H: The people are sitting outside on the grass

Neutral

- P: Two girls on a bridge dancing with the city skyline in the background
- H: The girls are sisters

Contradiction

- P: A group of soccer players are grabbing onto each other as they go for the ball
- H: A group of football players are playing a game

RTE Examples

Entailment

- P: A woman is kneeling on the ground taking a photograph
- H: A picture is being snapped

Contradiction

- P: People were watching the tournament in the stadium
- H: The people are sitting outside on the grass

Neutral

- P: Two girls on a bridge dancing with the city skyline in the background
- H: The girls are sisters

Contradiction

- P: A group of soccer players are grabbing onto each other as they go for the ball
- H: A group of football players are playing a game

Stanford Natural Language Inference Corpus (SNLI)

570k human-written sentence pairs

- Much larger than other RTE resources
- All sentences were human-generated

P sentences taken from Flickr30k corpus

H sentences provided by Amazon Mechanical Turk users (Turkers)

- Turkers given P and asked to write 3 H's
  - Definitely true given P, definitely false given P, might be true given P

A subset (10%) was subject to quality control checks

- 4 new annotators labeled the sentence pair + original author = 5 labels

Bowman, Samuel R. et al. "A large annotated corpus for learning natural language inference." Proceedings of EMNLP (2015)
Young, Peter, et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions." Transactions of the Association for Computational Linguistics 2 (2014): 67-78.

Data Collection & Model Fitting

SNLI Subsets

We selected 180 items from the quality control section for additional labeling from Amazon Mechanical Turk (1000 new labels per sentence pair)

90 “5GS” items (where 5/5 annotators agree on gold label)

- 30 Entailment, 30 Contradiction, 30 Neutral

90 “4GS” items (where 4/5 annotators agree)

- Same 30/30/30 split

6 data subsets: {5GS,4GS}, {Entailment, Contradiction, Neutral}

AMT Task

Initial screening for Turkers

- 97% or higher approval rating
- Located in America (as proxy for English-speaking)

Attention-check questions to check quality of responses

Turkers were given either 5GS or 4GS dataset to label (one P-H pair at a time)

Our IRT Models

Fit 1 IRT model per data subset

- Factor analysis of response patterns: 3 latent factors that matched labels
- Target rotation to associate factors and items
- Compare one- and two-factor 3PL models to confirm unidimensional structure

Each model measures ability to recognize that particular label

Prior to fitting models, remove sentence pairs with semantic or syntactic discrepancies (recall soccer - football example)

Our IRT Models

Fit 3PL model in each case

Test significance of c_i for each item

- Not significantly different than 0 → fit 2PL ICC for that item

Iterative Process

- Poor item fit to model ICC → remove
- a_i too low → remove
- Refit model and repeat process until no items are removed

Results

Turker Agreement

	4GS	5GS	Overall
Pairs with majority agreement	95.6%	96.7%	96.1%
Pairs with superma-jority agreement	61.1%	82.2%	71.7%

Table 2: Summary statistics from the AMT HITs.

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Turker Agreement

Fleiss' κ	4GS	5GS	Bowman et al. 2015
Contradiction	0.37	0.59	0.77
Entailment	0.48	0.63	0.72
Neutral	0.41	0.54	0.6
Overall	0.43	0.6	0.7

Table 3: Comparison of Fleiss' κ scores with scores from SNLI quality control sentence pairs.

Lalor, John P. et al. "Building an Evaluation Scale using Item Response Theory." Proceedings of EMNLP (2016)

Model Fitting Statistics

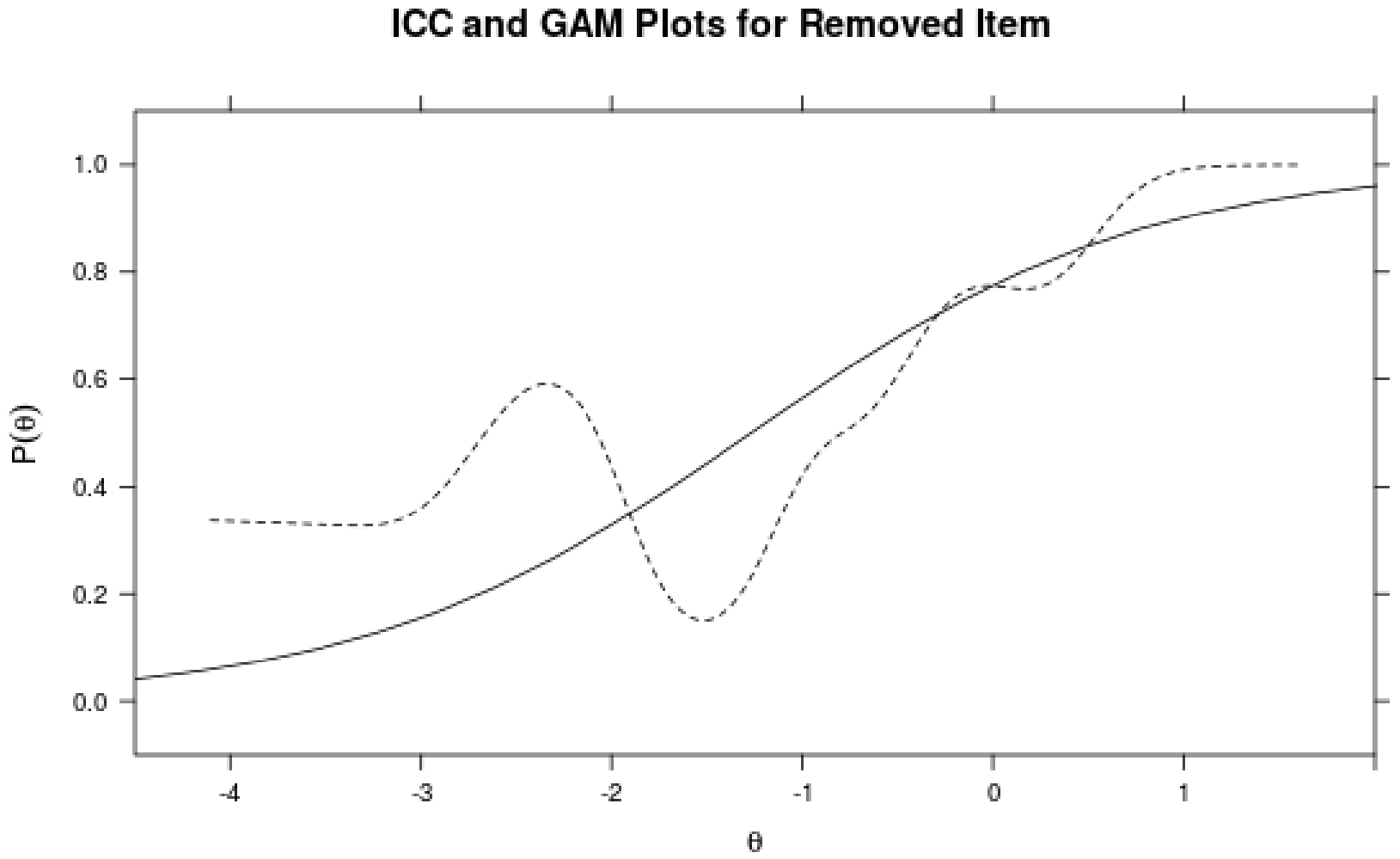
180 sentence pairs to start (30 per data subset)

- 6 removed before model fitting

124 items retained after model fitting (68.9%)

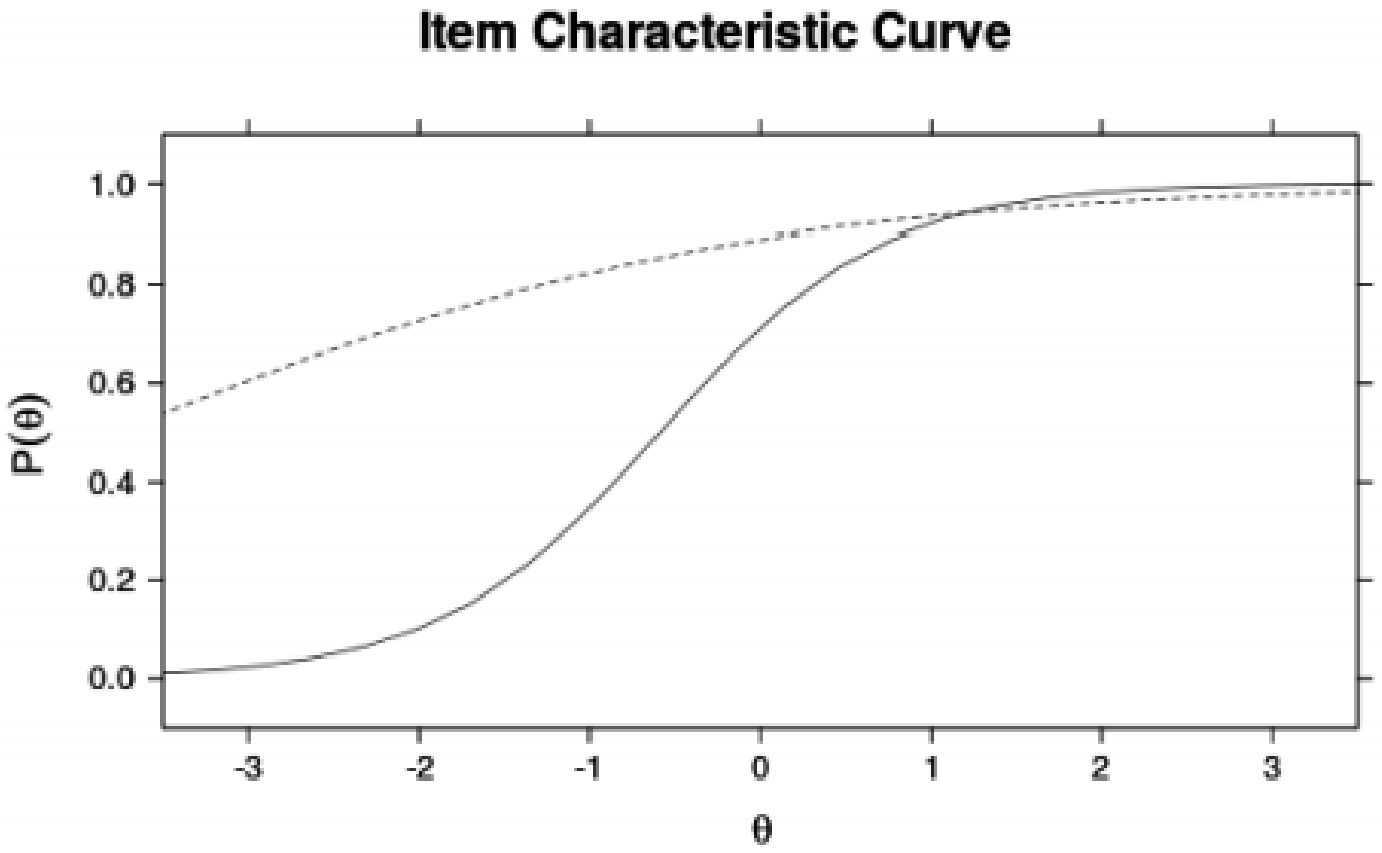
4GS Entailment: Could not build a good model of ability with the AMT response patterns, so all items from 4GS Entailment were removed

Item Removed – Bad Item Fit



Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Good & Bad Items



Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Additional Examples

Text	Hypothesis	Label
Retained - 4GS		
1. A toddler playing with a toy car next to a dog	A toddler plays with toy cars while his dog sleeps	Neutral
2. People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction
Retained - 5GS		
3. A person is shoveling snow	It rained today	Contradiction
4 Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral
5. A woman is kneeling on the ground taking a photograph	A picture is being snapped	Entailment
Removed - 4GS		
6. Two men and one woman are dressed in costume hats	The people are swingers	Neutral
7. Man sweeping trash outside a large statue	A man is on vacation	Contradiction
8. A couple is back to back in formal attire	Two people are facing away from each other	Entailment
9. A man on stilts in a purple, yellow and white costume	A man is performing on stilts	Entailment
Removed - 5GS		
10. A group of soccer players are grabbing onto each other as they go for the ball	A group of football players are playing a game	Contradiction
11. Football players stand at the line of scrimmage	The players are in uniform	Neutral
12. Man in uniform waiting on a wall	Near a wall, a man in uniform is waiting	Entailment

Table 1: Examples of retained & removed sentence pairs. The selection is not based on right/wrong labels but based on IRT model fitting and item elimination process. Note that no 4GS entailment items were retained (Section 4.2)

Additional Examples

Text	Hypothesis	Label
Retained - 4GS		
1. A toddler playing with a toy car next to a dog	A toddler plays with toy cars while his dog sleeps	Neutral
2. People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction
Retained - 5GS		
3. A person is shoveling snow	It rained today	Contradiction
4 Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral
5. A woman is kneeling on the ground taking a photograph	A picture is being snapped	Entailment
Removed - 4GS		
6. Two men and one woman are dressed in costume hats	The people are swingers	Neutral
7. Man sweeping trash outside a large statue	A man is on vacation	Contradiction
8. A couple is back to back in formal attire	Two people are facing away from each other	Entailment
9. A man on stilts in a purple, yellow and white costume	A man is performing on stilts	Entailment
Removed - 5GS		
10. A group of soccer players are grabbing onto each other as they go for the ball	A group of football players are playing a game	Contradiction
11. Football players stand at the line of scrimmage	The players are in uniform	Neutral
12. Man in uniform waiting on a wall	Near a wall, a man in uniform is waiting	Entailment

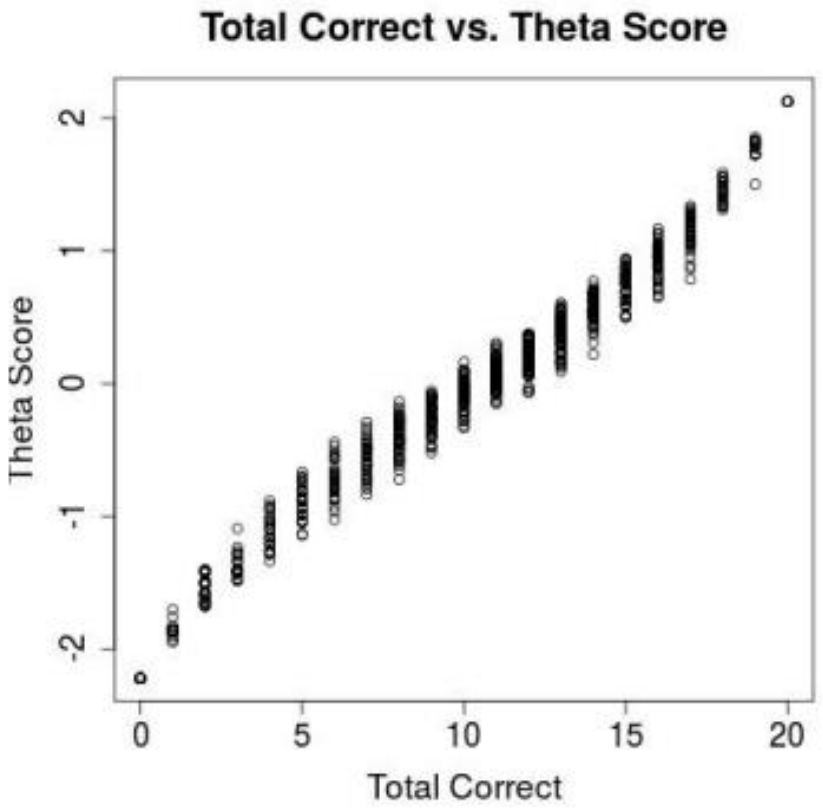
Table 1: Examples of retained & removed sentence pairs. The selection is not based on right/wrong labels but based on IRT model fitting and item elimination process. Note that no 4GS entailment items were retained (Section 4.2)

Additional Examples

Text	Hypothesis	Label
Retained - 4GS		
1. A toddler playing with a toy car next to a dog	A toddler plays with toy cars while his dog sleeps	Neutral
2. People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction
Retained - 5GS		
3. A person is shoveling snow	It rained today	Contradiction
4 Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral
5. A woman is kneeling on the ground taking a photograph	A picture is being snapped	Entailment
Removed - 4GS		
6. Two men and one woman are dressed in costume hats	The people are swingers	Neutral
7. Man sweeping trash outside a large statue	A man is on vacation	Contradiction
8. A couple is back to back in formal attire	Two people are facing away from each other	Entailment
9. A man on stilts in a purple, yellow and white costume	A man is performing on stilts	Entailment
Removed - 5GS		
10. A group of soccer players are grabbing onto each other as they go for the ball	A group of football players are playing a game	Contradiction
11. Football players stand at the line of scrimmage	The players are in uniform	Neutral
12. Man in uniform waiting on a wall	Near a wall, a man in uniform is waiting	Entailment

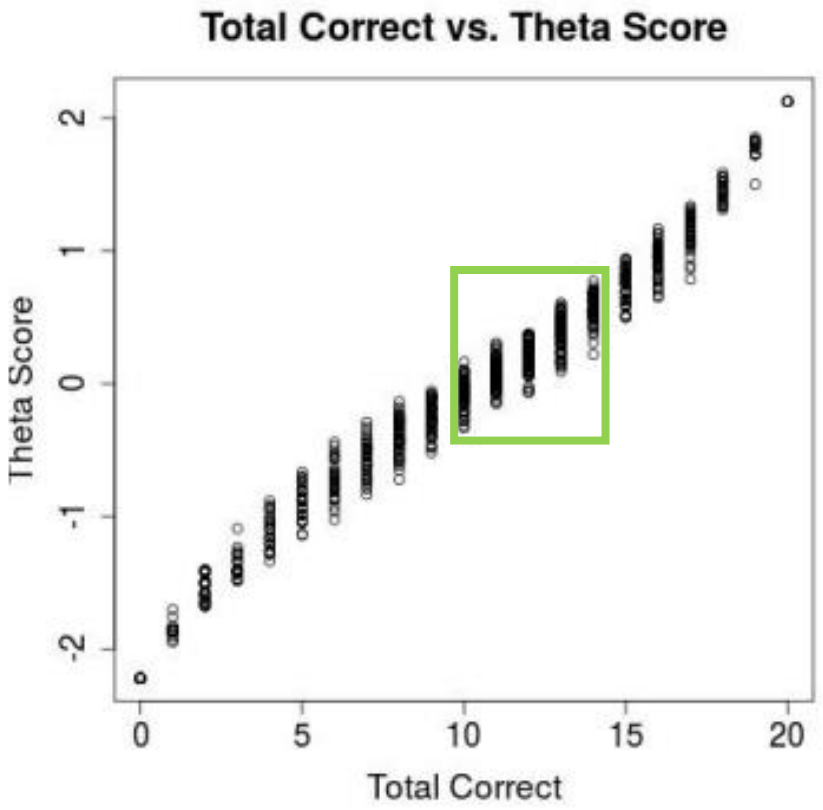
Table 1: Examples of retained & removed sentence pairs. The selection is not based on right/wrong labels but based on IRT model fitting and item elimination process. Note that no 4GS entailment items were retained (Section 4.2)

Results – AMT User Theta Scores



Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Results – AMT User Theta Scores



Same number of correct answers but different theta scores

Which items are correct matters, not just *how many*

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Item Parameters

Item Set	Min. Difficulty	Max. Difficulty	Min. Slope	Max. Slope
5GS				
Contradiction	-2.765	0.704	0.846	2.731
Entailment	-3.253	-1.898	0.78	2.61
Neutral	-2.082	-0.555	1.271	3.598
4GS				
Contradiction	-1.829	1.283	0.888	2.753
Neutral	-2.148	0.386	1.133	3.313

Table 4: Parameter estimates of the retained items

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Item Parameters

Item Set	Min. Diffi-culty	Max. Diffi-culty	Min. Slope	Max. Slope
5GS				
Contradiction	-2.765	0.704	0.846	2.731
Entailment	-3.253	-1.898	0.78	2.61
Neutral	-2.082	-0.555	1.271	3.598
4GS				
Contradiction	-1.829	1.283	0.888	2.753
Neutral	-2.148	0.386	1.133	3.313

Table 4: Parameter estimates of the retained items

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

IRT Evaluation: Test Case

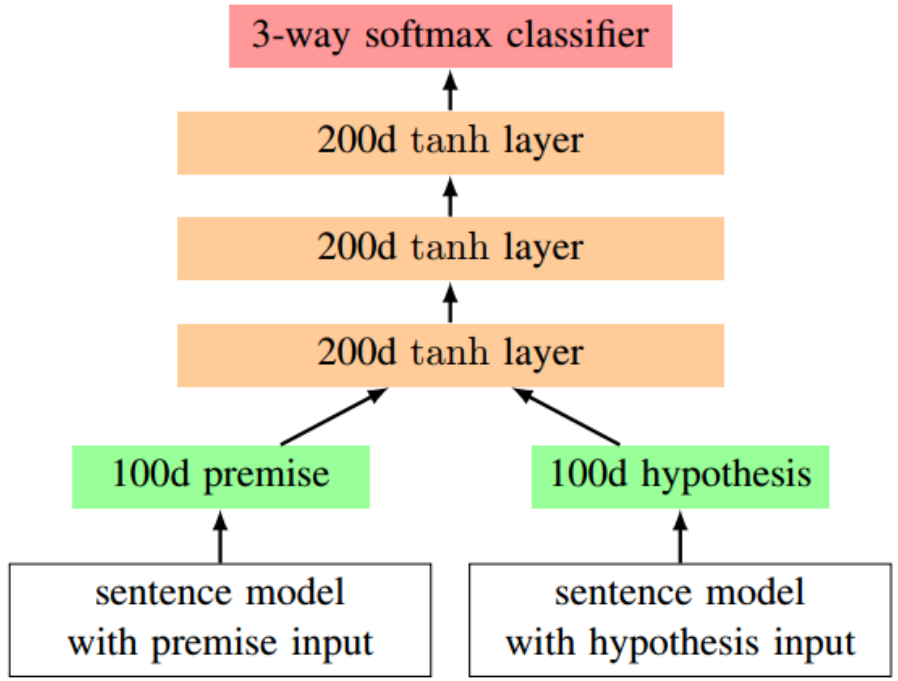
Evaluate an RTE model using IRT scales

- LSTM RNN released with original SNLI dataset

Train on full SNLI training set

Test using our IRT evaluation scales

What do the IRT scores tell us about the model?



Bowman, Samuel R. et al. “A large annotated corpus for learning natural language inference.” Proceedings of EMNLP (2015)

Results – Model Performance

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%

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Results – Model Performance

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%

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Results – Model Performance

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%

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Results – Model Performance

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%

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Results – Model Performance

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%

Lalor, John P. et al. “Building an Evaluation Scale using Item Response Theory.” Proceedings of EMNLP (2016)

Conclusion

Use IRT to build a test set

- Model characteristics of individual items

Compare results to human performance

High accuracy does not imply good performance in terms of human population

Limitations of IRT

Large amount of data required

- AMT helps reduce cost of data collection
- Classical Test Theory is an alternative method, but is test-centric

Fitting IRT models is a manual process

- Can treat item removal criteria as hyperparameters

Ongoing and Future work

Confirm the reliability and consistency of IRT for RTE

Is IRT useful for other NLP tasks?

Other metrics that consider ambiguity in language

- Hellinger Distance

IRT to measure health literacy

Ongoing Work

Hellinger Distance

Treat distribution of human-generated labels as gold standard

Performance is measured as distance between gold standard and NLP model output probabilities

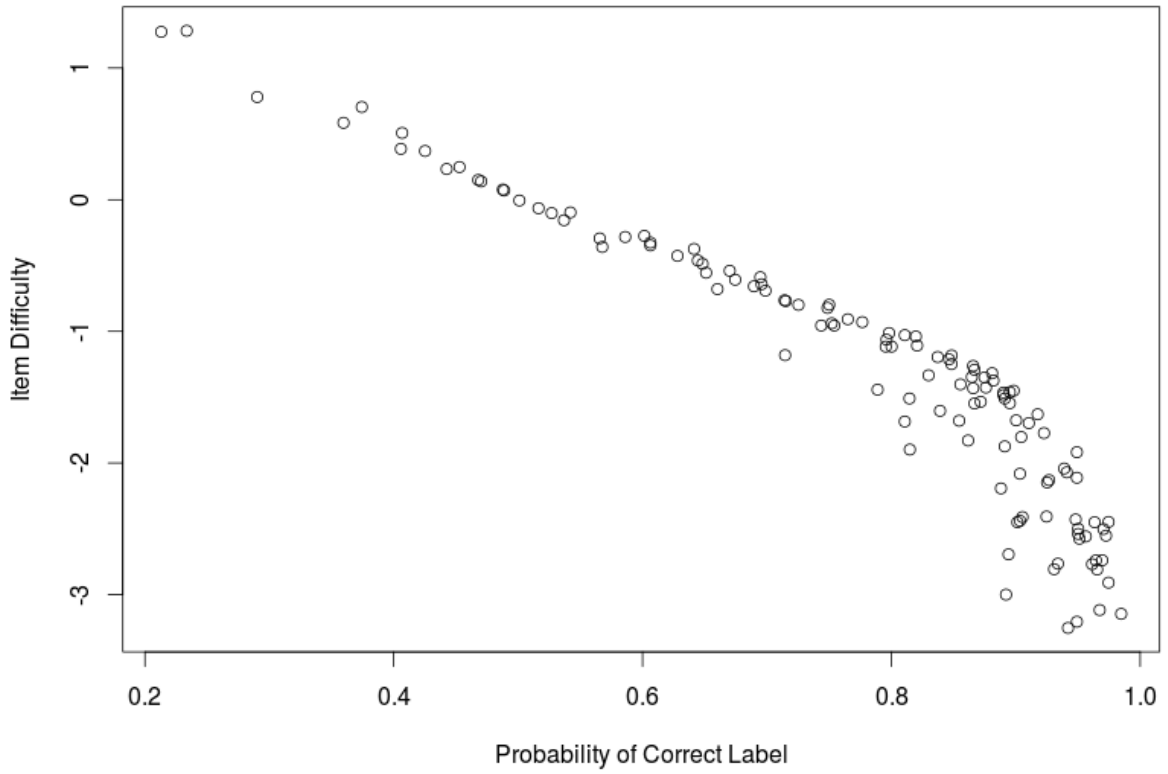
Compare model to “wisdom of the crowd”

Hellinger Distance Score

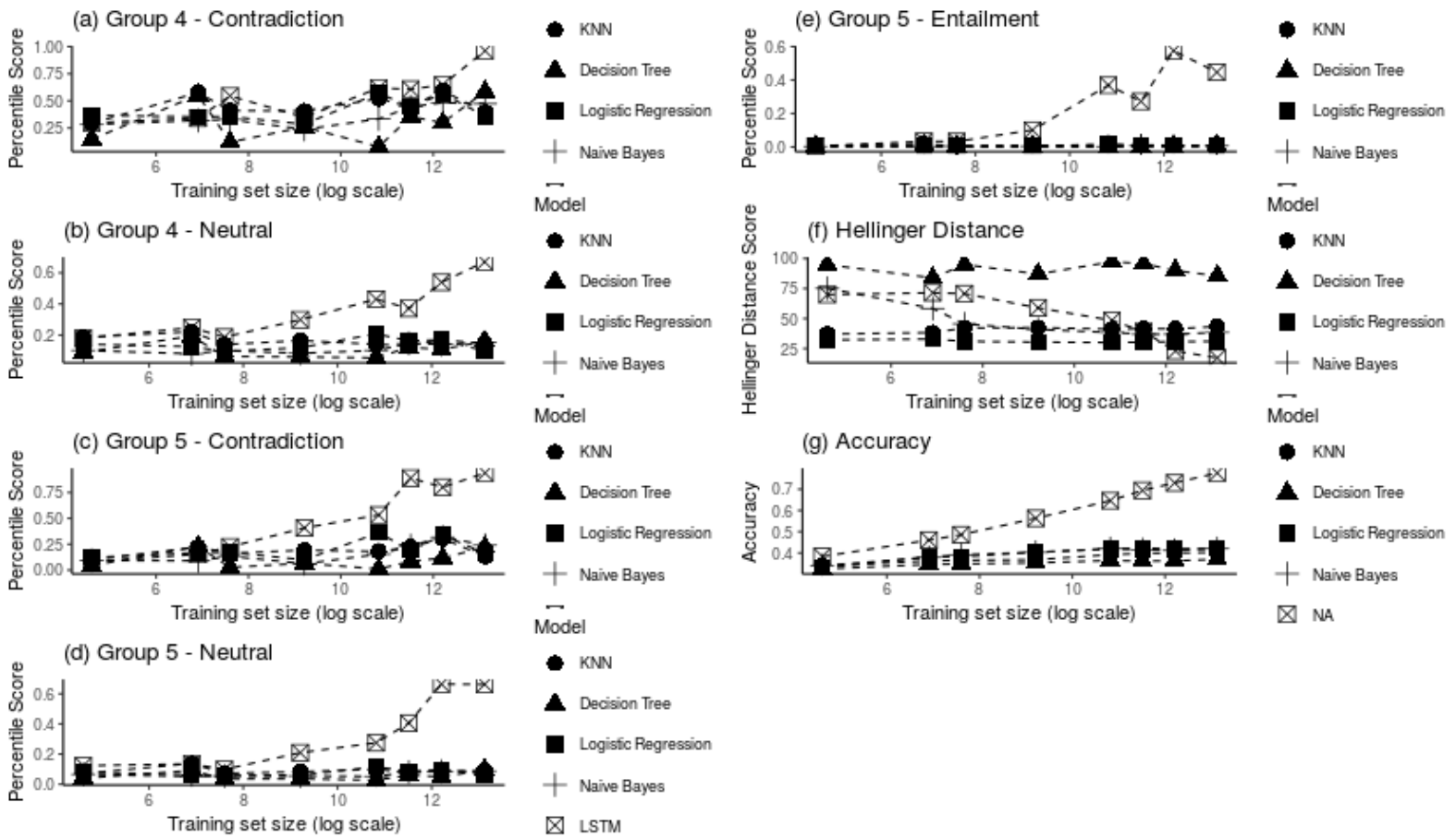
For each item: $HD(P,Q) = \frac{1}{\sqrt{2}} \left\| \sqrt{P} - \sqrt{Q} \right\|_2$

Test set score: $HD(X_{test}) = \sum_{x \in X_{test}} HD(P_x, Q_x)^2$

Probability Correct and Item Difficulty

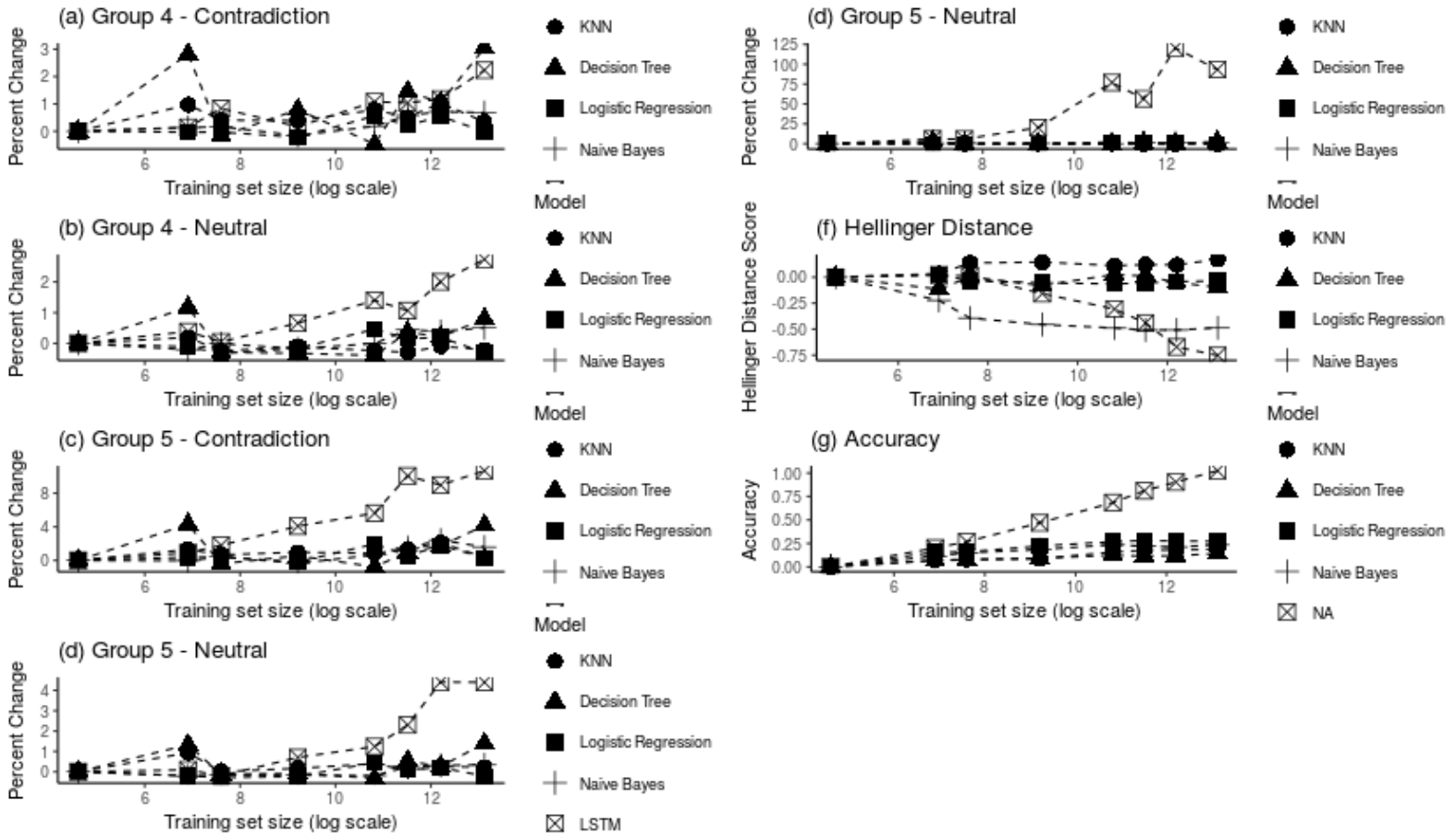


Comparing NN to other ML Models



Performance Change Across Training Sizes

Comparing NN to other ML Models



Cumulative Change Across Training Sizes

Thank you!