Comparing Human and DNN-Ensemble Response Patterns for Item Response Theory Model Fitting

John P. Lalor [1,2*] Hao Wu [3] Hong Yu [1,4]

[1]University of Massachusetts [2]University of Notre Dame [3]Vanderbilt University [4]Bedford VA

[*]Email: john.lalor@nd.edu, Web: http://jplalor.github.io

Introduction

Item Response Theory (IRT) models for natural language processing tasks can provide valuable information about model performance and behavior. However, a significant bottleneck to the IRT model building process is the need to obtain human response patterns (RPs) to fit the models. Can we replace human RPs with RPs from an ensemble of neural networks?

Item Response Theory

IRT models are designed to estimate latent ability parameters (θ) of subjects and latent item parameters such as difficulty of items (b). The probability that subject j will answer item i correctly is:

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

The probability that subject j will answer item i incorrectly is

$$p(y_{ij} = 0|\theta_j, b_i) = 1 - p(y_{ij} = 1|\theta_j, b_i) \quad (2)$$

The likelihood of a data set of RPs Y from J subjects to a set of I items is:

$$p(Y|\theta, b) = \prod_{j=1}^{J}\prod_{i=1}^{I} p(Y_{ij} = y_{ij}|\theta_j, b_i) \quad (3)$$

The item parameters are typically estimated by marginal maximum likelihood (MML) via an Expectation-Maximization (EM) algorithm [1], in which subject parameters are considered random effects $\theta_i \sim N(0, \sigma_\theta^2)$ and marginalized out. Once item parameters are learned, subjects' θ parameters are scored typically with maximum a posteriori (MAP) estimation. For the human and machine RP models, we fit a Rasch model using the mirt R package [3]. We then calculate the correlation between the fit parameters to determine if the items' difficulty parameters were consistent.

Data

In order to determine whether an IRT model fit using machine RPs is reliable and interpretable, we first need to compare the model to one learned using human response patterns. By comparing IRT models fit with human and machine RPs we can look at the learned item parameters for both models to identify correlations in item difficulties. That is, are items that are easy for humans also easy for machines? We also expect that certain properties of IRT models hold true when they are fit with machine RPs (e.g. that raw accuracy and latent ability are highly correlated). To do this we use the human response pattern data collected to learn IRT models in prior work [5, 4]. In the prior work the authors collected human annotations for examples selected from the Stanford Natural Language Inference (SNLI) and Stanford Sentiment Treebank (SSTB) datasets from 1000 Amazon Mechanical Turk workers [2, 7]. For each Turker a response pattern was generated to indicate which items the Turkers labeled correctly based on the gold standard label.

References

[1] R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443--459, 1981.

[2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and D. Christopher Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[3] Phil Chalmers, Joshua Pritikin, Alexander Robitzsch, and Mateusz Zoltak. mirt: Multidimensional Item Response Theory, November 2015.

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

[5] John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *EMNLP*, 2016.

[6] Tsendsuren Munkhdalai and Hong Yu. Neural semantic encoders. *EACL 2017*, 2017.

[7] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, D. Christopher Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

Generating Response Patterns

In order to generate machine RPs for our comparisons, we trained a set of LSTM models with varying degrees of training set sizes and noise [2]. More specifically, we trained 1000 LSTM models for NLI classification using the SNLI data set and 1000 LSTM models for binary SA classification using the SSTB dataset [2, 7]. For each model m_i, we randomly sampled a subset of the task training set, x^i_{train}. For each training set x^i_{train}, we corrupted the training labels for a randomly selected percentage of the training set. For each training set pair that was selected for label corruption, the gold standard label was replaced with an incorrect label. For SNLI, one of the two incorrect labels was chosen at random, and for SSTB the correct label was replaced with the incorrect label. For each model and training set pair, we trained the model, used the held out validation set for early stopping, and wrote the model's graded (correct/incorrect) output on the IRT test set to disk as that model's *response pattern*. The set of response patterns for all of the models is our input dataset for the IRT model.

We also looked at a more complex model to determine if the learned parameters would differ given the different model architectures. For our more complex model we used the Neural Semantic Encoder model (NSE), a memory-augmented RNN [6].

Discussion

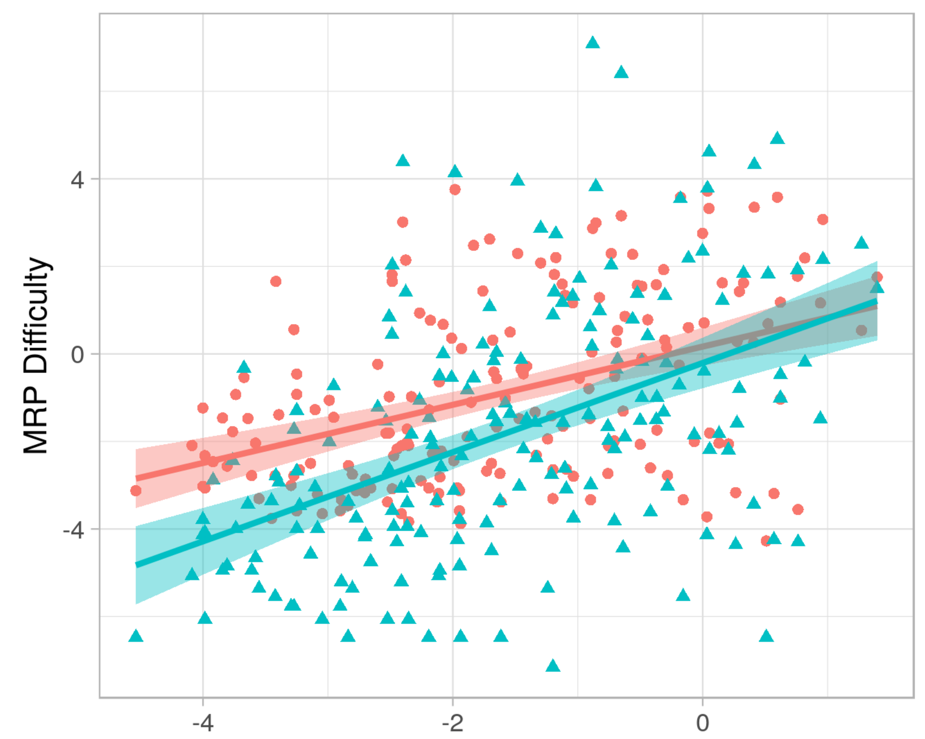
For both SNLI and SSTB, we find that there is a positive correlation between the item difficulties of IRT models fit using human and machine RPs. In addition, the more complex NSE model has a higher correlation score with the human-learned difficulty parameters than the LSTM model. This shows that creating more complex DNN architectures does have bearing on how the model identifies difficult items with regards to human expectations.

Where are the differences?

- SNLI
  - Contradiction or neutral (row 1)?
  - Multi-step reasoning (row 2): what happens after the unwrapping?
- SSTB
  - Name dropping to indicate quality (row 3)
  - Focus (or lack of) on specific keywords (row 4)

Results: Human-Machine Comparisons and Analysis of Disagreements

(a) SNLI
Spearman ρ: 0.409 (LSTM) and 0.496 (NSE).



(b) SSTB
Spearman ρ: 0.332 (LSTM) and 0.392 (NSE).

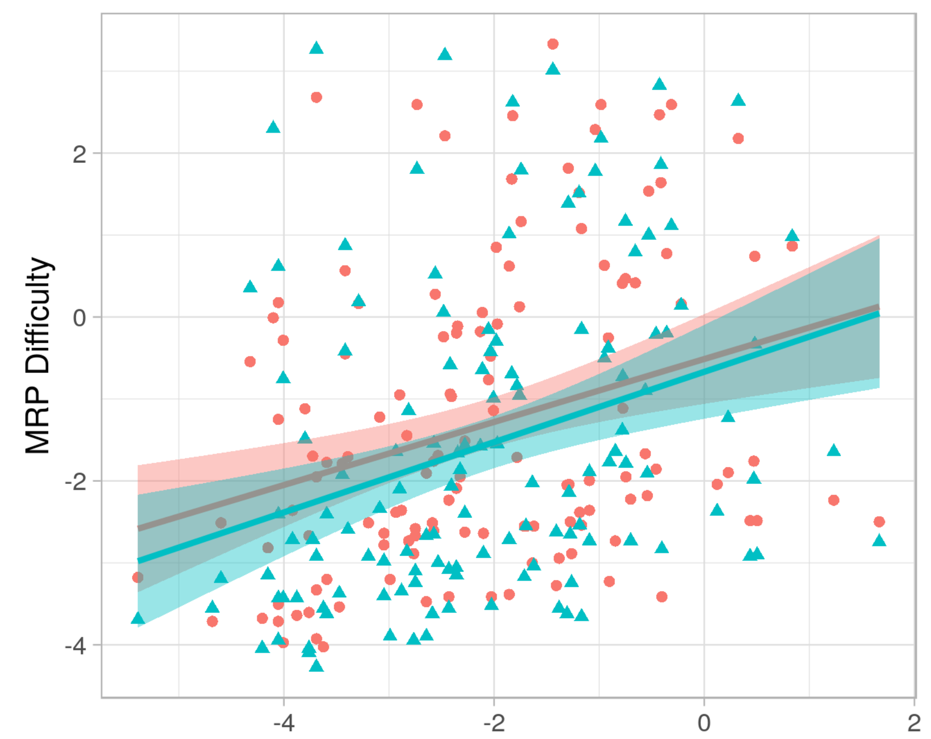


Figure: Item difficulty parameters for the human RPs (x-axis) and machine RPs (y-axis) models for NLI (1a) and SA (1b).

Task	Label	Item Text	Difficulty ranking Humans	LSTM	NSE
SNLI	Contradiction	*P:* Two dogs playing in snow. *H:* A cat sleeps on floor	168	1	5
	Entailment	*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
	Entailment	*P:* A man with a dog is seated at the base of a statue. *H:* The man and the dog are by the statue	12	131	97
SSTB	Positive	Only two words will tell you what you know when deciding to see it: Anthony. Hopkins.	9	103	110
	Negative	...are of course stultifyingly contrived and too stylized by half. Still, it gets the job done--a sleepy afternoon rental.	128	46	41

Table: Examples from the SNLI and SSTB datasets where the ranking in terms of difficulty varies widely between human and DNN models. In all cases difficulty is ranked from easy to hard.